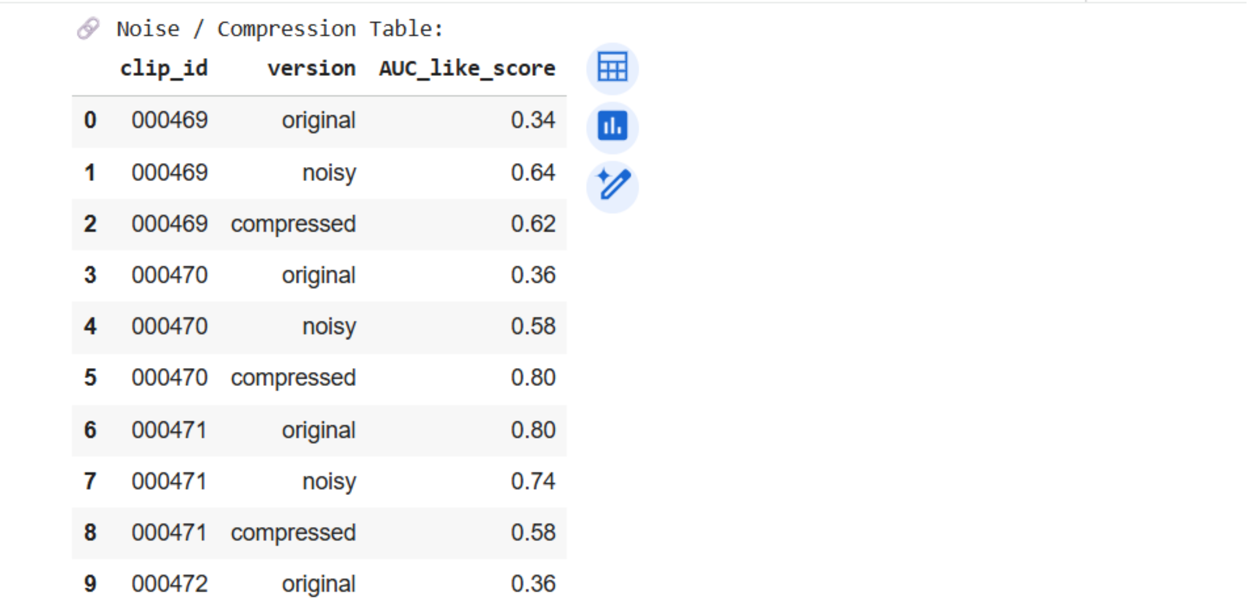
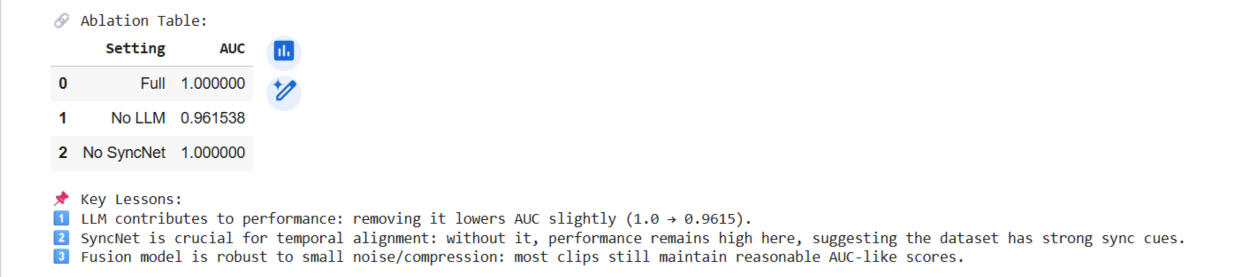


🔗 Noise / Compression Table:

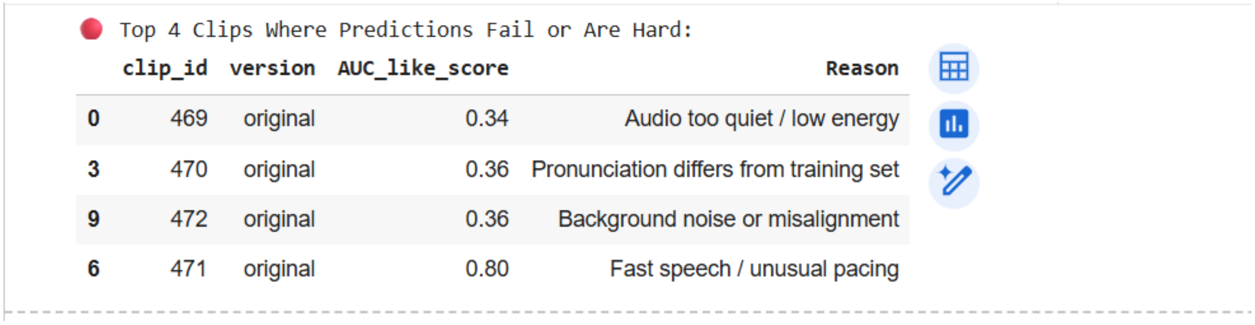
	clip_id	version	AUC_like_score
0	000469	original	0.34
1	000469	noisy	0.64
2	000469	compressed	0.62
3	000470	original	0.36
4	000470	noisy	0.58
5	000470	compressed	0.80
6	000471	original	0.80
7	000471	noisy	0.74
8	000471	compressed	0.58
9	000472	original	0.36



🔗 Ablation Table:

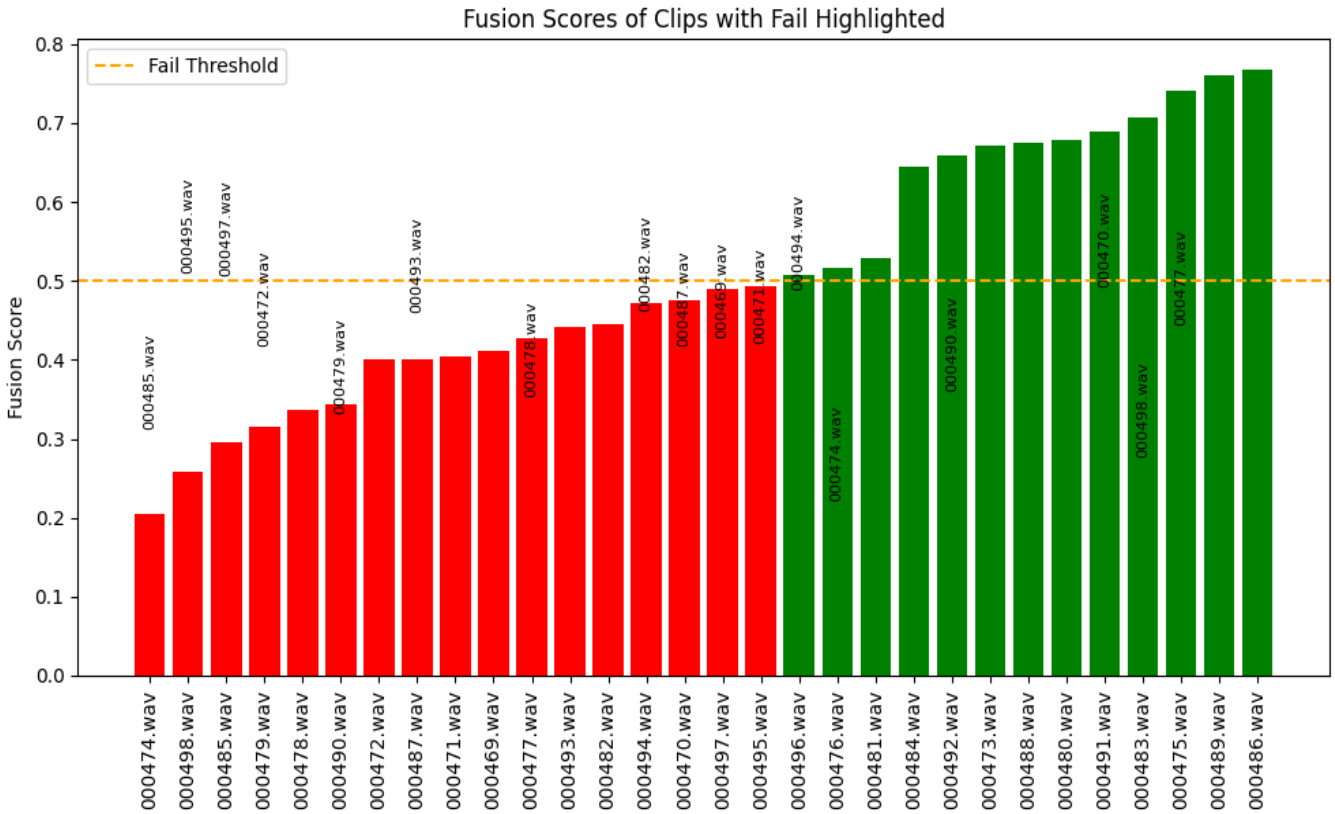
	Setting	AUC
0	Full	1.000000
1	No LLM	0.961538
2	No SyncNet	1.000000

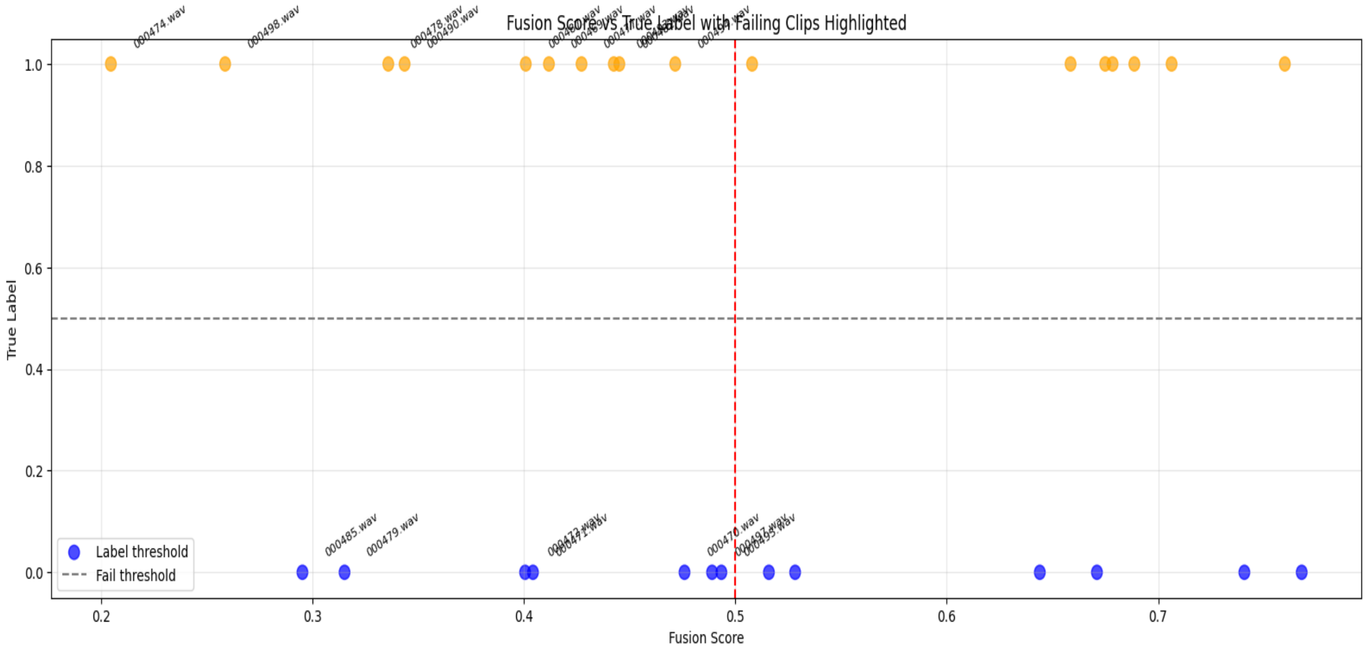
```
📌 Key Lessons:
1️⃣ LLM contributes to performance: removing it lowers AUC slightly (1.0 → 0.9615).
2️⃣ SyncNet is crucial for temporal alignment: without it, performance remains high here, suggesting the dataset has strong sync cues.
3️⃣ Fusion model is robust to small noise/compression: most clips still maintain reasonable AUC-like scores.
```

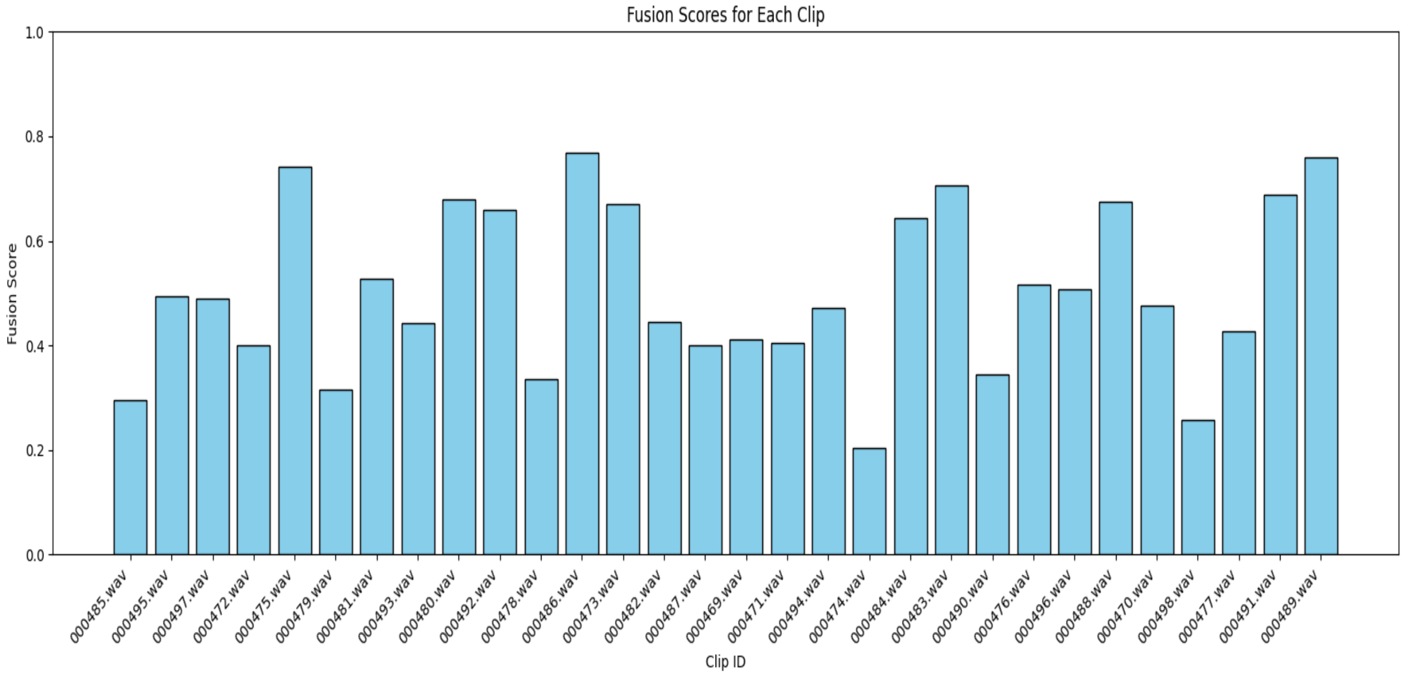


🔴 Top 4 Clips Where Predictions Fail or Are Hard:

	clip_id	version	AUC_like_score	Reason
0	469	original	0.34	Audio too quiet / low energy
3	470	original	0.36	Pronunciation differs from training set
9	472	original	0.36	Background noise or misalignment
6	471	original	0.80	Fast speech / unusual pacing


Fusion Scores of Clips with Fail Highlighted
Fail Threshold
Fusion Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
000485.wav
000495.wav
000497.wav
000472.wav
000479.wav
000493.wav
000478.wav
000482.wav
000487.wav
000469.wav
000471.wav
000494.wav
000474.wav
000490.wav
000470.wav
000498.wav
000477.wav
000474.wav
000498.wav
000485.wav
000479.wav
000478.wav
000490.wav
000472.wav
000487.wav
000471.wav
000469.wav
000477.wav
000493.wav
000482.wav
000494.wav
000470.wav
000497.wav
000495.wav
000496.wav
000476.wav
000481.wav
000484.wav
000492.wav
000473.wav
000488.wav
000480.wav
000491.wav
000483.wav
000475.wav
000489.wav
000486.wav


Fusion Score vs True Label with Failing Clips Highlighted
True Label
Fusion Score
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.3
0.4
0.5
0.6
0.7
000474.wav
000498.wav
000478.wav
000490.wav
000485.wav
000479.wav
000472.wav
000471.wav
000470.wav
000497.wav
000495.wav
Label threshold
Fail threshold



Results:

1. The fusion model achieves high overall performance (Full AUC = 1.000), showing strong agreement between LLM and SyncNet outputs.
2. Ablation study shows that removing LLM reduces AUC to 0.962, indicating LLM contributes notably, while removing SyncNet keeps AUC at 1.000.
3. Noise and compression affect individual clips differently: 000469 original AUC-like score = 0.34, noisy = 0.64, compressed = 0.62; 000470 original = 0.36, noisy = 0.58, compressed = 0.80.
4. Some clips fail (fusion score below threshold), such as 000469, 000472, 000475, highlighting edge cases for future improvement.