

OLMoASR: OPEN MODELS AND DATA FOR TRAINING ROBUST SPEECH RECOGNITION MODELS

Huong Ngo^{1,2*} Matt Deitke¹ Martijn Bartelds³ Sarah Pratt²
 Josh Gardner[†] Matt Jordan^{1,†} Ludwig Schmidt^{2*,3,†}

¹Allen Institute for AI, ²University of Washington, ³Stanford University

ABSTRACT

Improvements in training data scale and quality have led to significant advances, yet its influence in speech recognition remains underexplored. In this paper, we present a large-scale dataset, OLMoASR-POOL, and series of models, OLMoASR, to study and develop robust zero-shot speech recognition models. Beginning from OLMoASR-POOL, a collection of 3M hours of English audio and 17M transcripts, we design text heuristic filters to remove low-quality or mistranscribed data. Our curation pipeline produces a new dataset containing 1M hours of high-quality audio-transcript pairs, which we call OLMoASR-MIX. We use OLMoASR-MIX to train the OLMoASR suite of models, ranging from 39M (tiny.en) to 1.5B (large.en) parameters. Across all model scales, OLMoASR achieves comparable average performance to OpenAI’s Whisper on short and long-form speech recognition benchmarks. Notably, OLMoASR-medium.en attains a 12.8% and 11.0% word error rate (WER) that is on par with Whisper’s largest English-only model Whisper-medium.en’s 12.4% and 10.5% WER for short and long-form recognition respectively (at equivalent parameter count). OLMoASR-POOL, OLMoASR-MIX, OLMoASR models, and filtering, training and evaluation code will be made publicly available to further research on robust speech processing.

1 INTRODUCTION

Foundation models trained on web-scale data have changed the landscape of AI. Scaling up models for language, vision-language, and speech has led to breakthroughs such as GPT (Brown et al., 2020), CLIP (Radford et al., 2021), and Whisper (Radford et al., 2023), and the generalization capabilities of these new models has enabled a wide range of new applications. Training data is key to these advances: modern AI models rely on large training sets harvested from the Web that combine both broad data collection with detailed curation. For instance, the latest language models are now trained on trillions of text tokens produced by sophisticated data pipelines (Grattafiori et al., 2024; Liu et al., 2024; Li et al., 2024; OLMo et al., 2024; Liu et al., 2023).

The importance of web-scale training data has led to increasing interest in datasets, including several efforts to build open datasets for training foundation models. In the text domain, researchers have introduced a multitude of datasets such as C4 (Raffel et al., 2023), the Pile (Gao et al., 2020), RedPajama (Weber et al., 2024), RefinedWeb (Penedo et al., 2023), Dolma (Soldaini et al., 2024), DCLM (Li et al., 2025), FineWeb (Penedo et al., 2024a), Nemotron-CC (Su et al., 2025), etc. In addition, researchers have proposed a wide range of data curation methods (Li et al., 2024; Su et al., 2024; Penedo et al., 2024a; 2023; Wettig et al., 2025). Together, these efforts have enabled multiple open source language models that in some cases are competitive with closed-source models and serve as an important starting point for open research. Similarly, the open image-text datasets such as YFCC, LAION, and DataComp have served as a catalyst for research on multimodal learning, leading to reproductions of frontier commercial models such as OpenCLIP (Cherti et al., 2023). The speech domain, however, is currently lagging behind the other modalities: where there are important efforts such as OWSM (Peng et al., 2023) and YODAS (Li et al., 2023), there is currently no

[†]Equal senior contributions. Authors are listed alphabetically by last name.

*Part of work done while at University of Washington.

open-source reproduction of the full-scale Whisper (Radford et al., 2023) models, nor a publicly available training set to begin such an effort. This is despite the widespread use and significant impact of the Whisper model¹, and the stated importance of large-scale, high-quality data to Whisper’s performance (Radford et al., 2023).

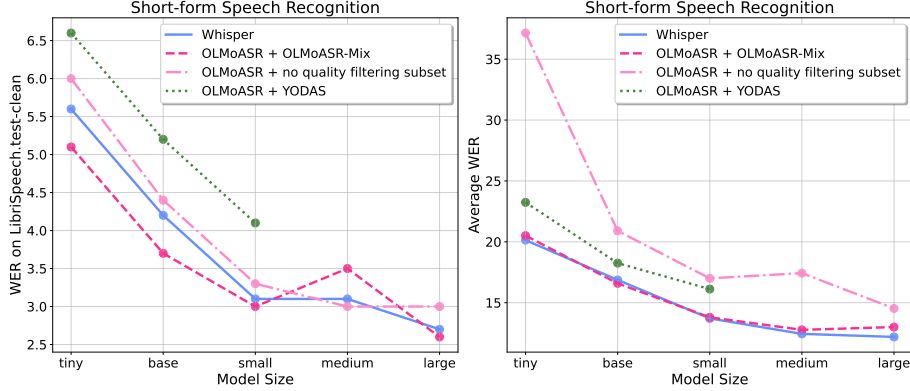


Figure 1: Performance on LibriSpeech.test-clean (left) and average performance across 14 short-form speech recognition benchmarks (right) of each baseline for all possible model scales.

We address this shortcoming in the open data ecosystem by introducing OLMoASR-POOL, a dataset with 3M hours of audio and associated transcripts taken from the public internet. Starting from these audio-text pairs, we build a careful data curation pipeline that allows us to assemble a high-quality subset for training state-of-the-art robust, zero-shot speech recognition models. As a result of this pipeline, we propose OLMoASR-MIX, a dataset with 1M hours of audio and accompanying transcripts, that surpasses the scale of data used to train the initial Whisper models. Our data scale matches the amount of weakly labeled data used to train the second and third versions of the Whisper models.

We validate the quality of OLMoASR-MIX by training a range of models following the Whisper architecture and training recipe. The resulting family of models, OLMoASR, closely matches the quality of Whisper on a wide range of benchmarks and across multiple compute scales up to the largest Whisper model scale (see Figure 1 and Figure 3). In addition, our models outperform other open data speech recognition models such as OWSM (see Table 8), wav2vec, and HuBERT (see Table 9) across a range of short- and long-form speech recognition benchmarks. Our experiments show that our data curation pipeline is key to the success of our models: compared to a baseline that was trained on a filtered OLMoASR-POOL to remove non-English audio-transcript pairs, our actual training set OLMoASR-MIX consistently improves performance across compute scales (see Figure 3). A key step in our pipeline is removing repeating lines, which improves performance by 14.5% WER (percentage points). A dataset quality ablation also demonstrates that training an OLMoASR model on a weakly-supervised, web-scale data collection like OLMoASR-MIX results in better performance across many evaluation sets compared to training on academic datasets.

We publicly release the IDs of the audio-text pairs in OLMoASR-POOL and OLMoASR-MIX as a starting point for research on speech training data. Our hope is that this will enable new research on data curation and speech recognition, similar to the LAION-5B project for multimodal learning and DataComp-LM for language modeling. In addition, providing a web-scale speech recognition training set increases transparency around current approaches to AI training and enables research on bias in datasets, fairness, privacy, and data auditing. Given the sensitive nature of training data, we strongly recommend that OLMoASR-POOL and OLMoASR-MIX should only be used for academic research purposes in its current form. We advise against any applications in deployed systems without carefully investigating the legal, privacy, and fairness risks associated with OLMoASR-POOL and OLMoASR-MIX.

¹Whisper is, for example, OpenAI’s most-starred repository on GitHub, and various official versions of Whisper have garnered at least 17M downloads on Hugging Face as of the date of this publication.

Our training data is available at <https://huggingface.co/datasets/allenai/OLMoASR-Pool>, code can be found at <https://github.com/allenai/OLMoASR> and models are available at <https://huggingface.co/datasets/allenai/OLMoASR>.

2 DATA

2.1 MOTIVATION

Whisper (Radford et al., 2023) demonstrated an approach of scaling speech recognition datasets to achieve strong generalization and robustness. While the related work focused on studying the impact of data scaling on zero-shot generalization, not much is known about the impact of dataset design and the dataset itself was never made publicly accessible.

Open Whisper-style Speech Model (OWSM) (Peng et al., 2023; 2024a;b) is an effort to reproduce Whisper with open-source tools, but trained on a mix of academic datasets. (Tian et al., 2024) investigates the effects of data quality on OWSM models using the same mix. However, studying and training on such data pool does not enable rigorous investigation into Whisper’s zero-shot capability. Moreover, the performance of those models demonstrate that despite improvements to the architecture or training recipe, dataset composition plays a central role in supplying the model’s generalization and robust capabilities.

To address this knowledge gap, we conduct experiments on a collection of weakly-supervised audio-transcript data that is on the same scale as Whisper’s dataset to analyze how different dataset design choices affect a speech recognition model’s downstream performance. Model architecture, training code and evaluation setup are controlled and only the data is changed. More specifically, we use the same architecture, tokenizer and evaluation setup as Whisper. As Whisper did not publish their training and data processing code, we construct a training loop and data processing pipeline to the best of our abilities to match what Whisper used. To validate that our training loop was correct, we monitor the model’s training and validation loss curves.

2.2 CURATION

In this section, we describe the curation choices made to achieve OLMoASR-MIX and quantify the impact of the curation layer. Firstly, to ensure that the audio and text language matches, we perform audio-text language alignment (Section 2.2.1). Next, we experiment with different text-based heuristics (Section 2.2.2) to remove low-quality audio-text pairs. We will also explain what type of low-quality data we are targeting at each layer. Finally, we perform fuzzy decontamination and deduplication (Appendix C) on the transcripts to remove contaminated or duplicated audio-text pairs. Figure 2 visually illustrates each step in removing low quality data and denotes their respective percentages of data removed.

All experiments are performed on the OLMoASR-tiny.en model and compared to a baseline that has only been trained on data filtered with the audio-text language alignment filter. We use this baseline as it does not target the quality of transcripts. This will be referred to as the “no quality filtering” baseline from this point onward. To assess them, we use the word error rate (WER) metric which calculates the percentage of words that were incorrectly predicted when compared to a reference text.

2.2.1 AUDIO-TEXT LANGUAGE ALIGNMENT

To ensure that we are training on only English audio-text pairs, a spoken language identification model, VoxLingua107 (Valk & Alumäe, 2021), is used to tag the audio sample with the spoken language, and `pycld2` to tag the corresponding transcript sample with the text language. The top-1 predicted language from both models are chosen as the tagged languages. We then remove audio-text pairs where the tagged audio and text language are not both English.

2.2.2 TEXT HEURISTICS

There is no guarantee that the transcripts are manual transcriptions of the audio from the public internet. In fact, many publicly accessible transcripts are produced by speech recognition sys-

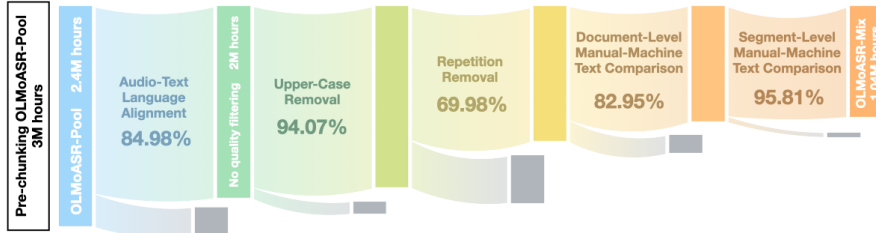


Figure 2: Construction of OLMoASR-Mix from OLMoASR-POOL. Segmentation reduces OLMoASR-POOL from 3M to 2.4M hours. Percentages are relative to most recent filtered subset, based on the number of hours or segments.

tems. Recent work has illustrated inferior performance from training on automatic transcripts for speech recognition (Li et al., 2023) or a mix of human and machine-labeled data on translation systems (Fernandes et al., 2023). Through manual examination, we identify text characteristics of machine-generated transcripts that can be used as filtering heuristics.

Exploratory analysis uncovered a non-trivial number of audio-text pairs where the transcriptions are unfaithful and unrelated to the audio. There are also instances of partial transcriptions, and temporally misaligned transcriptions. To remove them, each audio-text pair is scored based on the WER between the manually uploaded and an associated machine-generated text, then omitting pairs where the score is lower than a specific threshold.

Text casing. Through examining audio-text pairs, we noticed that a lot of machine-generated transcripts contain text that is mostly made up of lower or upper case characters. We designed a case detector to loop through each transcript line and keep counts of the respective cases. The case type with the highest frequency is the resulting case tag of the audio-text pair. In Table 1, removing audio-text pairs that have been tagged with upper or lower case types improves short-form WER by 4.8% after removing 32.0% of the data.

Filtering strategy	Data hours	Percent remaining (%)	Short-form WER
No quality filtering	2,010,447	-	37.2
Upper-case or lower-case removal	1,367,506	68.0	32.4

Table 1: Filtering out transcripts with upper or lower case improves WER performance on short-form transcription. Short-form WER refers to the average performance across 14 short-form speech recognition datasets. Percent remaining is based on number of hours or segments remaining from filter relative to the no quality filtering strategy.

Presence of repeating lines. Another issue with machine-generated transcripts is the repetition of lines, which can misalign audio and text. To detect these, we check if each line matches the previous one exactly. Table 2 shows that the removal of repeats reduces the short-form WER by 14.4% while removing 39.9% of the data.

We also test combining casing and repeat filters. Table 2 shows that filtering by both repeats, and lower and upper case yields a WER 0.7% higher than just repeats and mostly uppercase text, while removing more data. Therefore, our final curation only filters based on the presence of repeating lines and mostly upper case text.

Filtering strategy	Data hours	Percent remaining (%)	Short-form WER
No quality filtering	2,010,447	-	37.2
Repeating lines removal	1,207,676	60.1	22.7
Repeating lines removal and upper-case removal	1,139,722	56.7	21.9
Repeating lines removal and upper and lower case removal	944,106	47.0	22.6

Table 2: Filtering out transcripts with repeating lines improves WER performance on short-form transcription. Short-form WER refers to the average performance across 14 short-form speech recognition datasets. Percent remaining is based on number of hours or segments remaining from filter relative to the no quality filtering strategy.

Manual-machine text comparison. Unfaithful or misaligned transcripts can cause the model to learn from poorly matched audio-text pairs. To filter these, we compare a manual transcript with its machine-generated version using WER. Although automatic transcripts are less precise, they reliably capture speech utterances, making them effective for identifying low-quality data. Pairs with WER above a set threshold are removed.

We use two variants: manual-machine *document-level* and *segment-level* comparison. The document-level filter mainly detects unrelated transcripts, but might confound minor differences with misalignments. Manual inspection also showed that sections of poorly-aligned transcripts can be recovered, so we also utilize a segment-level filter for more fine-grained filtering. Through experiments, we determined thresholds of 0.5 for document-level and 0.7 for segment-level filtering.

Table 3 illustrates that employing this filter improves WER performance on short-form transcription by 16.5% after removing 54.8% of the data.

Filtering strategy	Data hours	Percent remaining (%)	Short-form WER
No quality filtering	2,010,447	-	37.2
Manual-machine text comparison	908,923	45.2	20.7

Table 3: Employing manual-machine text comparison filter improves WER performance on short-form transcription. Short-form WER refers to the average performance across 14 short-form speech recognition datasets. Percent remaining is based on number of hours or segments remaining from filter relative to the no quality filtering strategy.

3 MODEL AND TRAINING

3.1 MODEL

To fully understand the impact of our data curation methodology on producing robust speech recognition systems with strong zero-shot capabilities, we utilize Whisper’s model architecture and tokenizer. We have only modified the architecture code to use FlashAttention (Dao et al., 2022) in the attention module and incorporate the causal and padding mask for batch training.

3.2 TRAINING DETAILS

In contrast to the Whisper training procedure, we train with a larger batch size, reconfigure the learning rate and warmup scheduler, and total steps trained accordingly. This was done to leverage available compute and maximize efficient distributed training. Moreover, we retain the same maximum learning rate for all scales and do not perform hyperparameter tuning. For OLMOASR-tiny.en, OLMOASR-base.en and OLMOASR-small.en we train with Distributed Data Parallel (DDP) using FP16 with dynamic loss scaling. In contrast, OLMOASR-medium.en and OLMOASR-large.en

were trained with Fully Sharded Data Parallel (FSDP) using bfloat16 with dynamic loss scaling and activation checkpointing. We found that training with FSDP using bfloat16 provided better training stability than with FP16.

4 RESULTS AND DISCUSSION

4.1 EVALUATION METHODOLOGY

To properly assess how our dataset design approach contributes to OLMoASR’s zero-shot generalization ability, we evaluate our model on a suite of 21 datasets that have not been used for training, 14 short-form and 7 long-form sets. To maintain comparable evaluation with Whisper (Radford et al., 2023), we use greedy decoding for short-form and beam search for long-form. The evaluation sets will assess the model’s capabilities in contexts such as audio book recordings, lectures, calls and meetings. Moreover, these datasets contain speech of short and long utterances, different accents, high and low signal clarity. We also study how useful OLMoASR-MIX is as a robustness intervention, utilizing effective and relative robustness from (Taori et al., 2020).

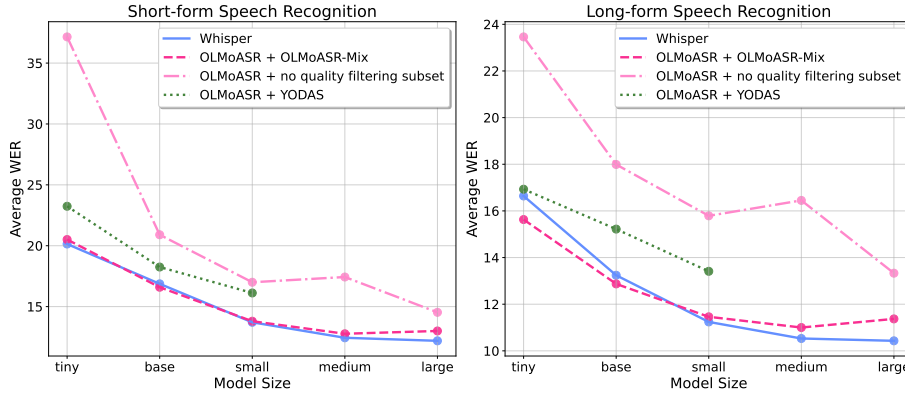


Figure 3: Average performance across 14 short-form speech recognition benchmarks (left) and across 7 long-form speech recognition benchmarks (right) of each baseline for all possible model scales.

4.2 ZERO-SHOT PERFORMANCE ACROSS DATASETS

Primary results. Average performance across 14 short and 7 long-form evaluation sets can be found in Figure 3. Short-form performance from each dataset can be found in Table 4 and long-form results can be found in Table 5. Below, we establish core findings from our main baseline.

OLMoASR-MIX enables OLMoASR’s competitive zero-shot capability. From Figure 3, OLMoASR is comparable to Whisper (Radford et al., 2023), the current state-of-the-art zero-shot ASR, with the largest average performance gap being 0.4% for short-form and 1% for long-form.

For short-form transcription, OLMoASR performs on-par with Whisper at tiny to small scales. However, the gap widens at 769M and 1.5B, which may be due to lack of hyperparameter tuning or differences in data scale. Specifically, OLMoASR-large.en was trained on 440K hours of English data per pass, while Whisper used 680K hours of multilingual data. For a more fair comparison, we re-trained OLMoASR-large.en on 680K hours of English data which reduces the gap from 0.8% to 0.4%. This is denoted as OLMoASR-large.en-v2 on 4.

For long-form transcription, OLMoASR-tiny.en and OLMoASR-base.en outperform Whisper’s equivalents, and OLMoASR-small.en is on par with Whisper-small.en. At larger scales, the performance gap reappears for similar reasons as in short-form transcription.

Data curation is vital to achieve strong zero-shot generalization. OLMoASR on all model scales benefits from data curation, especially OLMoASR-tiny.en for short-form and long-form,

Model	LibriSpeech.test-clean	LibriSpeech.test-other	TED-LIUM3	WSJ	CallHome	Switchboard	CommonVoice5.1	Arctic	CORAAL	CHiME6	AMI-IHM	AMI-SDM	VoxPopuli.en	Fleurs.en.us	Average
OLMoASR (Open weights, code, data) vs. Whisper (Open weights, closed training code, data)															
OLMoASR-tiny.en	5.1	12.3	5.5	5.6	23.9	18.7	25.1	19.3	25.7	45.2	24.2	55.4	11.6	9.7	20.5
Whisper tiny.en	5.6	14.6	6.0	5.0	24.1	17.8	26.3	20.0	23.9	41.3	23.7	50.3	11.7	11.6	20.1
OLMoASR-base.en	3.7	9.0	4.6	4.3	20.5	14.0	18.5	13.6	21.5	38.0	20.4	47.8	9.7	6.7	16.6
Whisper base.en	4.2	10.2	4.9	4.6	20.9	15.2	19.0	13.4	22.6	36.4	20.5	46.7	10.0	7.6	16.9
OLMoASR-small.en	3.0	7.0	4.2	3.8	16.7	13.2	13.1	9.6	19.6	30.6	18.7	39.9	8.7	5.0	13.8
Whisper small.en	3.1	7.4	4.0	3.3	18.2	15.7	13.1	9.7	20.2	27.6	17.5	38.0	8.1	6.0	13.7
OLMoASR-medium.en	3.5	5.7	5.0	3.6	14.3	12.7	11.3	7.5	18.7	28.5	16.9	38.3	8.4	4.4	12.8
Whisper medium.en	3.1	6.3	4.1	3.3	16.2	14.1	10.6	7.6	17.5	25.3	16.4	37.2	7.4	5.0	12.4
OLMoASR-large.en	2.6	5.9	4.5	3.7	16.5	12.7	11.1	7.9	18.7	30.7	16.4	38.8	8.1	4.5	13.0
OLMoASR-large.en-v2	2.7	5.6	4.2	3.6	15.0	11.7	11.1	7.8	18.1	29.4	17.1	38.0	8.0	4.2	12.6
Whisper large-v1	2.7	5.6	4.0	3.1	15.8	13.1	9.5	6.7	19.4	25.6	16.4	36.9	7.3	4.6	12.2
Whisper large-v2	2.7	5.2	4.0	3.9	17.6	13.8	9.0	6.2	16.2	25.5	16.9	36.4	7.3	4.4	12.1
Whisper large-v3	2.0	3.9	3.9	3.5	14.0	13.2	8.4	5.9	18.7	26.8	16.0	34.2	9.5	4.0	11.7
Whisper large-v3-turbo	2.2	4.2	3.5	3.5	13.2	12.9	9.7	6.3	18.6	27.3	16.1	35.2	12.2	4.4	12.1

Table 4: Short-form English transcription WER (%) with greedy decoding, comparing between OLMoASR and Whisper models.

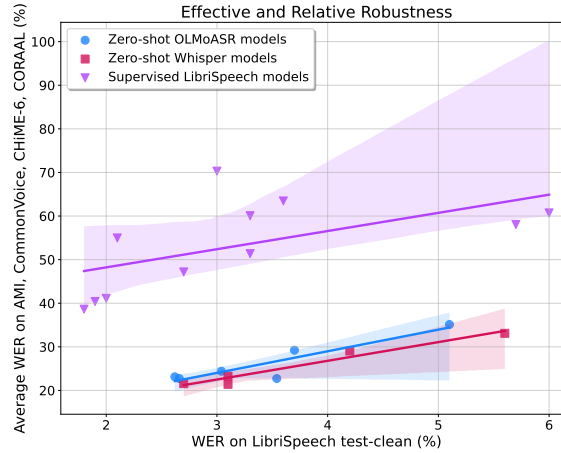


Figure 4: We plot 11 supervised models trained on LibriSpeech without any robustness interventions and demonstrate their WER on a reference test set and the average WER across 5 out-of-distribution evaluation sets. We also plot zero-shot OLMoASR models to compare to the standard models, and Whisper models to demonstrate OLMoASR’s similar robustness capability.

and OLMoASR-medium.en for long-form. This can be observed from the performance discrepancy between OLMoASR trained on the no quality filtering subset and OLMoASR-MIX on short and long-form in Figure 3.

4.3 ROBUSTNESS GAINED FROM WEB-SCALE DATA

Effective robustness. Following (Taori et al., 2020), effective robustness measures how much a model outperforms the expected baseline on out-of-distribution data, given its in-distribution performance. A positive gap indicates stronger robustness than a standard model.

To evaluate OLMoASR’s effective robustness, we use LibriSpeech test-clean as the in-distribution set and five out-of-distribution sets: AMI (AMI-IHM, AMI-SDM), CommonVoice, CHiME-6, and CORAAL, covering diverse speakers and conditions.

Model	TED-LIUM3	Meanwhile	Kincaid46	Rev16	Earnings-21	Earnings-22	CORAAL	Average
OLMoASR (Open weights, code, data) vs. Whisper (Open weights, closed training code, data)								
OLMoASR-tiny.en	4.8	12.6	13.6	14.0	14.2	20.0	30.2	15.6
Whisper tiny.en	5.5	12.8	13.8	15.1	17.0	22.0	30.3	16.6
OLMoASR-base.en	3.9	10.2	11.2	12.0	11.1	15.6	26.1	12.9
Whisper base.en	4.6	9.4	11.2	13.2	12.5	16.6	25.2	13.2
OLMoASR-small.en	3.6	7.4	10.2	11.5	10.1	14.0	23.4	11.5
Whisper small.en	4.6	6.0	9.4	12.0	10.8	14.0	21.9	11.2
OLMoASR-medium.en	3.3	6.9	9.4	12.5	9.5	13.5	21.9	11.0
Whisper medium.en	3.6	5.2	8.9	11.9	10.2	13.3	20.6	10.5
OLMoASR-large.en	3.5	8.8	10.0	11.5	9.9	13.5	22.4	11.4
OLMoASR-large.en-v2	3.6	10.0	10.1	11.1	9.8	13.5	22.1	11.5
Whisper large-v1	3.8	5.3	8.8	11.0	10.3	13.4	20.4	10.4
Whisper large-v2	3.5	5.1	8.8	11.3	9.7	12.6	19.6	10.1
Whisper large-v3	3.2	5.2	8.3	10.2	9.4	12.8	19.4	9.8
Whisper large-v3-turbo	3.1	5.2	8.4	9.5	9.5	12.6	19.3	9.7

Table 5: Long-form English transcription WER (%) with beam search and temperature fallback, comparing between OLMoASR and Whisper models.

Figure 4 shows that although zero-shot OLMoASR models have higher WER on LibriSpeech test-clean than supervised models, they significantly outperform them on the out-of-distribution benchmarks.

Relative robustness. Effective robustness on its own is insufficient to characterize the robustness of a model. Hence, we use relative robustness to directly measure the performance difference between a model with and without a robustness intervention. From Figure 4, we can examine the relative robustness OLMoASR has compared to supervised LibriSpeech models which illustrates that OLMoASR out-performs the other models on the out-of-distribution datasets.

OLMoASR-MIX is a useful robustness intervention. From analyzing OLMoASR’s effective and relative robustness, OLMoASR exhibits positive effective and relative robustness, making OLMoASR-MIX and the curation methodology to extract it a beneficial robustness solution.

5 ABLATIONS

5.1 DATASET SCALING

For our main experiments, we trained on 440K hours, but OLMoASR-MIX contains 1M hours. To examine the effect of data scaling, we trained OLMoASR-74M on subsampled portions of OLMoASR-MIX: 4.8%, 21.1%, 42.1%, 65.4%, 84.5% and 100% (about 50K, 220K, 440K, 680K, 880K and 1M hours), keeping total seen data and hyperparameters constant.

Figure 5 shows that for short-form speech recognition, WER drops by 0.9% when increasing data from 50K to 220K hours ($4\times$), but plateaus from 21.1% to 84.5%. Using the full dataset yields an additional 1.5% WER improvement. For long-form, OLMoASR shows minimal gains across scales. This suggests that beyond moderate scaling ($1.5\times$), gains diminish: a $20\times$ scale-up gives only a 2.1% WER boost for short-form and 0.6% for long-form. This may be due to OLMoASR-74M being too small, the need for more data curation, longer training, or larger models.

Our work has shown that training on OLMoASR-MIX leads to strong robustness and zero-shot capabilities. Can we also quantify the performance and robustness gap between OLMoASR and other models trained on a different data mix? To address this question, we evaluate OLMoASR that has been trained on an academic dataset mix and OLMoASR that has been trained on a dataset mix containing manual and automatic transcripts.

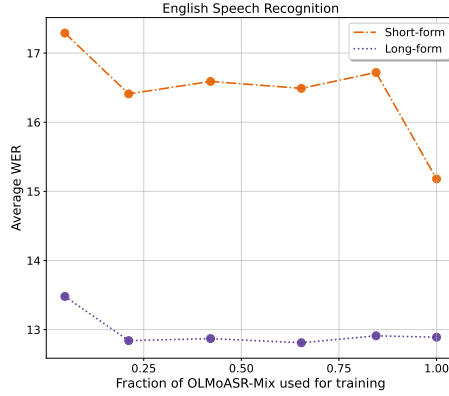


Figure 5: We plot the average performance of OLMoASR-74M on 14 short-form and 7 long-form evaluation sets, while varying the total data trained on. The fraction of OLMoASR-Mix used for training is based on number of hours.

5.2 RESULTS FROM TRAINING ON ACADEMIC DATASETS

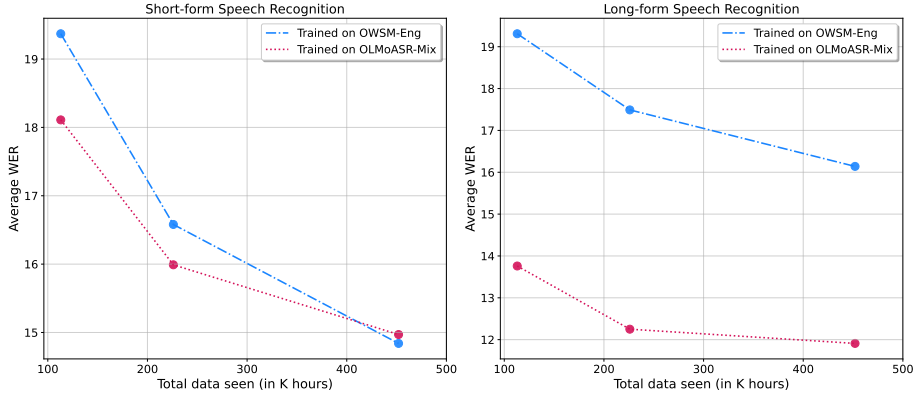


Figure 6: Short-form speech recognition performance of OLMoASR-244M trained on OWSM-Eng and OLMoASR-Mix for varying total amount of data seen. The baseline trained on OLMoASR-Mix trains for one epoch on subsampled subsets of the data, while OLMoASR-244M trained on OWSM-Eng does one, two and four passes through its 113K English subset.

To compare training on academic vs. web-scale data, we train OLMoASR-small.en on OWSM-Eng (the English subset of OWSM) and OLMoASR-Mix with the same total data seen: 113K, 226K, and 452K hours and evaluate both on short and long-form speech recognition.

Figure 6 shows that OLMoASR-Mix consistently yields lower WER for short-form speech except at 452K hours, where the gap is only 0.2%. The difference is more pronounced for long-form, highlighting better generalization from unsegmented long-form data in OLMoASR-POOL versus short-form academic corpora. The smaller short-form gap is partly because OWSM-Eng includes training splits of some evaluation sets, making its evaluation not fully zero-shot.

Since OWSM-Eng overlaps with many test sets except CHIME-6 and CORAAL, we assess out-of-distribution robustness using them and LibriSpeech as a reference. Figure 7 shows that OLMoASR-Mix-trained models achieve lower WER than OWSM-Eng-trained ones, outperforming expected baselines on CHIME-6 and CORAAL.

Overall, training on OLMoASR-Mix improves performance and robustness over academic data at the same scale.

5.3 RESULTS FROM TRAINING ON MANUAL AND AUTOMATIC DATA MIX

YODAS (Li et al., 2023) is a large-scale, multilingual speech dataset containing over 500K hours of YouTube audio across 100+ languages designed to support supervised and self-supervised learning. For our ablation, we train OLMoASR with the 190K hours English subset of YODAS on model scales ranging from tiny to small. The models are trained for the same amount of total data seen as OLMoASR on the full OLMoASR-Mix.

Figure 3 demonstrates that while YODAS is also a web-scale dataset, OLMoASR trained on OLMoASR-Mix out-performs the YODAS-trained model on all model scales with the largest difference being 2.7%.

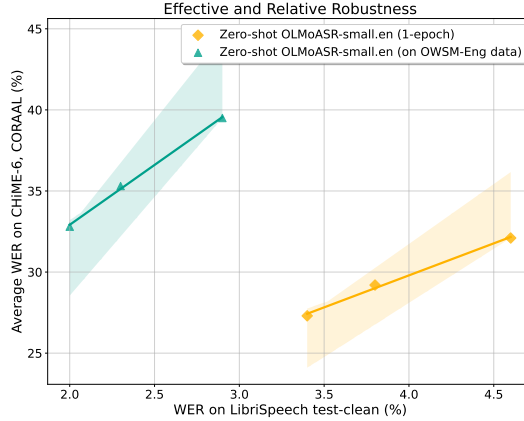


Figure 7: We plot 3 OLMoASR trained on OWSM-Eng data without any robustness interventions and demonstrate their WER on a reference test set (LibriSpeech test-clean) and the average WER across 2 out-of-distribution evaluation sets (CHiME-6, CORAAL). We also plot the performance of zero-shot OLMoASR models to compare to the former.

6 RELATED WORK

Large-scale English ASR Datasets English ASR datasets have grown dramatically in scale. LibriSpeech (Panayotov et al., 2015) remains a benchmark with 960 hours of read speech. GigaSpeech (Chen et al., 2021) expands this to 10,000 hours after filtering from 33,000 hours. The People’s Speech (Galvez et al., 2021) offers 30,000 hours from internet sources (excluding YouTube). Proprietary datasets like those for Whisper (Radford et al., 2023) and USM (Zhang et al., 2023) are much larger, ranging from 100K to 1M hours. YODAS (Li et al., 2023) helps close this gap by providing 190,000 hours of English audio within a 480,000-hour multilingual YouTube corpus.

Large-scale English ASR Models ASR performance benefits from more and better data (Baevski et al., 2020; Radford et al., 2023). Self-supervised learning (SSL) uses large unlabeled audio for pre-training, then fine-tunes on transcripts (Zhang et al., 2022; 2023; Communication et al., 2023), but fine-tuning may limit robustness (Radford et al., 2023). Supervised training on diverse data enhances generalization (Chan et al., 2021; Likhomanenko et al., 2021). Whisper, trained on 680K hours, is well-known but proprietary. OWSM (Peng et al., 2023; 2024b; Tian et al., 2024) provides an open alternative with up to 180K hours (73K English). OWLS (Chen et al., 2025) further explores scaling laws for multilingual ASR up to 360K hours and shows clear benefits from scaling data and model size, especially for non-English.

Data Quality and Data-centric Learning Recent work across language, vision, and multimodal domains shows that better data can greatly boost model performance. Llama 2 and 3 (Touvron et al., 2023b; Grattafiori et al., 2024) improved mainly through better data over Llama 1 (Touvron et al., 2023a). Similar trends hold for text (DCLM (Li et al., 2024), Nemotron-CC (Su et al., 2024)) and multimodal models (DataComp (Gadre et al., 2023), DeepSeek-VL2 (Wu et al., 2024), Bunny (He

et al., 2024)). A key principle in data-centric ML is to run *controlled experiments* with fixed architectures and training, varying only the data to isolate its impact—an approach central to works like DataComp (Gadre et al., 2023; Li et al., 2024).

REFERENCES

- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. Speechstew: Simply mix all available speech recognition data to train one large neural network, 2021. URL <https://arxiv.org/abs/2104.02133>.
- Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, pp. 3670–3674, 2021. doi: 10.21437/Interspeech.2021-1965.
- William Chen, Jinchuan Tian, Yifan Peng, Brian Yan, Chao-Han Huck Yang, and Shinji Watanabe. Owls: Scaling laws for multilingual speech recognition and translation models, 2025. URL <https://arxiv.org/abs/2502.10373>.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinеш Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changan Wang, Jeff Wang, and Skyler Wang. Seamlessm4t: Massively multilingual & multimodal machine translation, 2023. URL <https://arxiv.org/abs/2308.11596>.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. URL <https://arxiv.org/abs/2205.14135>.
- Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. Scaling laws for multilingual neural machine translation, 2023. URL <https://arxiv.org/abs/2302.09650>.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datacomp: In

- search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.
- Daniel Galvez, Greg Diamos, Juan Torres, Keith Achorn, Juan Cerón, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. The people’s speech: A large-scale diverse english speech recognition dataset for commercial usage. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/202cb962ac59075b964b07152d234b70-Paper-round1.pdf.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL <https://arxiv.org/abs/2101.00027>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalk, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Se-woong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kol- lar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2025. URL <https://arxiv.org/abs/2406.11794>.
- 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, 2023. doi: 10.1109/ASRU57964.2023.10389689.
- Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. Rethinking evaluation in asr: Are our models robust enough? In *Interspeech 2021*, pp. 311–315, 2021. doi: 10.21437/Interspeech.2021-1758.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*, 2023.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

- 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, 2015. doi: 10.1109/ICASSP.2015.7178964.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.
- Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024a.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024b. URL <https://arxiv.org/abs/2406.17557>.
- Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee-Weon Jung, Soumi Maiti, and Shinji Watanabe. 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, 2023. doi: 10.1109/ASRU57964.2023.10389676.
- Yifan Peng, Yui Sudo, Muhammad Shakeel, and Shinji Watanabe. Owsm-ctc: An open encoder-only speech foundation model for speech recognition, translation, and language identification, 2024a. URL <https://arxiv.org/abs/2402.12654>.
- Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee weon Jung, and Shinji Watanabe. Owsm v3.1: Better and faster open whisper-style speech models based on e-branchformer. In *Interspeech 2024*, pp. 352–356, 2024b. doi: 10.21437/Interspeech.2024-1194.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. Pmlr, 2021.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/radford23a.html>.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL <https://arxiv.org/abs/1910.10683>.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxu Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research, 2024. URL <https://arxiv.org/abs/2402.00159>.
- Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeibi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. *arXiv preprint arXiv:2412.02595*, 2024.

- Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset, 2025. URL <https://arxiv.org/abs/2412.02595>.
- Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification, 2020. URL <https://arxiv.org/abs/2007.00644>.
- Jinchuan Tian, Yifan Peng, William Chen, Kwanghee Choi, Karen Livescu, and Shinji Watanabe. On the effects of heterogeneous data sources on speech-to-text foundation models. In *Interspeech 2024*, pp. 3959–3963, 2024. doi: 10.21437/Interspeech.2024-1938.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- Jörgen Valk and Tanel Alumäe. Voxlingua107: A dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 652–658, 2021. doi: 10.1109/SLT48900.2021.9383459.
- Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models, 2024. URL <https://arxiv.org/abs/2411.12372>.
- Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341*, 2025.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.
- Yu Zhang, Daniel S. Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, Zongwei Zhou, Bo Li, Min Ma, William Chan, Jiahui Yu, Yongqiang Wang, Liangliang Cao, Khe Chai Sim, Bhuvana Ramabhadran, Tara N. Sainath, Françoise Beaufays, Zhifeng Chen, Quoc V. Le, Chung-Cheng Chiu, Ruoming Pang, and Yonghui Wu. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1519–1532, 2022. doi: 10.1109/JSTSP.2022.3182537.
- Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. Google usm: Scaling automatic speech recognition beyond 100 languages, 2023. URL <https://arxiv.org/abs/2303.01037>.

A TRAINING DETAILS

Table 6 displays hyperparameters used to train all models. We trained OLMoASR-tiny.en, OLMoASR-base.en and OLMoASR-small.en on 1 H100 node and OLMoASR-medium.en, OLMoASR-large.en, and OLMoASR-large.en-v2 on 2 and 4 H100 nodes respectively.

Hyperparameter	Value
Updates	524288
Batch Size	512
Warmup Updates	1049
Max grad norm	1.0
Optimizer	AdamW
β_1	0.9
β_2	0.98
ϵ	10^{-6}
Weight Decay	0.1
Weight Init	Gaussian Fan-In
Maximum Learning Rate	1.5×10^{-3}
Learning Rate Schedule	Linear Decay

Table 6: Training hyperparameters.

B MODEL SIZES

Table 7 enumerates all the model sizes OLMoASR has and the associated parameter count.

Size	Parameters
tiny	39 M
base	74 M
small	244 M
medium	769 M
large	1550 M
large-v2	1550 M

Table 7: Model sizes and their parameter counts

C DEDUPLICATION AND DECONTAMINATION

We performed transcript level fuzzy deduplication using minhash. We used the parameters from FineWeb (Penedo et al., 2024b), where we used 5-grams of tokens and computed 112 hash functions, split into 14 buckets of 8 hashes each. If any pair of transcripts has the same 8 hashes in any one bucket, they are marked as duplicates. This procedure targets documents that have a Jaccard similarity of 75%. We performed this on 17M total transcripts and removed 505K transcripts for a total deduplication removal rate of 3%. Decontamination was performed by a simple n-gram search. In particular, we decontaminate the evaluation datasets of TED-LIUM3 against our training corpus. First we collect all n-grams of size 10 from the evaluation dataset and check for their presence in each training dataset transcript. If any n-gram is present, we mark the training document as contaminated and do not include it in our training sets. We apply this procedure to 17M transcripts and find only 286 contaminated transcripts.

D OWSM vs. OLMoASR PERFORMANCE TABLE

Table 8 shows the WER performance on short-form evaluation sets that OLMoASR, Whisper and OWSM have all been evaluated on.

E OLMoASR vs. OTHER OPEN-SOURCE MODELS

Table 9 illustrates the WER performance of OLMoASR relative to other open-source models.

Model	LibriSpeech.test-clean	LibriSpeech.test-other	TED-LIUM3	WSJ	Switchboard	VoxPopuli.en	Fleurs.en.us	Average
OLMoASR-base.en	3.7	9.0	4.6	4.3	14.0	9.7	6.7	7.4
Whisper base.en	4.2	10.2	4.9	4.6	15.2	10.0	7.6	8.1
OWSM-v3.1 base	3.6	9.1	7.8	5.3	22.9	12.0	14.8	10.1
OLMoASR-small.en	3.0	7.0	4.2	3.8	13.2	8.7	5.0	6.4
Whisper small.en	3.1	7.4	4.0	3.3	15.7	8.1	6.0	6.8
OWSM-v3.1 small	2.5	5.8	5.0	3.8	17.4	9.1	10.3	7.3
OWSM-v3.2 small	2.5	6.2	5.4	4.0	17.4	9.0	10.1	7.4
OLMoASR-medium.en	3.5	5.7	5.0	3.6	12.7	8.4	4.4	6.2
Whisper medium.en	3.1	6.3	4.1	3.3	14.1	7.4	5.0	6.2
OWSM-v3.1 medium	2.4	5.0	5.1	3.5	16.3	8.4	9.0	6.8
OWSM-CTC medium	2.4	5.2	4.9	4.2	16.9	8.6	9.9	7.0

Table 8: Short-form English transcription WER (%) with greedy decoding, comparing between OLMoASR, Whisper and OWSM models.

F CONTRIBUTIONS

- **Huong Ngo:** Programed and designed all main code infrastructures (data collection and processing, training, and evaluation), collected and processed all data, planned and executed all experiments on Ai2 compute cluster and UW Hyak Supercomputer Cluster, coordinated and performed evaluation and paper writing and revision.
- **Matt Deitke:** Provided advice on data collection and processing and training, paper writing and revision.
- **Martijn Bartelds:** Provided advice on training and evaluation, paper writing and revision.
- **Sarah Pratt:** Provided visual graphics support for paper.
- **Josh Gardner:** Provided advice on data collection and processing, model training, experiment design and project direction, paper writing and revision.
- **Matt Jordan:** Performed deduplication and decontamination, provided advice on data processing, experiment design and project direction, paper writing and revision.
- **Ludwig Schmidt:** Provided advice on data collection and processing, model training, experiment design and project direction, paper writing and revision.

Model	LibriSpeech test-clean	LibriSpeech test-other	TED-LIUM3	WSJ	CallHome	Switchboard	Common Voice 5.1	Arctic	CORAAL	CHIME6	AMI-1HM	AMI-SDM	VoxPopuli.en	Fleurs.en.us	Average
OLMoASR (Open weights, code, data) vs. Whisper (Open weights, closed training code, data)															
OLMoASR-tiny.en	5.1	12.3	5.5	5.6	23.9	18.7	25.1	19.3	25.7	45.2	24.2	55.4	11.6	9.7	20.5
Whisper tiny.en	5.6	14.6	6.0	5.0	24.1	17.8	26.3	20.0	23.9	41.3	23.7	50.3	11.7	11.6	20.1
OLMoASR-base.en	3.7	9.0	4.6	4.3	20.5	14.0	18.5	13.6	21.5	38.0	20.4	47.8	9.7	6.7	16.6
Whisper base.en	4.2	10.2	4.9	4.6	20.9	15.2	19.0	13.4	22.6	36.4	20.5	46.7	10.0	7.6	16.9
OLMoASR-small.en	3.0	7.0	4.2	3.8	16.7	13.2	13.1	9.6	19.6	30.6	18.7	39.9	8.7	5.0	13.8
Whisper small.en	3.1	7.4	4.0	3.3	18.2	15.7	13.1	9.7	20.2	27.6	17.5	38.0	8.1	6.0	13.7
OLMoASR-medium.en	3.5	5.7	5.0	3.6	14.3	12.7	11.3	7.5	18.7	28.5	16.9	38.3	8.4	4.4	12.8
Whisper medium.en	3.1	6.3	4.1	3.3	16.2	14.1	10.6	7.6	17.5	25.3	16.4	37.2	7.4	5.0	12.4
OLMoASR-large.en	2.6	5.9	4.5	3.7	16.5	12.7	11.1	7.9	18.7	30.7	16.4	38.8	8.1	4.5	13.0
OLMoASR-large.en-v2	2.7	5.6	4.2	3.6	15.0	11.7	11.1	7.8	18.1	29.4	17.1	38.0	8.0	4.2	12.6
Whisper large-v1	2.7	5.6	4.0	3.1	15.8	13.1	9.5	6.7	19.4	25.6	16.4	36.9	7.3	4.6	12.2
Whisper large-v2	2.7	5.2	4.0	3.9	17.6	13.8	9.0	6.2	16.2	25.5	16.9	36.4	7.3	4.4	12.1
Whisper large-v3	2.0	3.9	3.9	3.5	14.0	13.2	8.4	5.9	18.7	26.8	16.0	34.2	9.5	4.0	11.7
Whisper large-v3-turbo	2.2	4.2	3.5	3.5	13.2	12.9	9.7	6.3	18.6	27.3	16.1	35.2	12.2	4.4	12.1
wav2vec2-base-100h	6.0	13.4	17.8	13.9	46.9	40.2	47.4	40.8	47.0	79.9	48.1	81.2	28.9	23.1	38.2
wav2vec2-base-960h	3.3	8.5	12.8	8.9	40.6	32.9	36.4	30.9	39.9	68.5	40.2	71.9	21.4	17.4	31.0
wav2vec2-large-960h-lv60-self	1.8	3.8	7.4	4.4	29.1	22.2	19.9	15.8	29.2	56.3	30.8	57.0	13.0	10.2	21.5
wav2vec2-large-960h	2.7	6.2	10.5	7.7	34.8	28.3	29.9	24.5	35.6	65.8	37.0	67.6	17.9	14.6	27.4
wav2vec2-large-robust-ft-libri-960h	2.6	5.3	9.2	6.1	23.4	19.8	20.3	16.2	29.4	58.1	31.7	61.6	15.1	11.8	22.2
asr-crdnn-rnnlm-librispeech	3.0	9.7	17.7	10.7	59.7	56.1	43.7	33.3	83.8	81.0	57.2	85.8	30.6	32.4	43.2
asr-transformer-transformerlm-librispeech	2.1	5.4	11.9	7.4	38.9	33.0	30.6	23.5	44.9	79.5	44.5	75.4	17.8	17.0	30.9
hubert-large-ls960-ft	2.0	4.1	8.4	5.4	29.6	22.8	20.8	16.0	32.0	60.0	33.7	59.1	14.4	10.9	22.8
hubert-xl-large-ls960-ft	1.9	3.5	8.3	5.4	29.3	22.2	19.8	14.8	31.5	58.5	33.3	58.9	14.2	10.5	22.3
s2t-large-librispeech-asr	3.3	8.1	14.9	9.4	54.5	40.3	38.1	30.7	50.2	79.2	53.4	79.5	21.6	18.0	35.8
s2t-medium-librispeech-asr	3.6	8.2	15.7	9.7	58.1	42.4	39.3	31.3	52.6	79.8	60.3	85.3	22.9	19.7	37.8
stt.en.conformer.ctc.large	2.1	4.2	4.4	2.1	11.3	8.2	7.4	4.0	13.5	30.5	15.9	39.9	6.7	8.2	11.3
stt.en.conformer.transducer.xlarge	1.5	2.8	4.3	1.2	12.0	7.4	4.3	1.5	19.9	36.8	20.5	48.6	6.0	6.3	12.4
unispeech-sat-base-100h-libri-ft	5.7	13.8	17.7	13.6	46.5	40.0	45.3	38.6	44.7	74.8	47.8	77.7	29.8	22.4	37.0

Table 9: Short-form English transcription WER (%) with greedy decoding, comparing between OLMoASR , Whisper models and other open-source models