Table 3: Model performance in Scenario 1

Model	**Macro F1↑**	**Micro F1↑**	**Exact Match↑**
gemini-2.0-flash-lite	0.34	0.46	0.48
gemini-2.0-flash	0.33	0.46	0.46
gpt-4o-mini-audio	0.21	0.20	0.20
gpt-4o-audio	0.16	0.44	0.44
gpt-4o-mini-transcribe	0.34	0.56	0.56
gpt-4o-transcribe	0.37	0.47	0.47
whisper-gpt4o	0.37	0.47	0.47
qwen2.5-omni-7b	0.79	0.71	0.71
qwen2.5-omni-3b	0.59	0.42	0.42
qwen2-audio-7b-instruct	0.21	0.45	0.45
qwen-audio-chat	0.00	0.00	0.00
phi-multimodal	0.25	0.32	0.32
granite-speech-3.3-8b	0.00	0.00	0.00
granite-speech-3.3-2b	0.00	0.00	0.00
granite-speech-3.2-8b	0.00	0.00	0.00
Finetuned Models with Asterisk			
qwen2.5-omni-7b (finetuned)	0.93	0.95	0.95
qwen2.5-omni-3b (finetuned)	0.76	0.89	0.89
qwen2-audio-instruct (finetuned)	0.01	0.30	0.30
Finetuned Models without Asterisk			
qwen2.5-omni-7b (finetuned)	0.81	0.67	0.67
qwen2.5-omni-3b (finetuned)	0.20	0.25	0.25
qwen2-audio-instruct (finetuned)	0.01	0.36	0.36
Fewshot Prompting			
gpt-4o-mini-audio	0.19	0.23	0.23
gpt-4o-audio	0.21	0.71	0.71
gpt-4o-mini-transcribe	0.44	0.72	0.72
gpt-4o-transcribe	0.45	0.72	0.72
whisper-gpt4o	0.24	0.72	0.72

Table 4: Model performance in Scenario 2

Model	**Macro F1↑**	**Micro F1↑**	**Exact Match↑**
gemini-2.0-flash-lite	0.01	0.01	0.01
gemini-2.0-flash	0.00	0.00	0.00
gpt-4o-mini-audio	0.03	0.07	0.07
gpt-4o-audio	0.01	0.02	0.02
gpt-4o-transcribe	0.03	0.13	0.13
whisper-gpt4o	0.00	0.00	0.00
qwen2.5-omni-7b	0.01	0.07	0.07
qwen2.5-omni-3b	0.00	0.01	0.01
qwen2-audio-7b-instruct	0.00	0.03	0.03
qwen-audio-chat	0.00	0.00	0.00
phi-multimodal	0.00	0.06	0.06
granite-speech-3.3-8b	0.00	0.03	0.03
granite-speech-3.3-2b	0.00	0.02	0.02
granite-speech-3.2-8b	0.00	0.04	0.04
Finetuned Models with Asterisk			
qwen2.5-omni-7b (finetuned)	0.06	0.44	0.44
qwen2.5-omni-3b (finetuned)	0.08	0.23	0.23
qwen2-audio-instruct (finetuned)	0.03	0.32	0.32
Finetuned Models without Asterisk			
qwen2.5-omni-7b (finetuned)	0.06	0.44	0.44
qwen2.5-omni-3b (finetuned)	0.08	0.23	0.23
qwen2-audio-instruct (finetuned)	0.16	0.44	0.44
Fewshot Prompting			
gpt-4o-mini-audio	0.03	0.09	0.09
gpt-4o-audio	0.04	0.12	0.12
gpt-4o-mini-transcribe	0.01	0.01	0.01
gpt-4o-transcribe	0.01	0.01	0.01
whisper-gpt4o	0.00	0.01	0.01

Table 5: Model performance in Scenario 3

Model	WER↓	MER↓	WIP↑
gemini-2.0-flash-lite	0.83	0.68	0.21
gemini-2.0-flash	0.94	0.71	0.24
gpt-4o-mini-audio	1.40	0.68	0.26
gpt-4o-audio	2.25	0.70	0.26
gpt-4o-mini-transcribe	1.54	0.75	0.19
gpt-4o-transcribe	1.31	0.74	0.23
whisper-gpt4o	2.84	0.75	0.18
qwen2.5-omni-7b	2.17	0.74	0.22
qwen2.5-omni-3b	4.98	0.75	0.22
qwen2-audio-7b-instruct	4.98	0.75	0.22
qwen-audio-chat	12.3	0.90	0.08
phi-multimodal	6.36	0.76	0.20
granite-speech-3.3-8b	13.50	0.93	0.05
granite-speech-3.3-2b	4.13	0.89	0.07
granite-speech-3.2-8b	6.64	0.66	0.14
Finetuned Models with Asterisk			
qwen2.5-omni-7b (finetuned)	1.40	0.52	0.41
qwen2.5-omni-3b (finetuned)	0.97	0.53	0.39
qwen2-audio-instruct (finetuned)	0.58	0.43	0.50
Finetuned Models without Asterisk			
qwen2.5-omni-7b (finetuned)	1.76	0.46	0.47
qwen2.5-omni-3b (finetuned)	0.95	0.49	0.43
qwen2-audio-instruct (finetuned)	0.52	0.38	0.56
Fewshot Prompting			
gpt-4o-mini-audio	1.58	0.65	0.28
gpt-4o-audio	1.73	0.62	0.30
gpt-4o-mini-transcribe	1.08	0.80	0.12
gpt-4o-transcribe	1.01	0.79	0.14
whisper-gpt4o	1.89	0.77	0.16

Table 6: Model performance in Scenario 4

Model	**Macro F1↑**	**Micro F1↑**	**Exact Match↑**
gemini-2.0-flash-lite	0.17	0.19	0.19
gemini-2.0-flash	0.20	0.46	0.46
gpt-4o-mini-audio	0.12	0.15	0.15
gpt-4o-audio	0.14	0.36	0.36
gpt-4o-mini-transcribe	0.28	0.42	0.46
gpt-4o-transcribe	0.32	0.41	0.41
whisper-gpt4o	0.33	0.43	0.41
qwen2.5-omni-7b	0.79	0.71	0.71
qwen2.5-omni-3b	0.56	0.44	0.44
qwen2-audio-7b-instruct	0.20	0.33	0.33
qwen-audio-chat	0.00	0.00	0.00
phi-multimodal	0.18	0.37	0.37
granite-speech-3.3-8b	0.00	0.00	0.00
granite-speech-3.3-2b	0.00	0.00	0.00
granite-speech-3.2-8b	0.00	0.01	0.01
Finetuned Models with Asterisk			
qwen2.5-omni-7b (finetuned)	0.93	0.95	0.95
qwen2.5-omni-3b (finetuned)	0.76	0.89	0.89
qwen2-audio-instruct (finetuned)	0.02	0.21	0.21
Finetuned Models without Asterisk			
qwen2.5-omni-7b (finetuned)	0.28	0.40	0.40
qwen2.5-omni-3b (finetuned)	0.24	0.36	0.36
qwen2-audio-instruct (finetuned)	0.05	0.27	0.27

Table 7: Model performance in Scenario 5

Model	**Macro F1↑**	**Micro F1↑**	**Exact Match↑**
gemini-2.0-flash-lite	0.19	0.43	0.43
gemini-2.0-flash	0.09	0.22	0.22
gpt-4o-mini-audio	0.10	0.39	0.39
gpt-4o-audio	0.20	0.49	0.49
gpt-4o-mini-transcribe	0.15	0.26	0.26
gpt-4o-transcribe	0.13	0.28	0.28
whisper-gpt4o	0.18	0.36	0.36
qwen2.5-omni-7b	0.79	0.71	0.71
qwen2.5-omni-3b	0.56	0.44	0.44
qwen2-audio-7b-instruct	0.08	0.10	0.10
qwen-audio-chat	0.00	0.00	0.00
phi-multimodal	0.09	0.13	0.13
granite-speech-3.3-8b	0.00	0.00	0.00
granite-speech-3.3-2b	0.00	0.00	0.00
granite-speech-3.2-8b	0.06	0.20	0.20
Finetuned Models with Asterisk			
qwen2.5-omni-7b (finetuned)	0.93	0.95	0.95
qwen2.5-omni-3b (finetuned)	0.76	0.89	0.89
qwen2-audio-instruct (finetuned)	0.00	0.07	0.07
Finetuned Models without Asterisk			
qwen2.5-omni-7b (finetuned)	0.16	0.34	0.34
qwen2.5-omni-3b (finetuned)	0.08	0.16	0.16
qwen2-audio-instruct (finetuned)	0.01	0.08	0.08