

Reproduction of huggingface/open_asr_leaderboard: Whisper large-v3-turbo on the seven English test sets

K-Veritas Team

Submitted by K-Veritas Team

kv:2610.00043v1 · Published 2026-10-21

Independent reproduction · Artifact evaluation · Natural language processing

Abstract

Ran transformers/run_whisper.sh at the pinned commit, unchanged: openai/whisper-large-v3-turbo with batch size 64 on the full test splits of AMI, GigaSpeech, VoxPopuli, Earnings-22, LibriSpeech test-clean and test-other and SPGISpeech from hf-audio/open-asr-leaderboard, reporting WER and RTFx per set, then normalizer/eval_utils.score_results. The dataset and model weights were downloaded before the sealed run, which ran offline. Folders for the other toolkits (nemo_asr, espnet, ctranslate2 and so on) are withheld as hash-only entries since this run does not use them. PyTorch 2.11 (CUDA 12.8), transformers 5.19, datasets 5.1 with torchaudio for audio decoding. One NVIDIA RTX 5060 Ti.

About this record

Title, authors and abstract were provided by the authors and reviewed for completeness, not for scientific soundness. Every number in this record is copied from the signed data of the sealed reports [2], which remain authoritative.

This cover page is not signed. Check it against the SHA-256 listed at <https://kveritas.org/records/2610.00043v1>, and verify each report with its code bundle: `kveritas verify <report.pdf> --bundle <bundle.zip>`.

1 Sealed reports

The experiments were sealed with K-Veritas [1], which binds each result to the code, hardware and time that produced it.

Report	Label	Sealed	Runs	Invocations	Signer
r1 [2]	Whisper large-v3-turbo: 7 Eng...	2026-10-11	1	1	Signed by the K-Veritas server

2 Results

r1: Whisper large-v3-turbo: 7 English ASR test sets

Run	Metric	Final value	Step
1	ami_cleaned_test_wer	14.31	
1	ami_cleaned_test_rtfx	56.9	
1	gigaspeech_cleaned_test_wer	9.21	
1	gigaspeech_cleaned_test_rtfx	99.64	
1	voxpathuli_cleaned_aa_test_wer	5.44	
1	voxpathuli_cleaned_aa_test_rtfx	166.15	
1	earnings22_test_wer	11.07	
1	earnings22_test_rtfx	105.52	
1	librispeech_test_clean_wer	1.99	
1	librispeech_test_clean_rtfx	105.38	

Run	Metric	Final value	Step
1	librispeech_test_other_wer	4.26	
1	librispeech_test_other_rtfx	93.43	
1	spgispeech_test_wer	3.07	
1	spgispeech_test_rtfx	127.46	

Run	Inline claim	Value	Phase
1	ami_cleaned_test_wer	14.31	
1	gigaspeech_cleaned_test_wer	9.21	
1	voxpathli_cleaned_aa_test_wer	5.44	
1	earnings22_test_wer	11.07	
1	librispeech_test_clean_wer	1.99	
1	librispeech_test_other_wer	4.26	
1	spgispeech_test_wer	3.07	
1	mean_wer_7_sets	7.05	

3 Setup

Report	Item	Value
r1	Operating system	linux / amd64
r1	CPU	Intel(R) Xeon(R) W-2145 CPU @ 3.70GHz, 16 cores
r1	GPU	NVIDIA GeForce RTX 5060 Ti, 16311 MiB
r1	Memory	31.0 GB
r1	Total run time	6426.3 s

4 Files

Each report and code bundle must hash to these values. A bundle belongs to the report whose signed data names its hash.

File	SHA-256	Size
r1.pdf	3f3514cd578961ae6c2dddb57dd7fc3860b35a759e46c06a343369253280c379	7.0 MB
r1.kvbundle.zip	e821bff26d4be24b05d388c1aa6be3b70ba320326597440f28e9813d63408e19	162 KB

5 Links and licence

Paper: <https://arxiv.org/abs/2510.06961>

Reproduction of:

https://github.com/huggingface/open_asr_leaderboard/tree/60e2e3b897ea6bd50adff236c8dc55569d485cdc

Record: CC-BY 4.0. Code: Apache-2.0, as declared by the authors. Tags: speech-recognition, whisper, leaderboard, wer.

Listed, not endorsed. Review checks the artifact, never the science. Reports are never edited; changing them makes a new version. Other fields are corrected in place, and every correction is logged in metadata.yaml.

References

[1] M. K. Keita and C. Homan. Computer Science Conferences Should Require Nonrepudiable Experimental Results. NeurIPS 2026 Position Paper Track. arXiv:2605.08586.

[2] Sealed experiment report r1 (Whisper large-v3-turbo: 7 English ASR test sets), kv:2610.00043v1/r1. Data hash 68cfae92f6b1c3c20a90ccc7a48432ab90ee8ca0f47051901f37f0970d06205b.

<https://kveritas.org/records/2610.00043v1/r1.pdf>