Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

SECTION 1 - Metric Definitions

Metric	Description	Interpretation	Ideal Range
WER (Word Error Rate)	Percentage of incorrectly predicted words	Measures transcription accuracy	0-100 (Lower = bet
CER (Character Error Rate)	Percentage of incorrectly predicted characters	Captures robustness to accents and noise	0-100 (Lower = bet

Note: WER and CER are primary indicators of ASR accuracy. Lower values indicate fewer misrecognized words and better noise resilience.

Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

SECTION 2 - Data Overview & Methodology

This report evaluates multiple ASR models on four distinct datasets: Helpline (Own), Common Voice (Swahili), FLEURS, and a Domain Test Dataset. Performance is measured using Word Error Rate (WER) and Character Error Rate (CER), calculated without special normalization. The models include both general-purpose systems (e.g., Whisper, MMS) and models fine-tuned on specific domains to assess the impact of specialization.

Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

SECTION 3 - Comparative Performance Tables

Dataset: Helpline Audio

Model	WER ?	CER ?	Rank
openchs/asr-whisper-large-v4	61.03	25.42	1
facebook/seamless-m4t-v2-large	62.20	32.29	2
openchs/asr-whisper-large-v3-helpline	67.66	28.79	3
facebook/mms-1b-all	69.30	29.41	4
openchs/asr-whisper-helpline-sw-v1	69.94	37.26	5
Sunbird/asr-whisper-large-v2-salt	103.68	77.26	6
openai/whisper-large-v3	124.94	74.46	7
openai/whisper-large-v2	226.47	139.65	8

Dataset: Mozilla Common Voice 23.0-Swahili

Model	WER ?	CER ?	Rank
facebook/seamless-m4t-v2-large	25.83	22.03	1
openchs/asr-whisper-helpline-sw-v1	31.87	24.87	2
openchs/asr-whisper-large-v3-helpline	38.06	27.39	3
facebook/mms-1b-all	39.91	24.25	4
openchs/asr-whisper-large-v4	46.17	29.44	5
openai/whisper-large-v3	72.17	38.26	6
Sunbird/asr-whisper-large-v2-salt	94.15	49.78	7
openai/whisper-large-v2	95.05	55.34	8

Dataset: FLEURS

Model	WER ?	CER ?	Rank
facebook/mms-1b-all	15.71	4.12	1
facebook/seamless-m4t-v2-large	24.85	8.81	2
openchs/asr-whisper-large-v3-helpline	25.14	8.19	3
openchs/asr-whisper-helpline-sw-v1	25.52	8.25	4
openchs/asr-whisper-large-v4	28.41	11.59	5
openai/whisper-large-v3	46.13	11.70	6
openai/whisper-large-v2	52.72	13.71	7
Sunbird/asr-whisper-large-v2-salt	87.33	30.04	8

Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

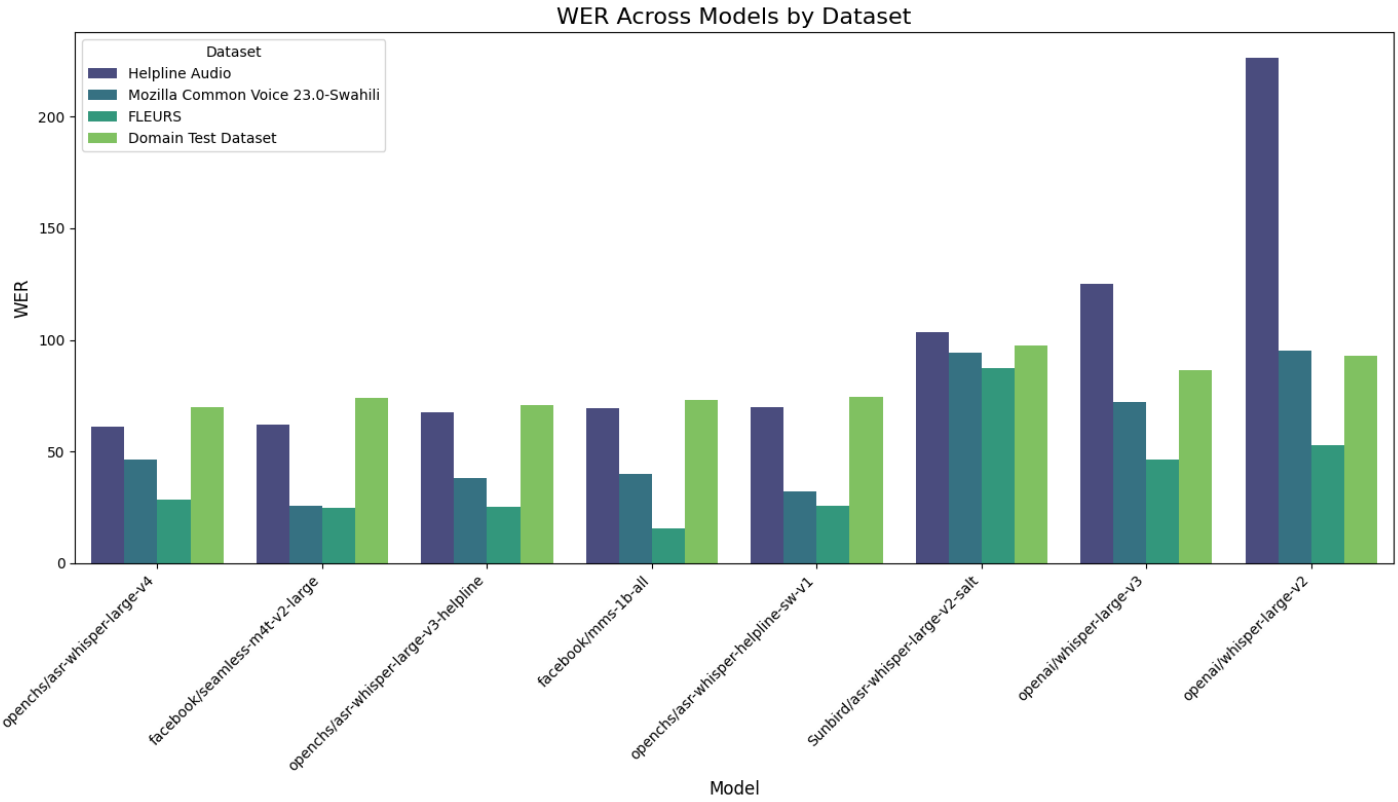
Dataset: Domain Test Dataset

Model	WER ?	CER ?	Rank
openchs/asr-whisper-large-v4	70.00	40.63	1
openchs/asr-whisper-large-v3-helpline	70.87	37.65	2
facebook/mms-1b-all	73.20	41.85	3
facebook/seamless-m4t-v2-large	74.03	43.17	4
openchs/asr-whisper-helpline-sw-v1	74.66	45.01	5
openai/whisper-large-v3	86.53	54.66	6
openai/whisper-large-v2	92.91	75.48	7
Sunbird/asr-whisper-large-v2-salt	97.62	70.40	8

Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

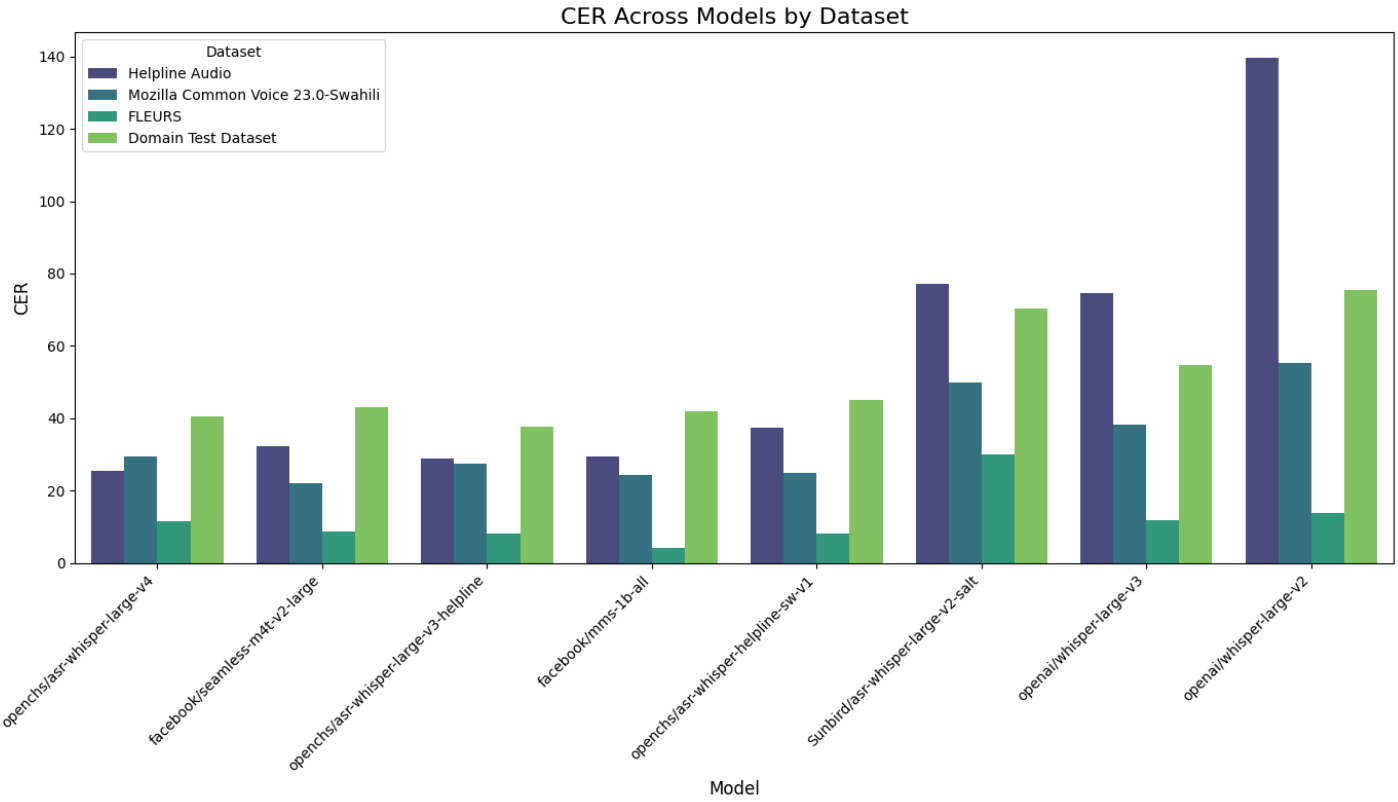
SECTION 4.1 - Metric-Wise Charts: WER



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

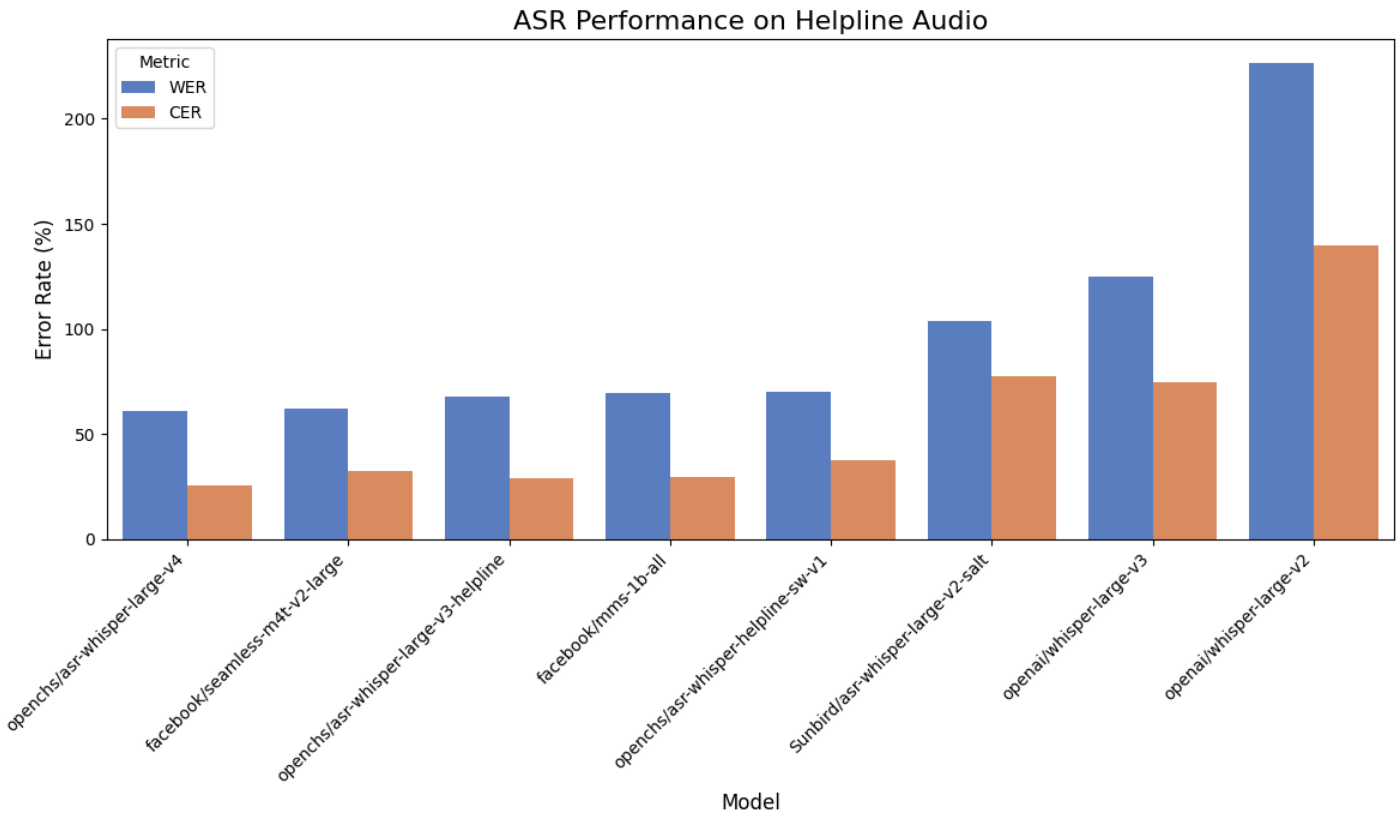
SECTION 4.1 - Metric-Wise Charts: CER



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

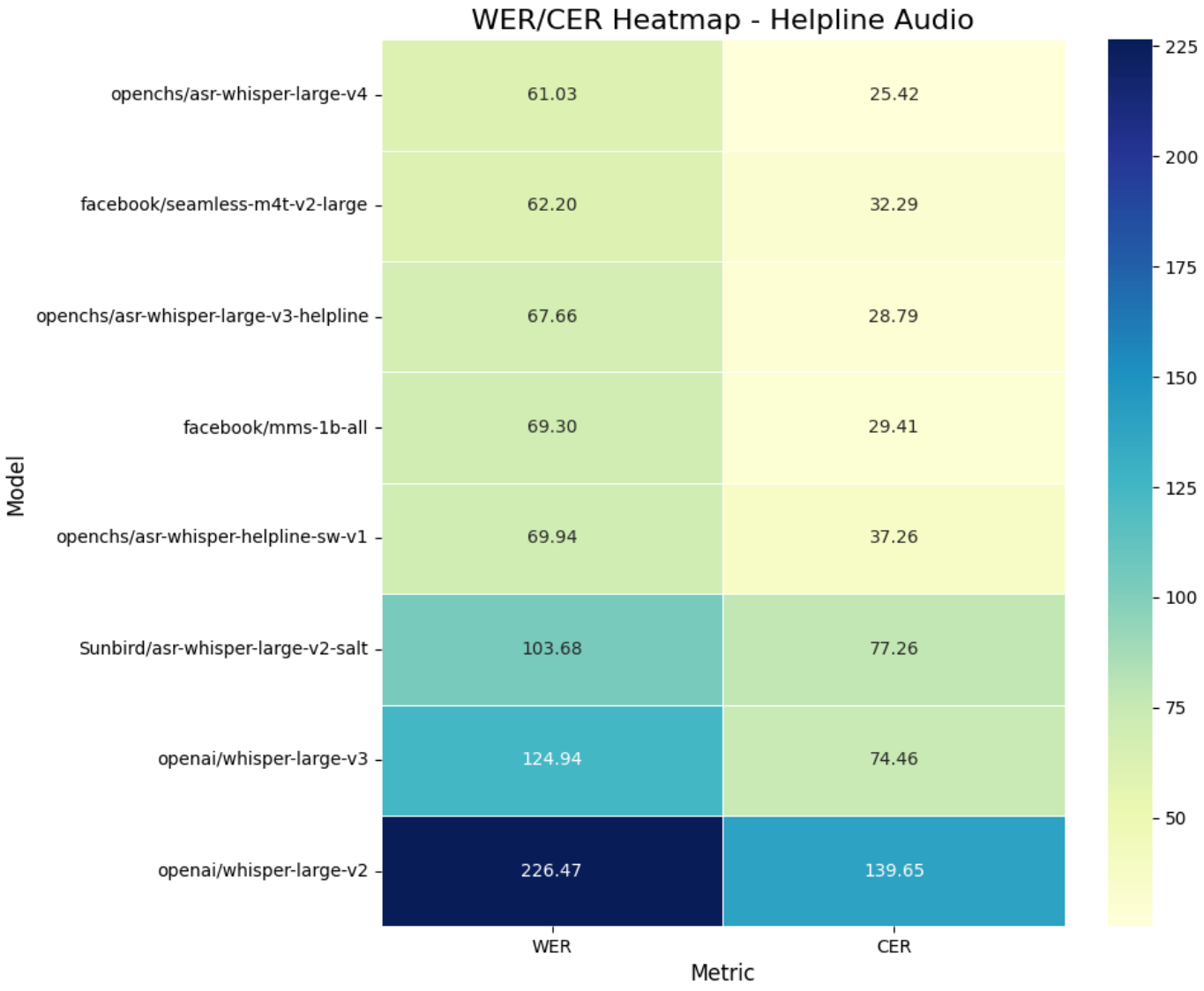
SECTION 4.2 - Dataset-Wise Chart: Helpline Audio



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

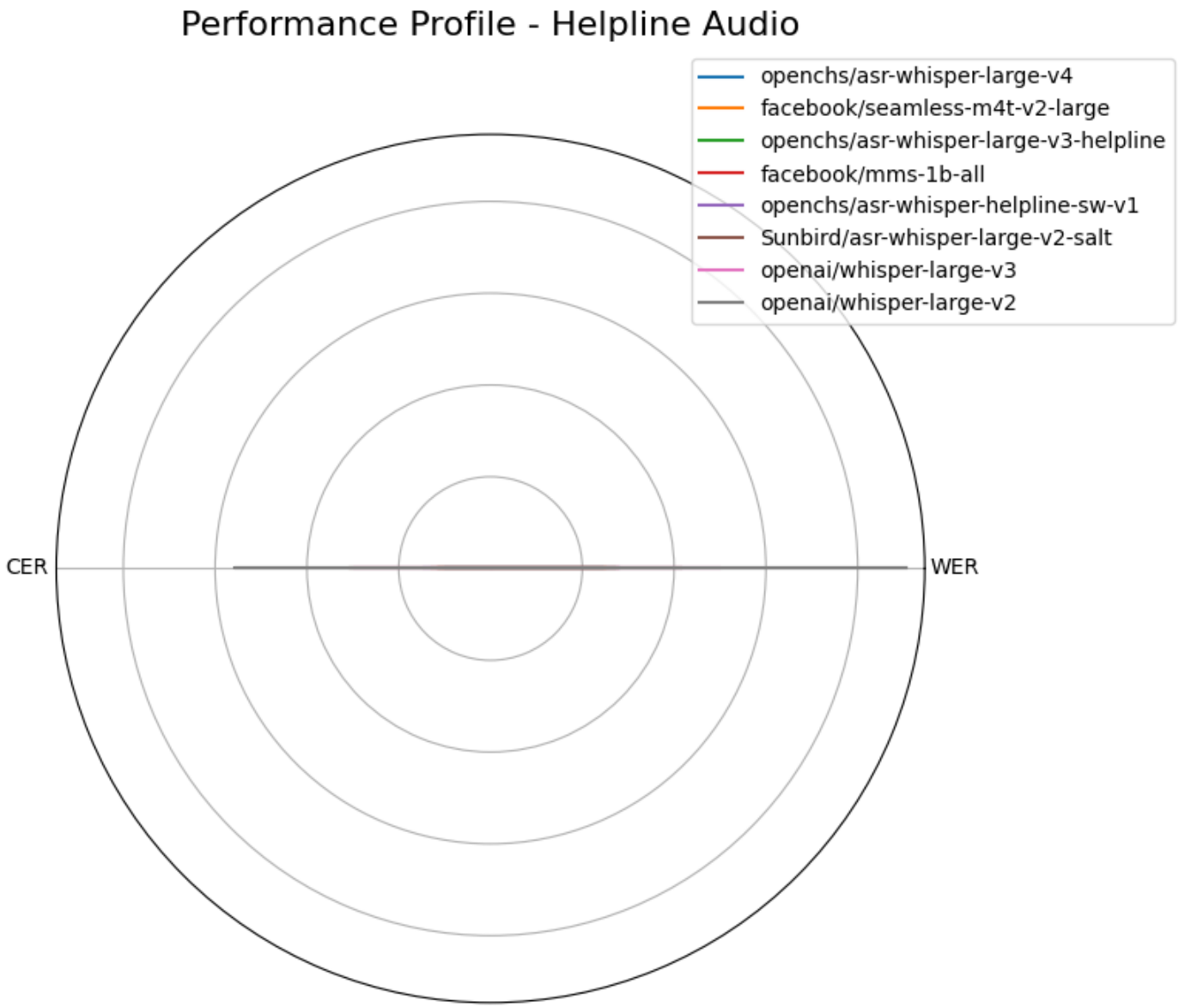
SECTION 4.3 - Dataset-Wise Heatmap: Helpline Audio



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

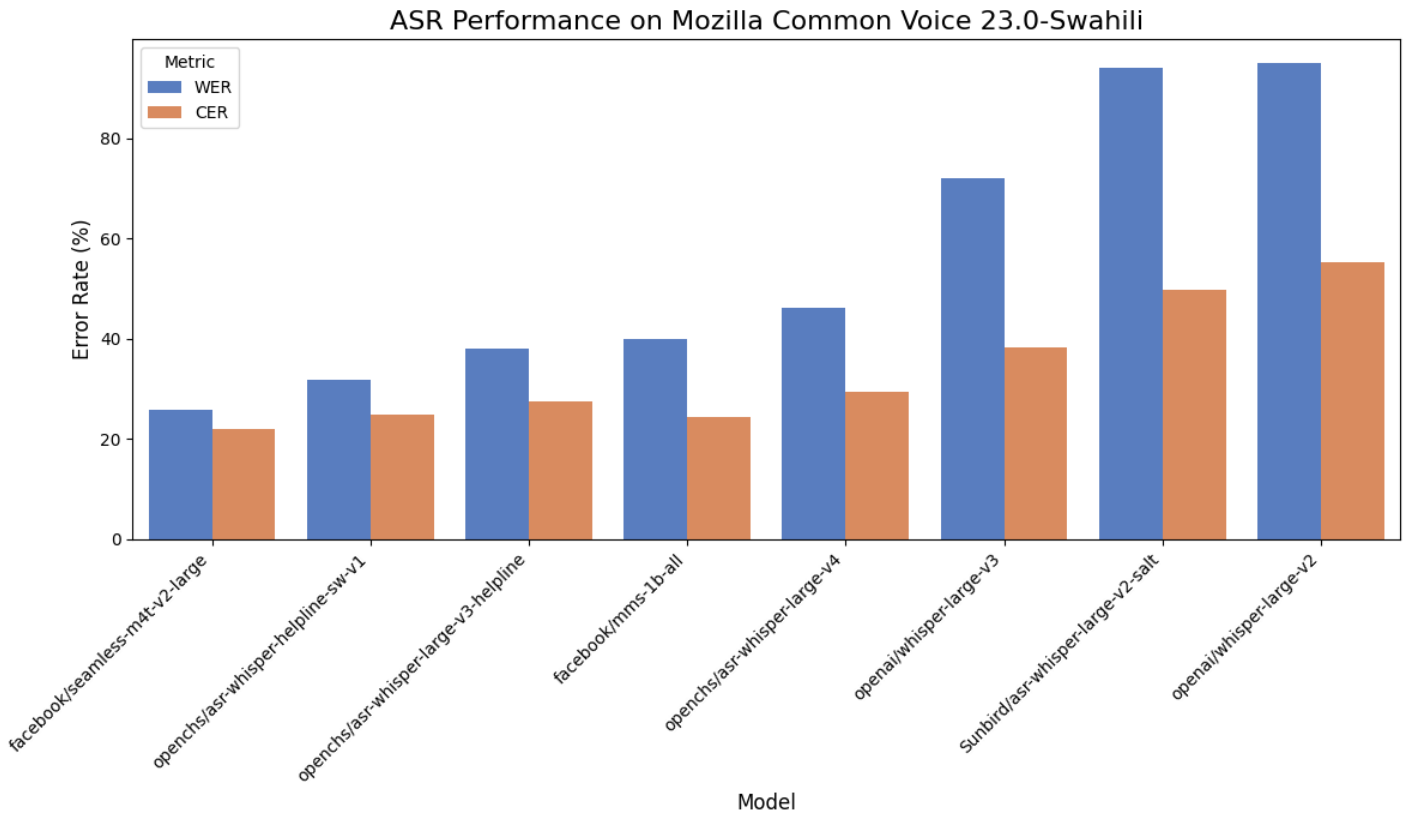
SECTION 4.4 - Radar Chart: Helpline Audio



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

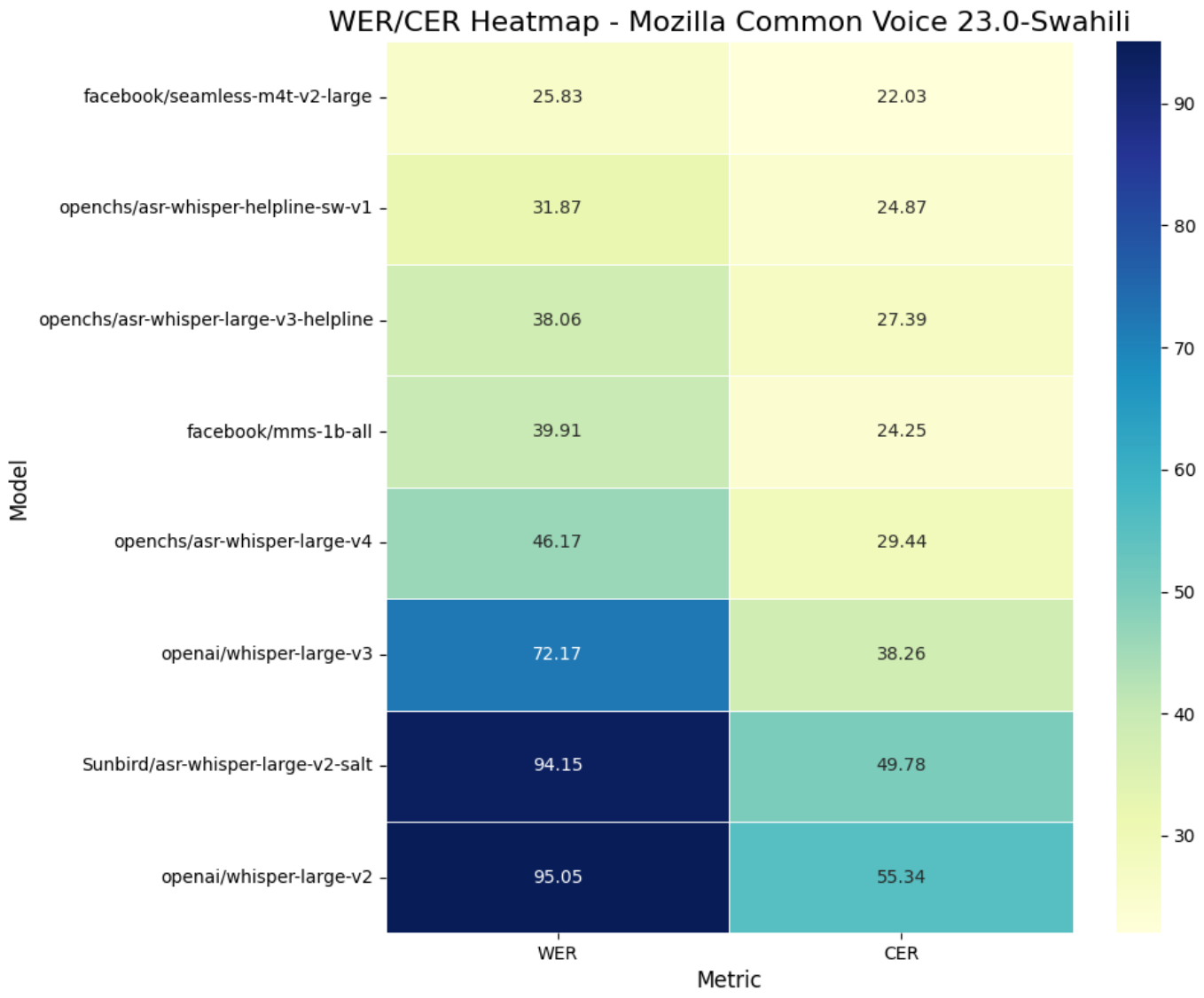
SECTION 4.2 - Dataset-Wise Chart: Mozilla Common Voice 23.0-Swahili



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

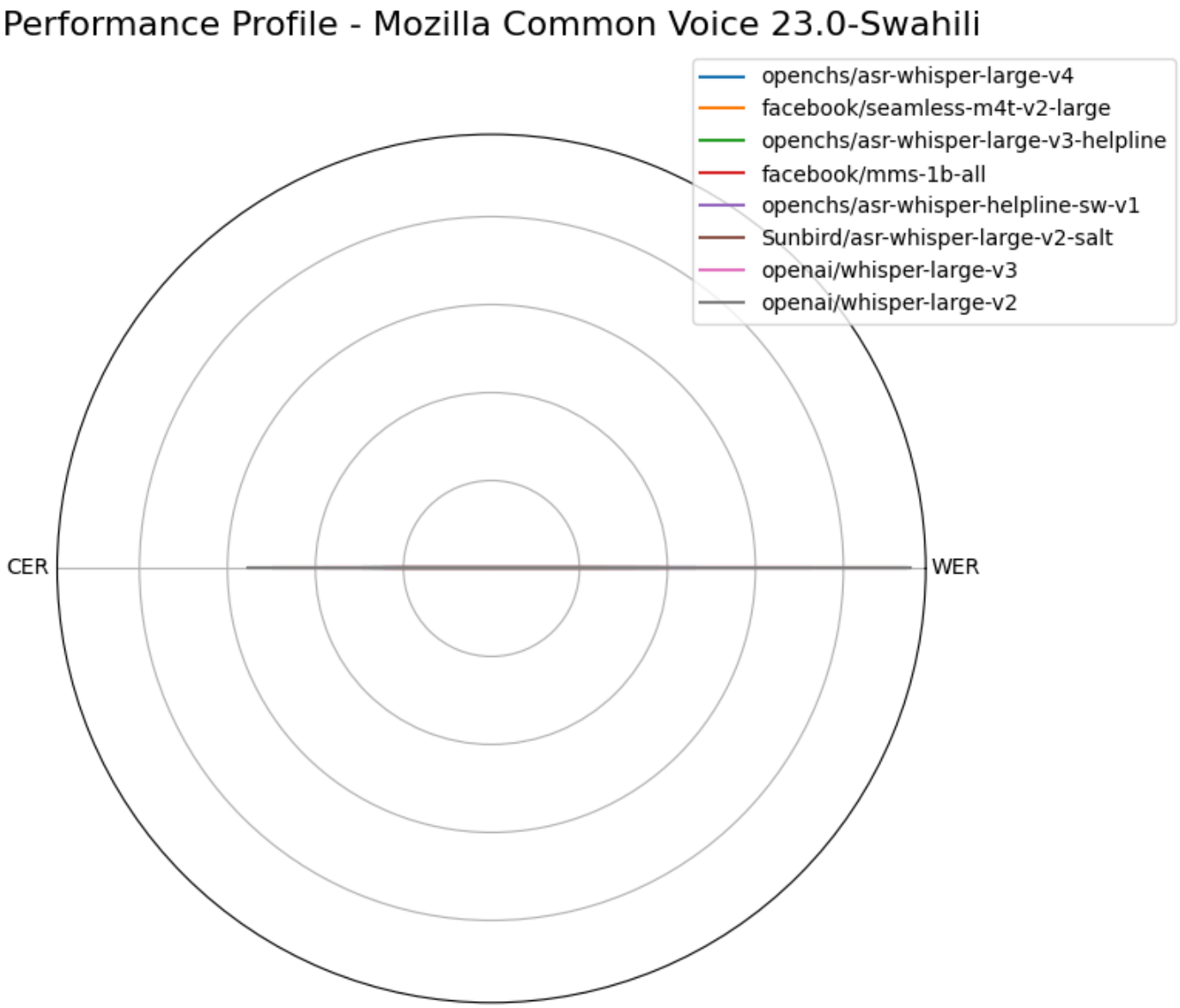
SECTION 4.3 - Dataset-Wise Heatmap: Mozilla Common Voice 23.0-Swahili



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

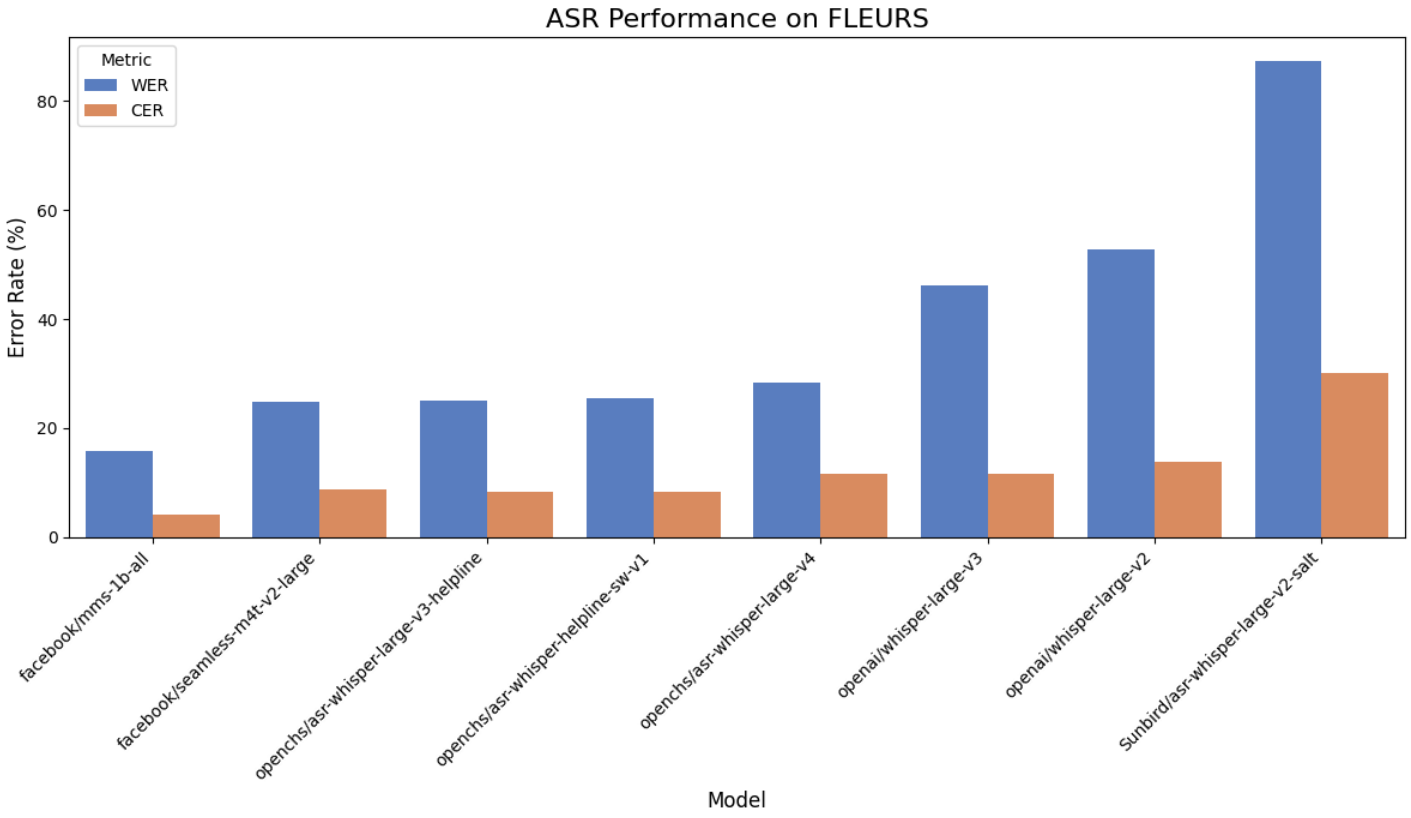
SECTION 4.4 - Radar Chart: Mozilla Common Voice 23.0-Swahili



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

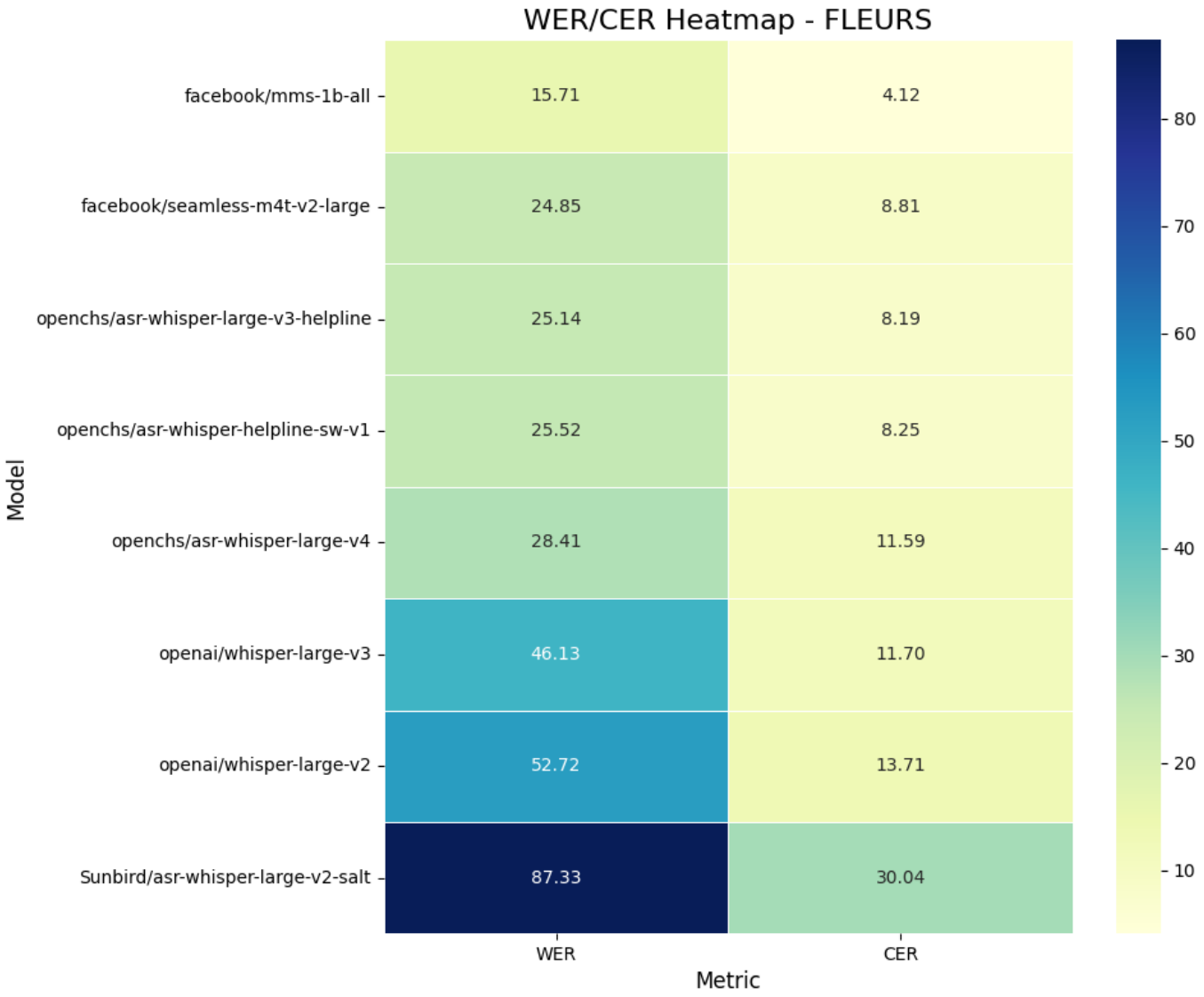
SECTION 4.2 - Dataset-Wise Chart: FLEURS



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

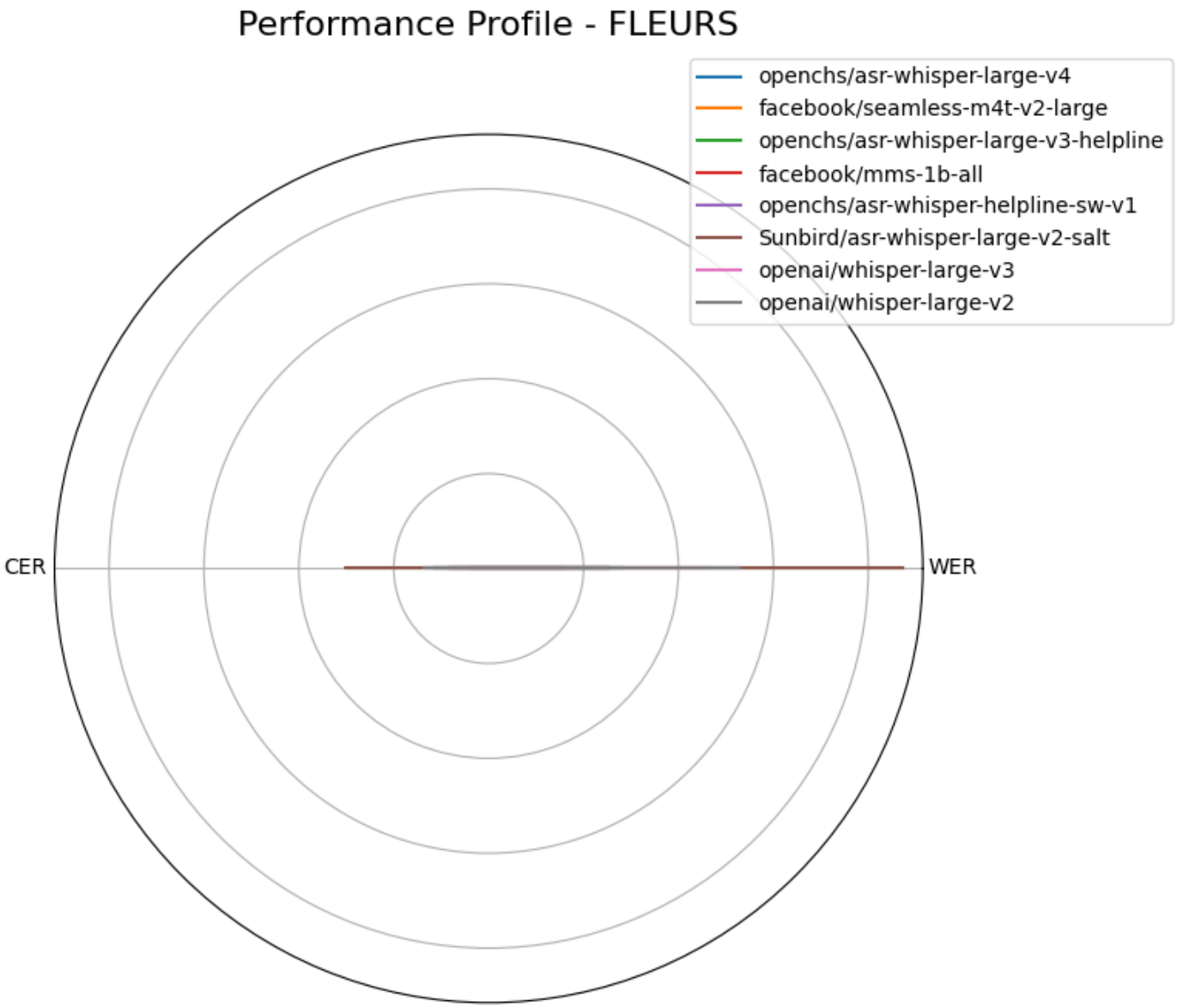
SECTION 4.3 - Dataset-Wise Heatmap: FLEURS



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

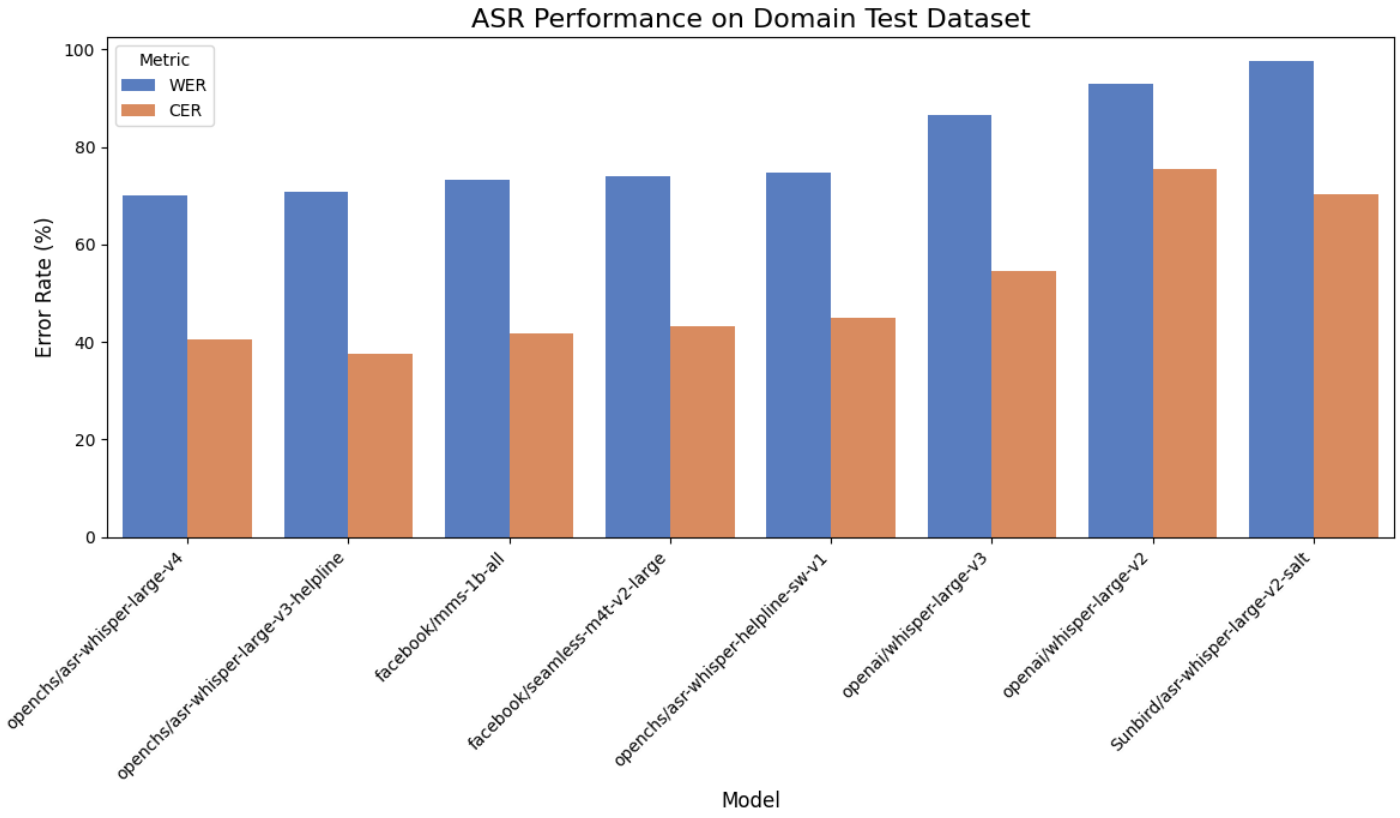
SECTION 4.4 - Radar Chart: FLEURS



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

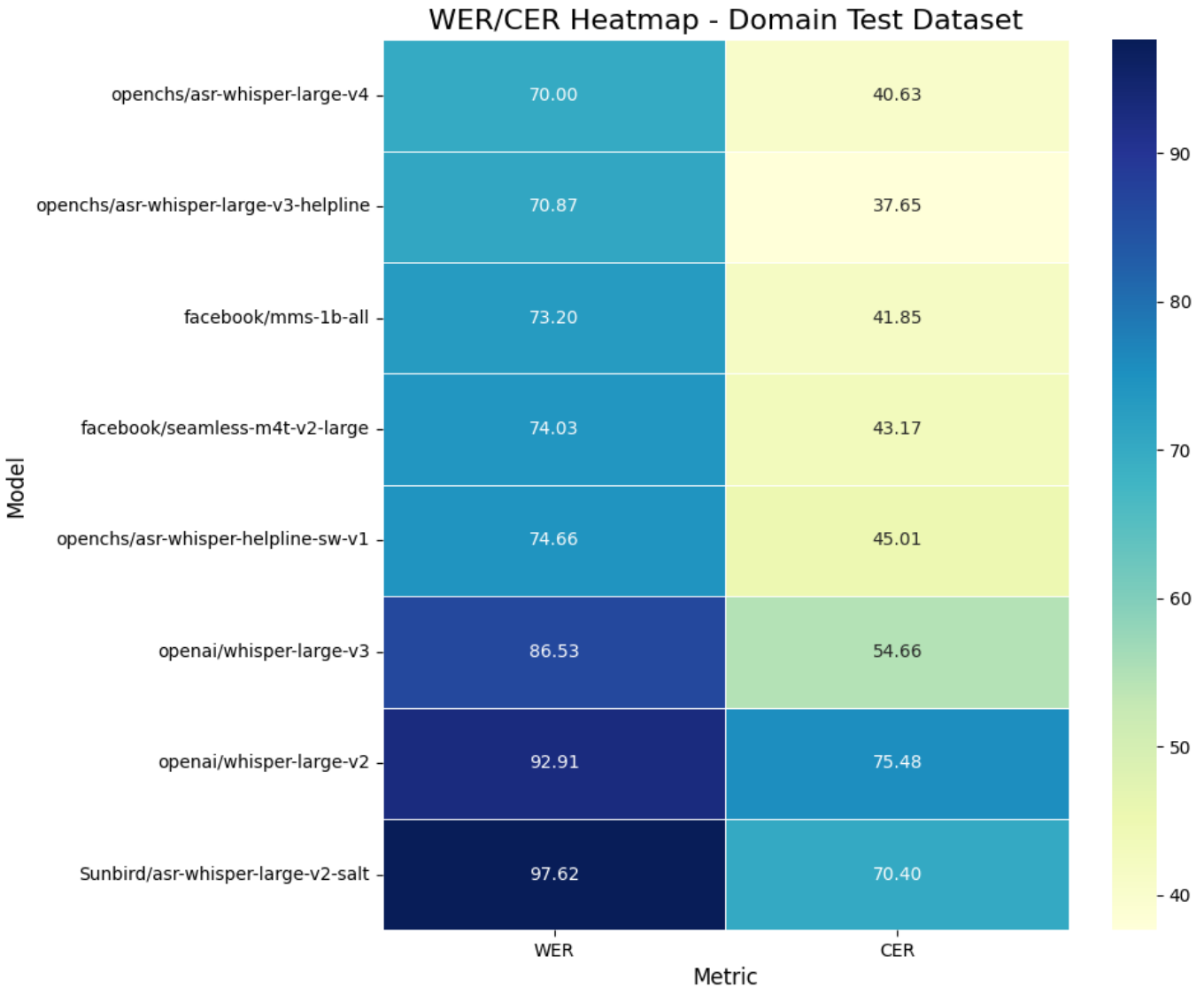
SECTION 4.2 - Dataset-Wise Chart: Domain Test Dataset



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

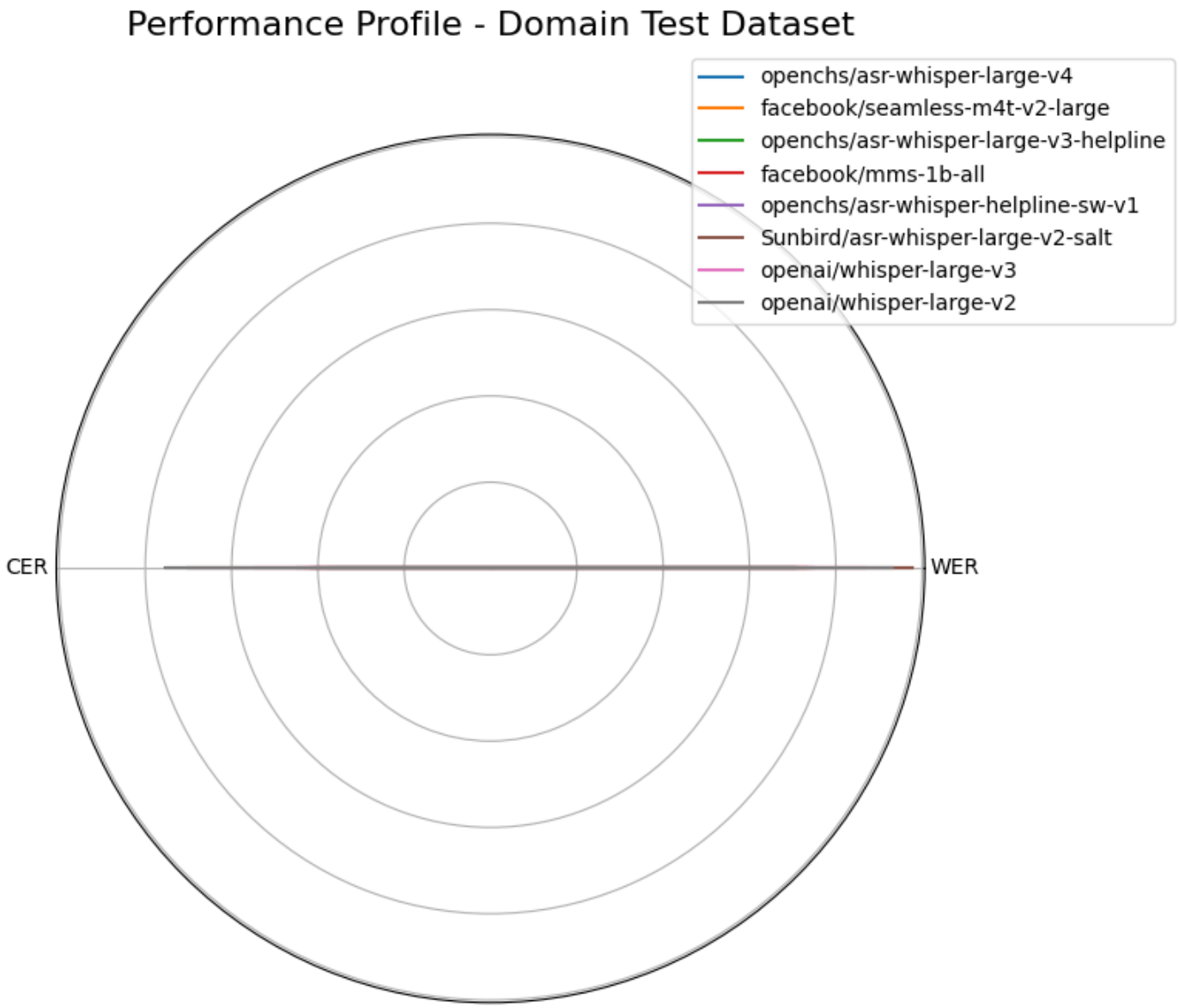
SECTION 4.3 - Dataset-Wise Heatmap: Domain Test Dataset



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

SECTION 4.4 - Radar Chart: Domain Test Dataset



Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

SECTION 5 - Analytical Summary

The fine-tuned models, particularly 'openchs/asr-whisper-large-v4', demonstrate superior performance on the domain-specific 'Helpline Audio' and 'Domain Test Dataset'. For instance, on the Helpline dataset, 'openchs/asr-whisper-large-v4' (61.03% WER) shows a significant reduction in errors compared to the general-purpose 'openai/whisper-large-v2' (226.47% WER). On benchmark datasets like FLEURS, large general-purpose models like 'facebook/mms-1b-all' still hold an advantage (15.71% WER). This indicates a clear trade-off: fine-tuning yields substantial gains for specific domains but may not surpass the top performers on general, clean audio benchmarks.

SECTION 6 - Limitations & Interpretation Caveats

- WER and CER may not fully capture intelligibility or context accuracy.

- Fine-tuned models may overfit to specific noise patterns in the training data.

- Datasets vary in noise levels, accents, and recording quality, affecting comparability.

- Benchmark datasets like FLEURS and Common Voice typically contain cleaner speech than real-world scenarios.

Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

SECTION 7 - Appendix / Reproducibility

Full Raw Data

Model	Dataset	WER	CER
openchs/asr-whisper-large-v4	Helpline Audio	61.03	25.42
facebook/seamless-m4t-v2-large	Helpline Audio	62.20	32.29
openchs/asr-whisper-large-v3-helpline	Helpline Audio	67.66	28.79
facebook/mms-1b-all	Helpline Audio	69.30	29.41
openchs/asr-whisper-helpline-sw-v1	Helpline Audio	69.94	37.26
Sunbird/asr-whisper-large-v2-salt	Helpline Audio	103.68	77.26
openai/whisper-large-v3	Helpline Audio	124.94	74.46
openai/whisper-large-v2	Helpline Audio	226.47	139.65
facebook/seamless-m4t-v2-large	Mozilla Common Voice 23.0-Swahili	25.83	22.03
openchs/asr-whisper-helpline-sw-v1	Mozilla Common Voice 23.0-Swahili	31.87	24.87
openchs/asr-whisper-large-v3-helpline	Mozilla Common Voice 23.0-Swahili	38.06	27.39
facebook/mms-1b-all	Mozilla Common Voice 23.0-Swahili	39.91	24.25
openchs/asr-whisper-large-v4	Mozilla Common Voice 23.0-Swahili	46.17	29.44
openai/whisper-large-v3	Mozilla Common Voice 23.0-Swahili	72.17	38.26
Sunbird/asr-whisper-large-v2-salt	Mozilla Common Voice 23.0-Swahili	94.15	49.78
openai/whisper-large-v2	Mozilla Common Voice 23.0-Swahili	95.05	55.34
facebook/mms-1b-all	FLEURS	15.71	4.12
facebook/seamless-m4t-v2-large	FLEURS	24.85	8.81
openchs/asr-whisper-large-v3-helpline	FLEURS	25.14	8.19
openchs/asr-whisper-helpline-sw-v1	FLEURS	25.52	8.25
openchs/asr-whisper-large-v4	FLEURS	28.41	11.59
openai/whisper-large-v3	FLEURS	46.13	11.70
openai/whisper-large-v2	FLEURS	52.72	13.71
Sunbird/asr-whisper-large-v2-salt	FLEURS	87.33	30.04
openchs/asr-whisper-large-v4	Domain Test Dataset	70.00	40.63
openchs/asr-whisper-large-v3-helpline	Domain Test Dataset	70.87	37.65
facebook/mms-1b-all	Domain Test Dataset	73.20	41.85
facebook/seamless-m4t-v2-large	Domain Test Dataset	74.03	43.17
openchs/asr-whisper-helpline-sw-v1	Domain Test Dataset	74.66	45.01
openai/whisper-large-v3	Domain Test Dataset	86.53	54.66
openai/whisper-large-v2	Domain Test Dataset	92.91	75.48
Sunbird/asr-whisper-large-v2-salt	Domain Test Dataset	97.62	70.40

Speech-to-Text Model Evaluation

Domain-Specific ASR Performance Comparison

Generated Charts

radar_chart.png

cer_bar_chart.png

mozilla_common_voice_23.0-swahili_heatmap.png

domain_test_dataset_perf_chart.png

wer_bar_chart.png

mozilla_common_voice_23.0-swahili_perf_chart.png

fleurs_heatmap.png

helpline_audio_perf_chart.png

fleurs_radar.png

helpline_audio_radar.png

domain_test_dataset_radar.png

wer_heatmap.png

fleurs_perf_chart.png

mozilla_common_voice_23.0-swahili_radar.png

wer_metric_chart.png

cer_metric_chart.png

helpline_audio_heatmap.png

domain_test_dataset_heatmap.png

Tools Used

Python (pandas, matplotlib, seaborn, fpdf2)