DALLA LINGUA DEI SEGNI ALLA VOCE: SVILUPPO DI UNA PIPELINE PER LA SINTESI DEL PARLATO ESPRESSIVO

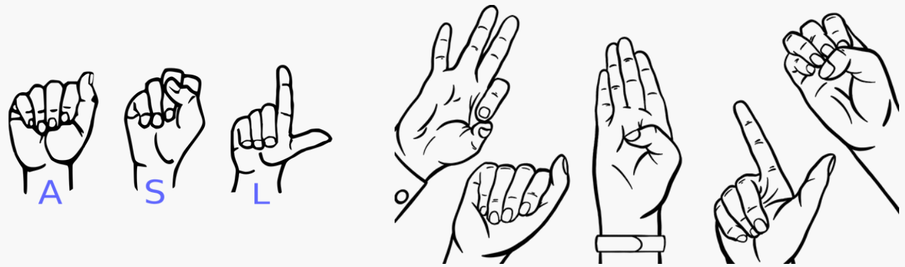
Relatore
Bacco Luca

Correlatori
Merone Mario
Sasso Daniele

Laureando
Ignazio Emanuele Piccichè

18 Dicembre, 2025

Oltre la Traduzione Letterale: Il Ruolo del Sentiment



Il Limite della SLT

La ricerca attuale si concentra sul tradurre **cosa** viene detto (il testo), ignorando spesso **come** viene detto (le Componenti Non-Manuali)

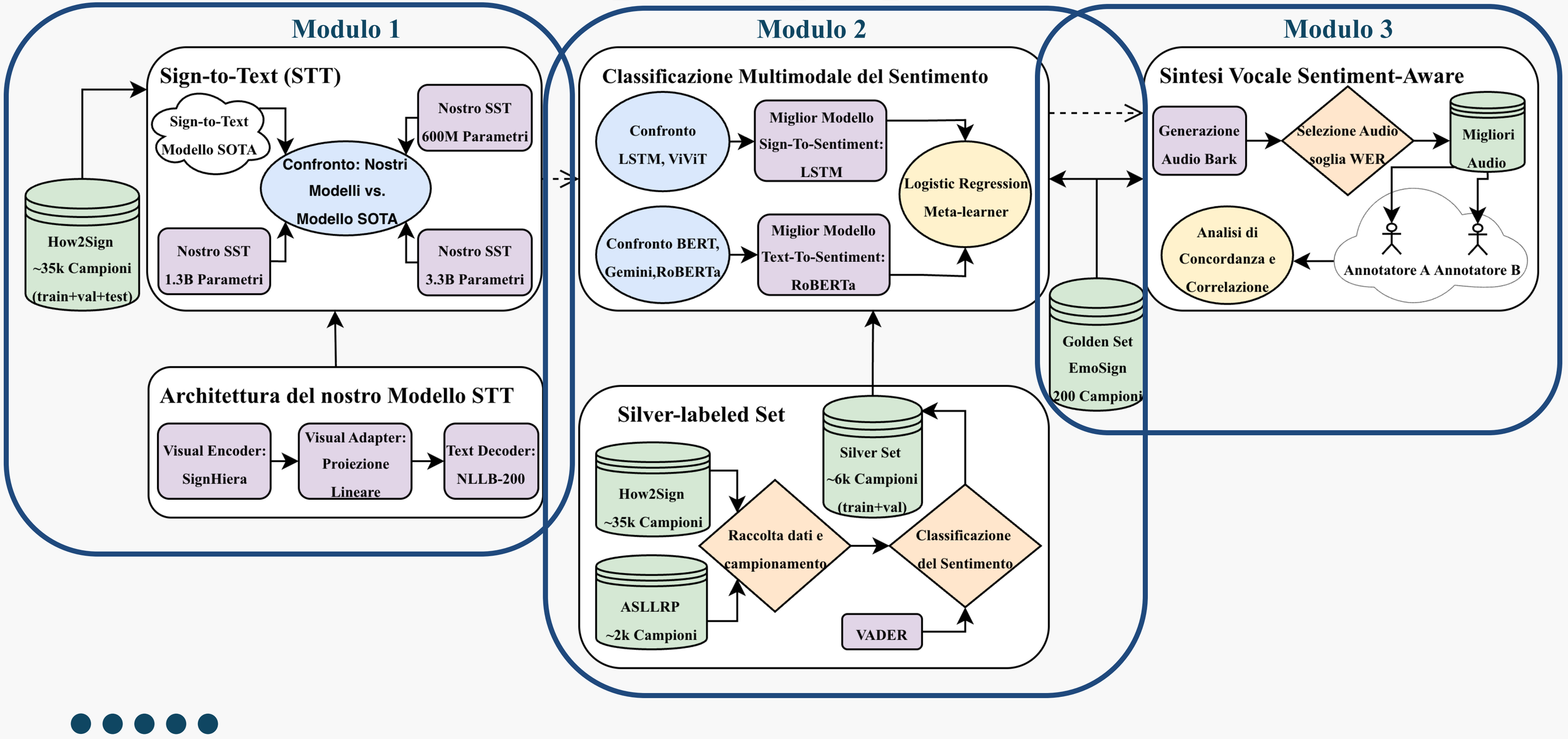
Il Problema della Voce

I sintetizzatori vocali standard producono un audio "piatto" (monotono) che non riflette l'intensità espressiva del segnante sordo.

La Visione

Sviluppare una pipeline che non si limiti a tradurre le parole, ma trasferisca l'**intenzione del Sentiment** dal segno alla voce sintetica.

Panoramica della Pipeline proposta



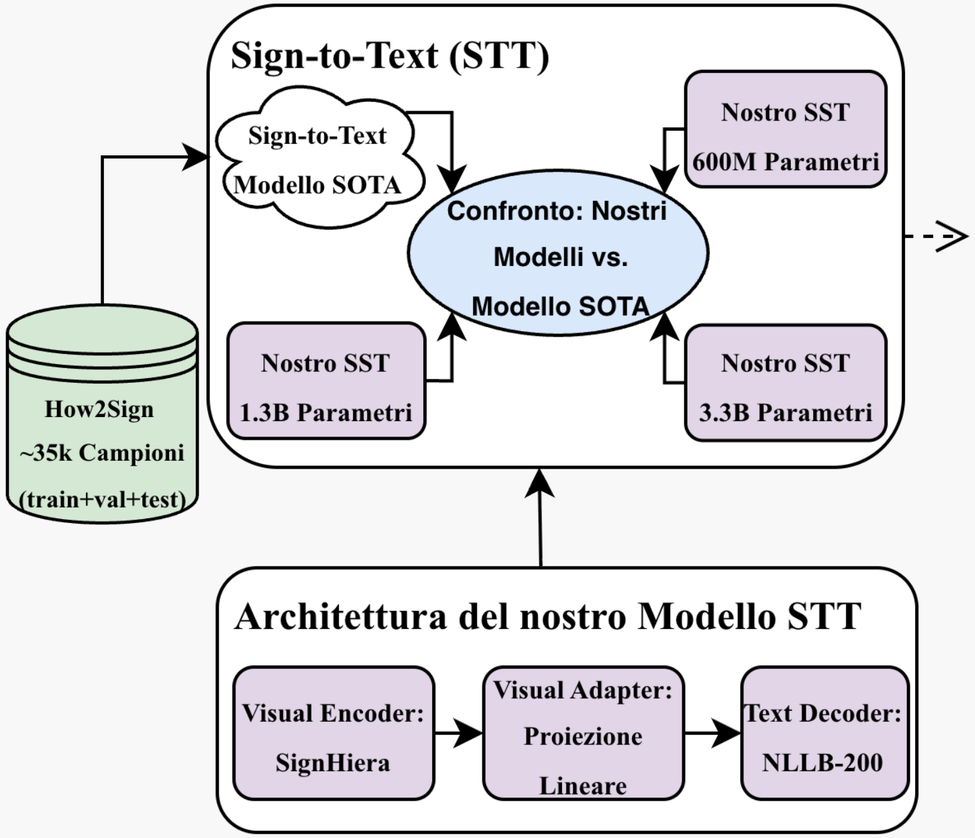
Modulo 1: Sign-to-Text

Adattamento Architetturale NLLB-200 + Visual Adapter (Video Domain).

Strategia Low-Resource Frozen Encoder (vs Full Fine-Tuning).

Benchmark comparativo: NLLB vs SOTA

Modello	Parametri	Note
NLLB-600M	600 Milioni	Variante leggera (Distilled)
NLLB-1.3B	**1.3 Miliardi**	**Modello con dimensione media**
NLLB-3.3B	3.3 Miliardi	Modello con maggiori parametri
SVVP-SLT (SOTA)	*53 Miliardi*	*Modello di riferimento (Meta)*



Modulo 2: Classificazione Multimodale del Sentimento

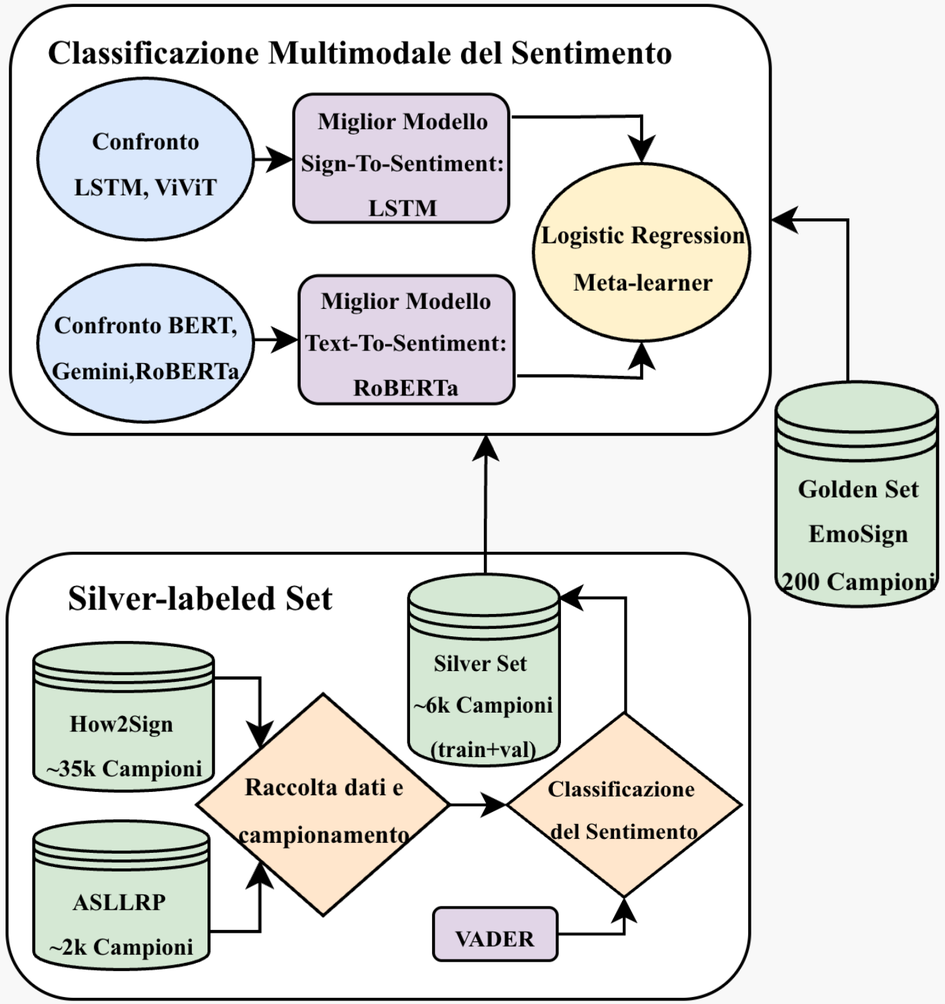
Silver Dataset (6k Campioni) Annotazione automatica su corpus ibrido (ASLLRP + How2Sign).

Gli "Esperti" Unimodali

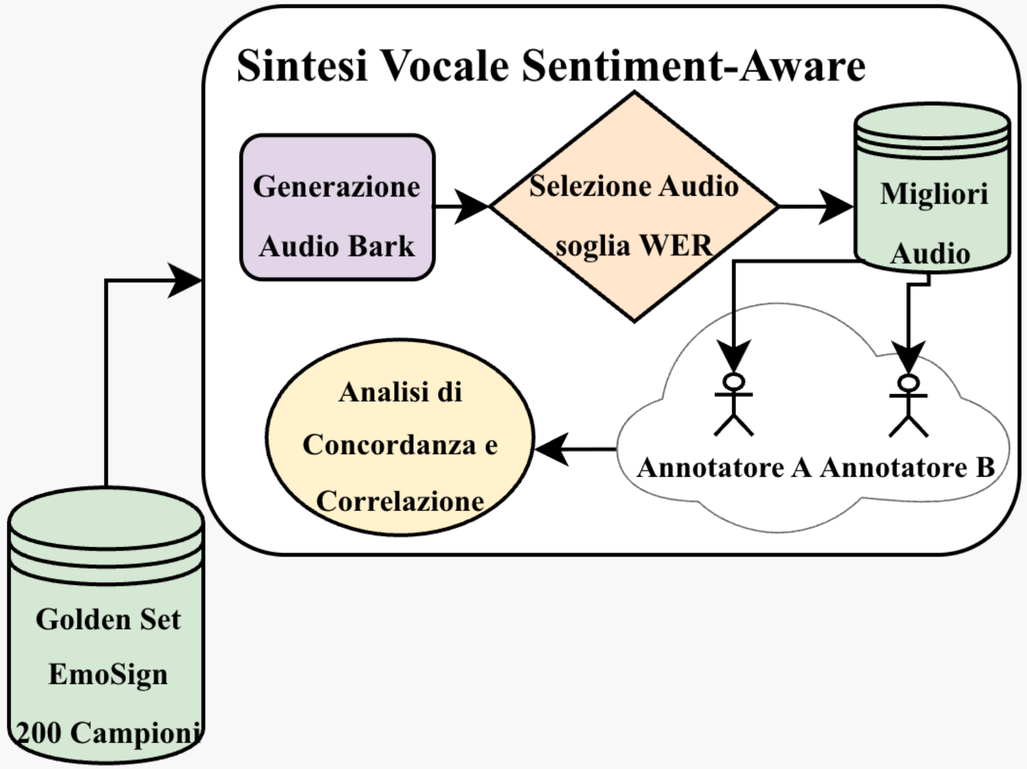
Esperto Visivo	
Modello	**Architettura**
LSTM	Analisi cinematica su scheletri
ViViT	Vision Transformer (Video)
ST-GCN	Graph Conv. Network

Esperto Testuale	
Modello	**Architettura**
RoBERTa (Twitter)	Sentiment (3 classi)
BERT (GoEmotions)	Emozioni (27 classi)
Gemini	LLM Generativo

Strategia "Late Fusion" Un **Meta-Learner** unisce le probabilità dei due esperti per la decisione finale.



Modulo 3: Sentiment-Aware TTS



Modello Generativo (Bark) Fonemi + Tratti Paralinguistici (risate, sospiri).

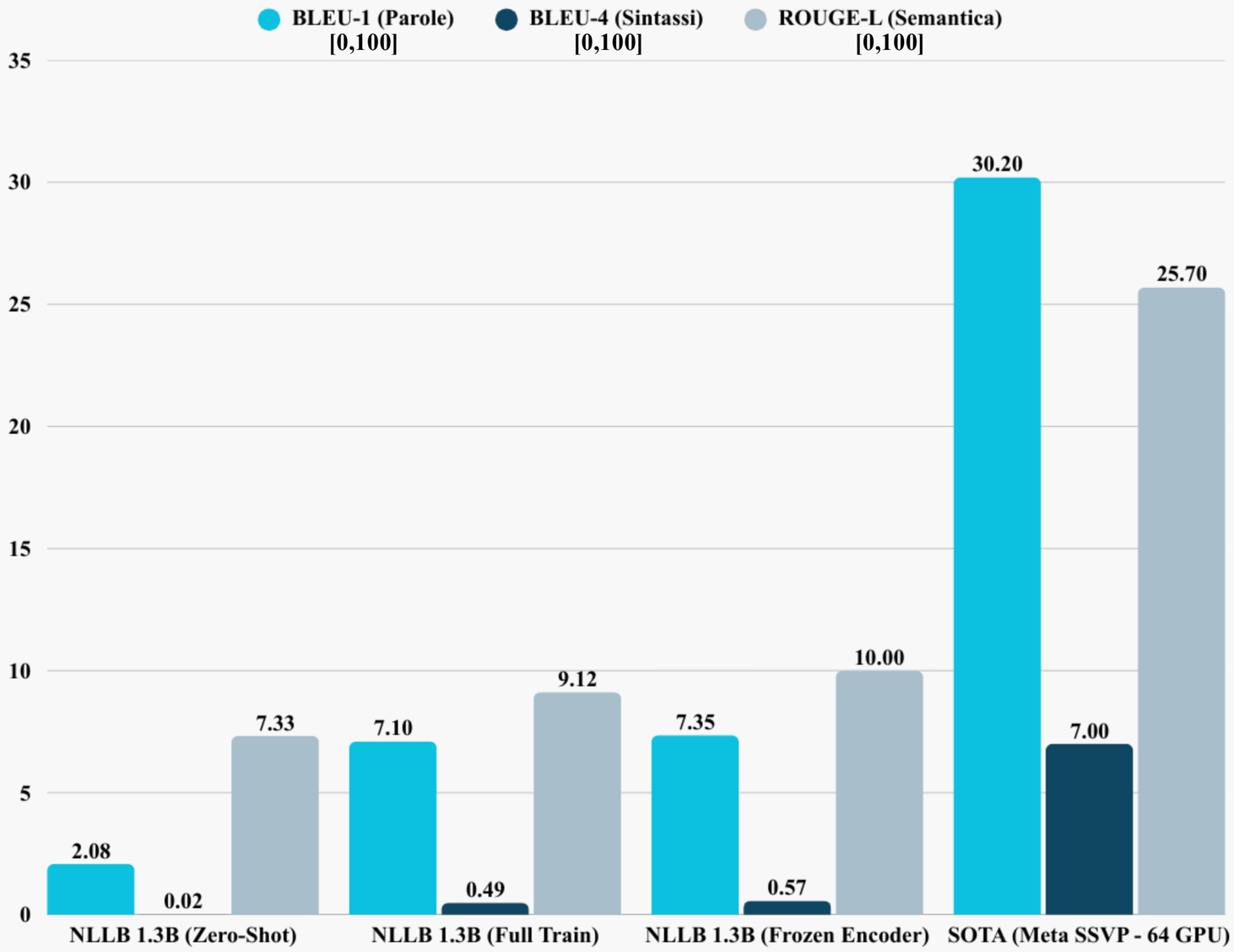
Prompting Dinamico Mappatura diretta:

- Scelta degli *speaker*
- Mappatura della *temperatura*
- *Token paralinguistici*

Validazione Ibrida

- **Tecnica:** Filtro automatico (WER) per scartare audio sintattica non corretti.
- **Umana:** Analisi percettiva con annotatori per valutare la Coerenza Emotiva
  - *Testo* + *Audio* con flag *artefatti sonori*

Risultati Sign-to-Text (Modulo 1)

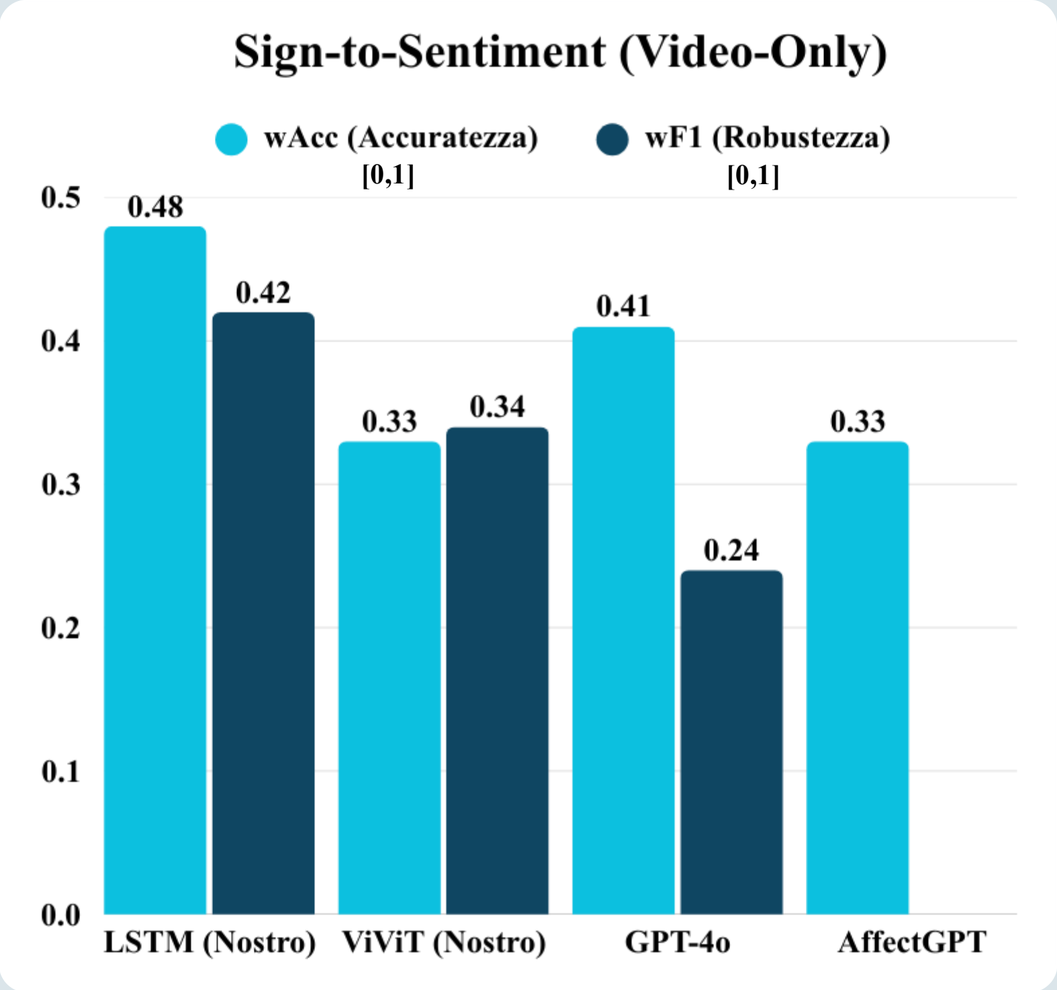


Gap Sintattico (BLEU-4): La sintassi complessa richiede pre-training massivo (milioni di video vs 35k).

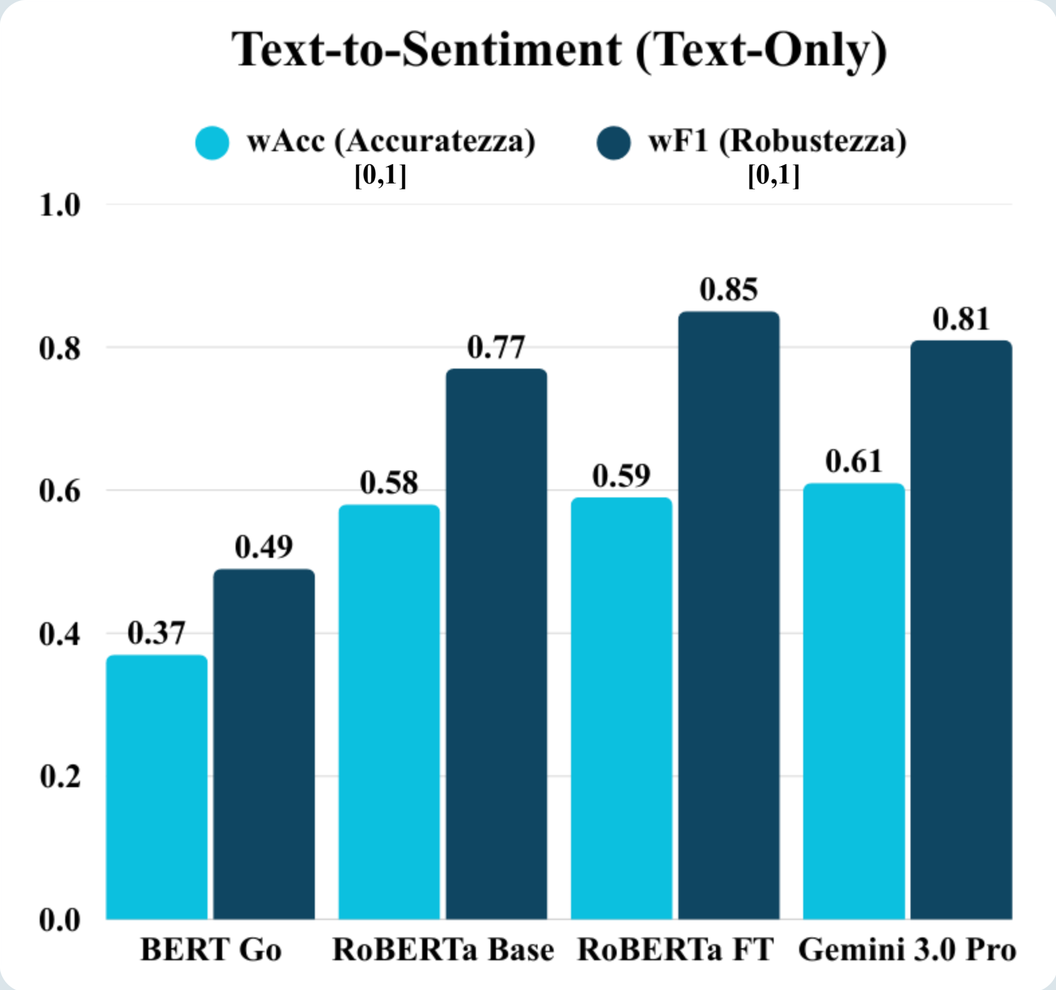
Tenuta Semantica (ROUGE-L):
Il modello *Frozen* cattura il significato (Score 10.0), superando il Full Train.

Efficienza: Risultati ottenuti prevenendo il *Catastrophic Forgetting* su singola GPU.

Risultati degli Stream Visivo e Testuale (Modulo 2)



L'LSTM specializzato supera i modelli generalisti.



Il Fine-Tuning specifico batte modelli di frontiera.

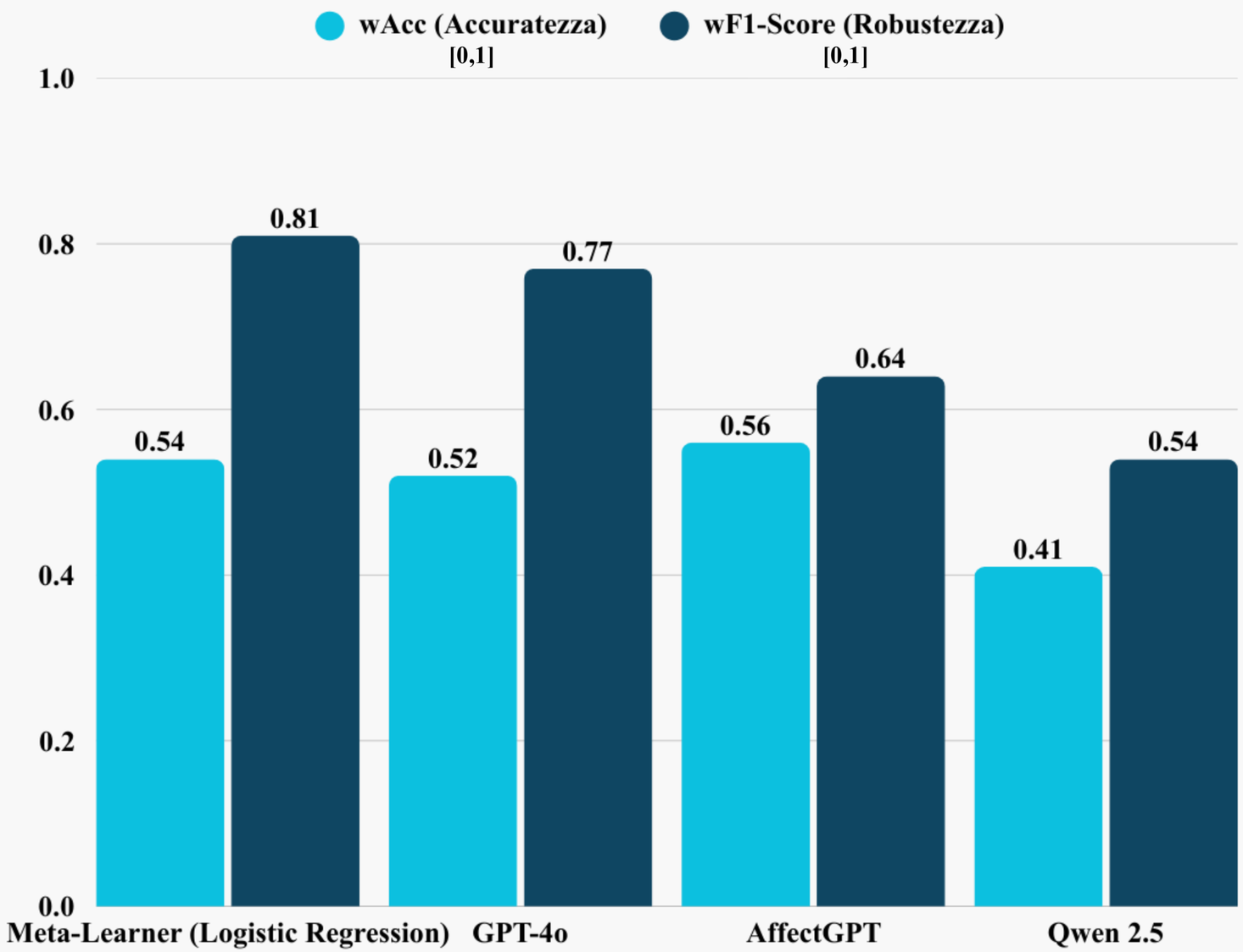
Risultati Meta-learner per il Riconoscimento del Sentimento

Model Selection

- **Architettura:** Logistic Regression (scelta via Grid Search).
- **Motivo:** La linearità del modello conferma la purezza dei segnali degli "esperti" (LSTM + RoBERTa).

Confronto con lo Stato dell'Arte

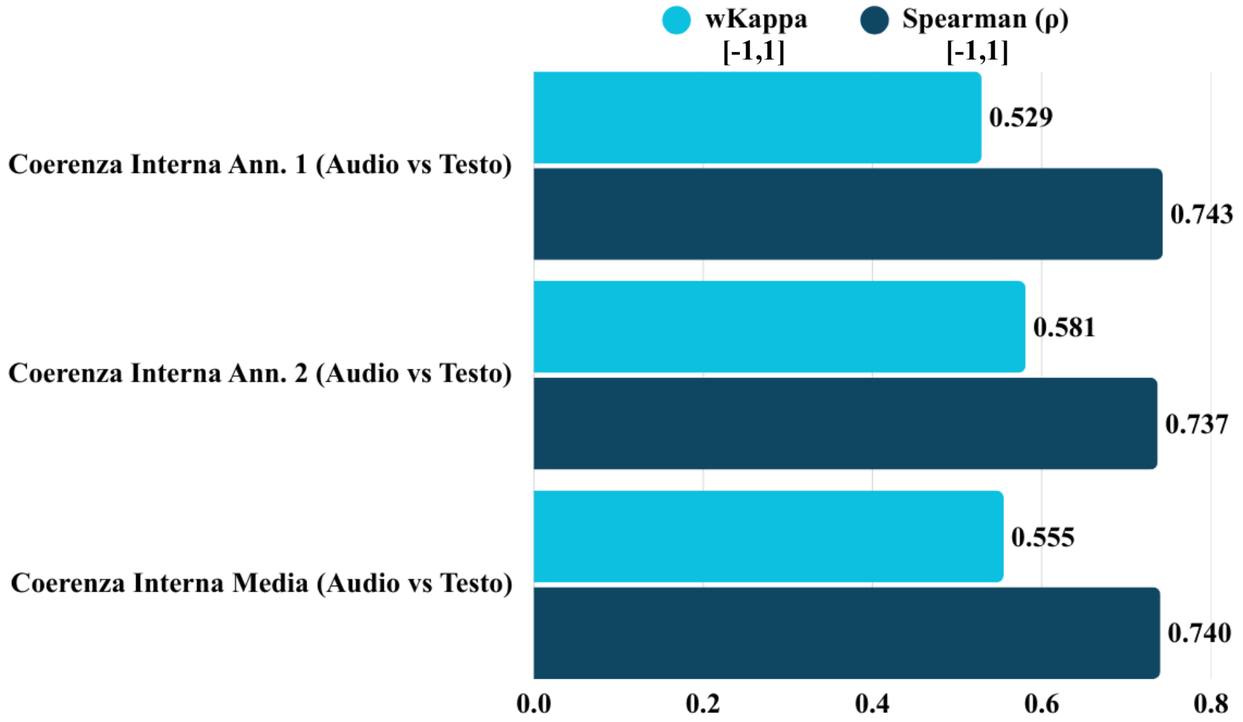
- **Nuovo SOTA:** Superato **GPT-4o** di **+4.1%** nella metrica chiave (wF1).
- **Analisi Competitor:** AffectGPT mostra un bias verso la classe dominante (alta Accuracy, basso F1), mentre il nostro modello è bilanciato.



Risultati Sintesi Vocale *Sentiment-Aware* (Modulo 3)

Audio scartati con filtro WER: ~46%

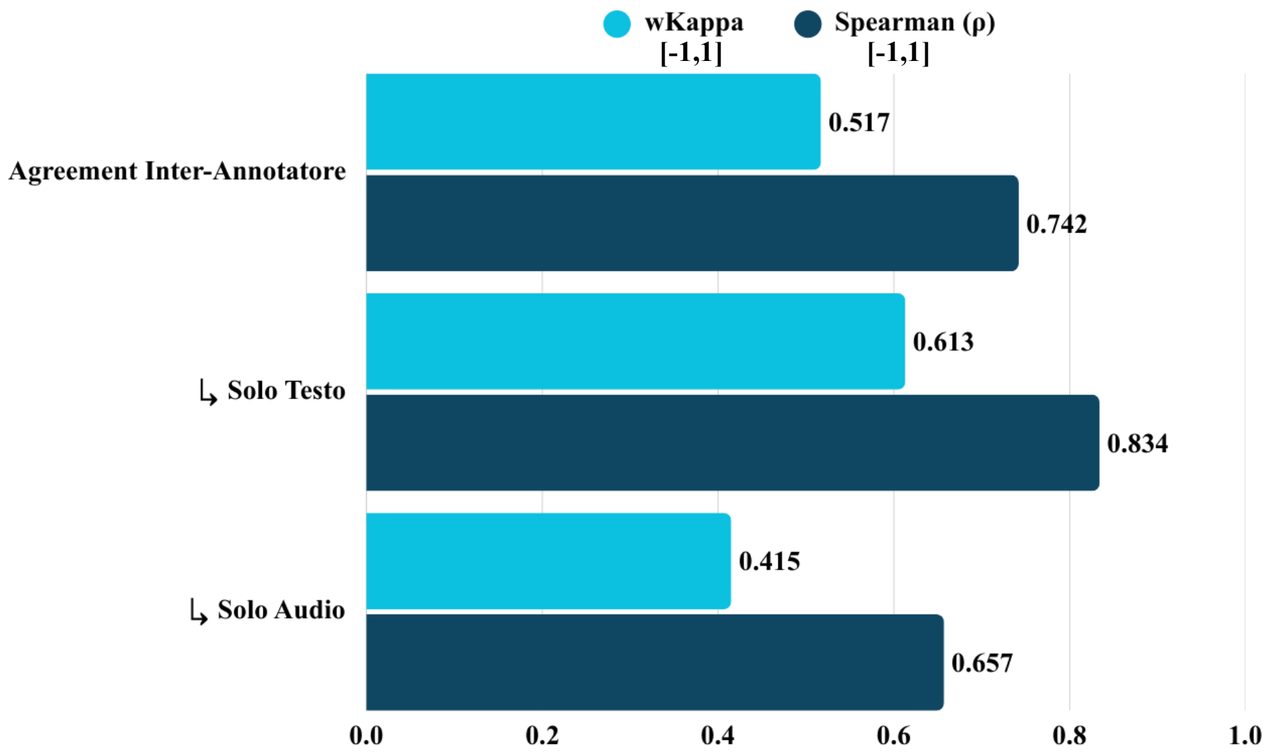
Audio scartati dagli annotatori: ~8%



Coerenza Audio-Testo:

La direzione emotiva (Positivo/Negativo) viene preservata fedelmente.

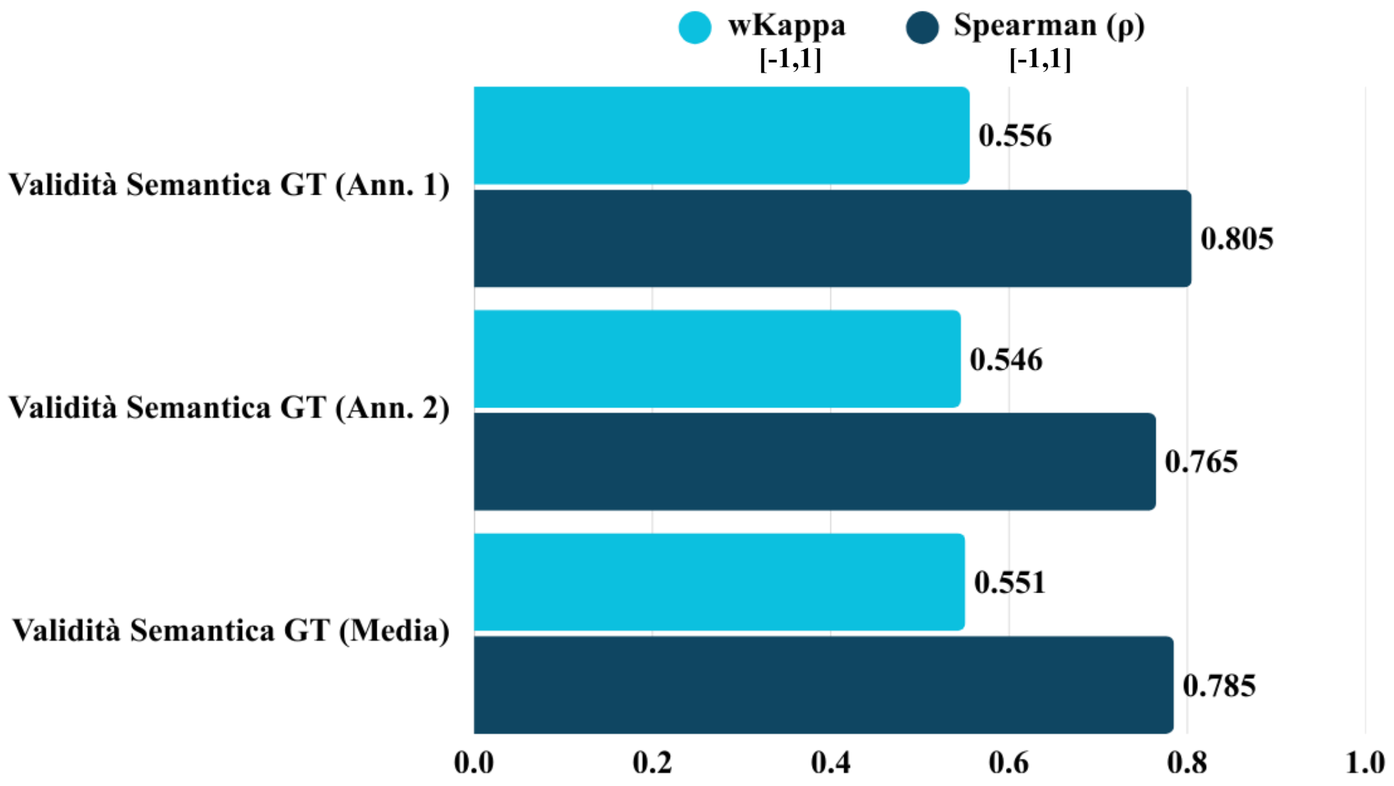
Risultati Sintesi Vocale *Sentiment-Aware* (Modulo 3)



Gap di Soggettività:

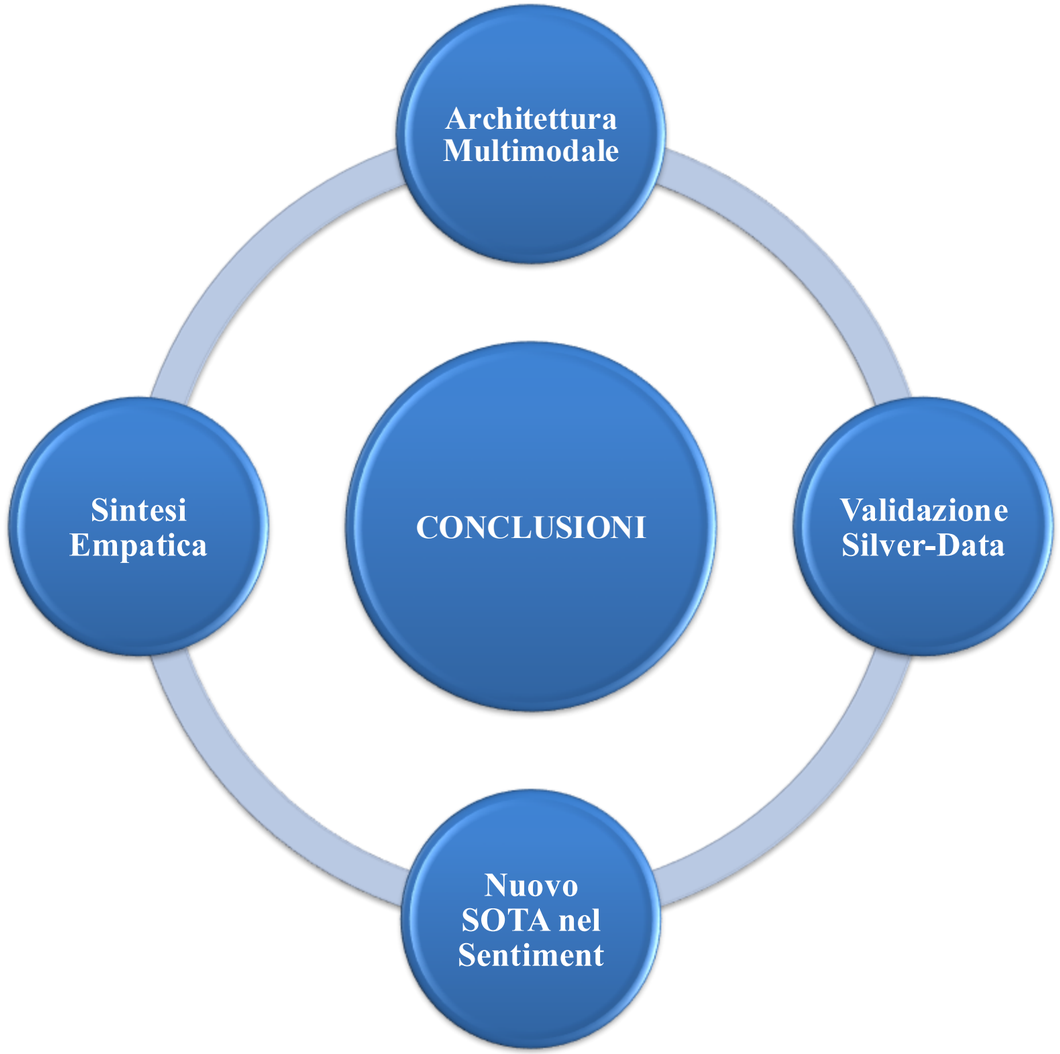
L'accordo scende fisiologicamente da **0.834 (Testo)** a **0.657 (Audio)**.

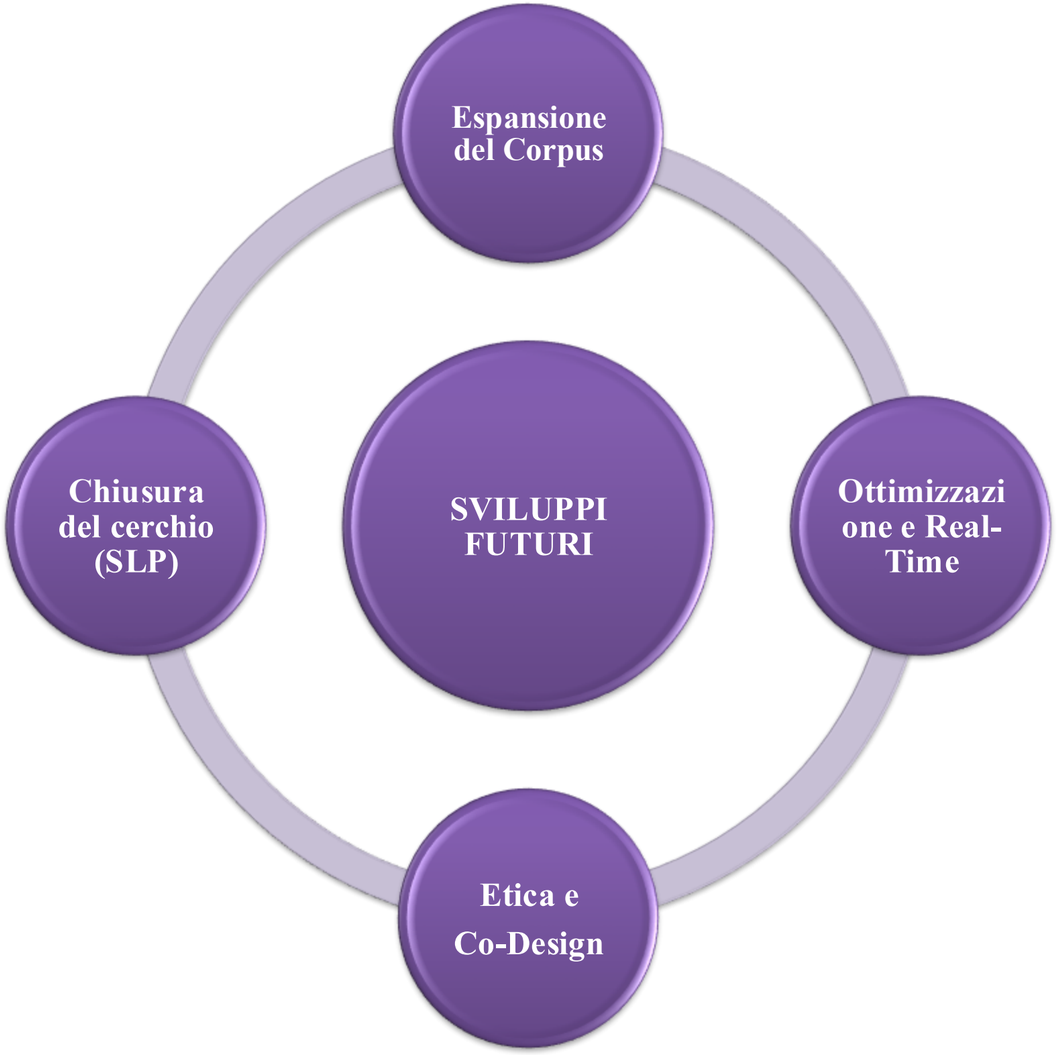
Risultati Sintesi Vocale *Sentiment-Aware* (Modulo 3)



Bias di Intensità: Il modello tende a smorzare gli estremi emotivi (compressione dinamica).

Conclusioni e Sviluppi Futuri





Grazie per l’attenzione

Appendice I

Metriche Mod. 1

Metrica	Focus Principale	Descrizione e Utilizzo
BLEU-1	Precisione (Unigrammi)	**Adeguatezza**: verifica se le singole parole sono presenti (fedeltà lessicale).
BLEU-2	Precisione (Bigrammi)	**Fluidità locale**: valuta l'ordine delle parole a breve raggio.
BLEU-3	Precisione (Trigrammi)	**Fluidità**: valuta sequenze più lunghe per una maggiore naturalezza.
BLEU-4	Precisione (4-grammi)	**Grammatica**: valuta la struttura di frasi complesse e la coerenza sintattica.
ROUGE-L	Recall (LCS)	**Struttura frase**: basato sulla *Sottosequenza Comune Più Lunga*. Cattura la struttura senza n-grammi fissi (ideale per riassunti).

Metriche Mod. 2

Metrica	Formula (Concetto)	A cosa serve / Quando usarla
Accuracy (Standard)	(Corretti / Totale)	**Visione d'insieme**: Ottima per dataset *bilanciati*
Balanced Accuracy	Media della Recall per ogni classe recall: T1/(T1+F1)	**Dataset sbilanciati**: Normalizza l'accuracy per evitare che la classe di maggioranza "nasconda" gli errori sulle classi rare.
F1-Score (Standard)	$2 * \frac{Precision * Recall}{Precision + Recall}$	**Equilibrio Prec/Recall**: Fondamentale quando Falsi Positivi e Falsi Negativi sono entrambi costosi. Focus su una classe specifica (es. "Positivo").
F1-Score (Weighted)	Media F1 pesata per il "supporto"	**Multi-classe sbilanciato**: Calcola l'F1 per ogni classe e fa la media pesando per il numero di esempi reali di quella classe. Restituisce un valore unico rappresentativo del sistema.

Metriche Mod. 3

Metrica	Focus / Tipo Dati	Cosa misura / A cosa serve
Cohen's Kappa (Standard)	Accordo tra giudici (Categorie nominali)	**Affidabilità oltre il caso**: Misura quanto due valutatori sono d'accordo (es. "Sì/No"), togliendo la probabilità che siano d'accordo per pura fortuna.
Cohen's Kappa (Weighted)	Accordo tra giudici (Categorie ordinali)	**Gravità dell'errore**: Come sopra, ma penalizza di più gli errori "grandi". Es: se la scala è 1-5, scambiare 1 con 5 è un errore più grave che scambiare 1 con 2.
Pearson Correlation (r)	Relazione Lineare (Dati continui)	**Dati distribuiti normalmente**: Misura quanto due variabili crescono insieme a ritmo costante. Valori da -1 a +1.
Spearman Correlation (rho)	Relazione Monotona (Ranghi / Ordinale)	**Ordine (Ranking)**: Non guarda i valori grezzi ma la "classifica" (rank). Capisce se Y cresce quando X cresce, anche se non in modo lineare (es. esponenziale). Meno sensibile agli outlier.

Appendice II

Note Mod. 1

- Il **Visual Encoder (SignHiera)** costituisce il primo stadio del sistema ed è responsabile dell'estrazione delle
  - **Feature spaziali e temporali**
  - **La "testa" del modello è rimossa** per restituire una sequenza di vettori latenti (**feature map**) che rappresentano la **semantica visiva** del video nel corso del tempo.
- **Visual Adapter**
  - **Multi-Layer Perceptron (MLP):** proiezioni lineari, attivazioni non lineari (ReLU) e regolarizzazione (Dropout).
  - Traformare le feature map da SignHIeria con **la stessa dimensione e distribuzione statistica attesa dallo spazio latente del modello di traduzione successivo**.
- **Translation Model (NLLB-200)**, un modello *Sequence-to-Sequence* pre-addestrato massivamente *su dati multilingue*.
  - **l'encoder** di NLLB accetta direttamente gli **embedding visivi** adattati, come se fossero parole di una lingua straniera.
  - **Il decoder** genera la **traduzione testuale** nella lingua target.
- **Pooling SONAR (testo e video)** vengono **mappate** in un **unico spazio vettoriale**.
  - **l'output dell'encoder** viene **compresso in un singolo vettore (Mean Pooling).**

Note Mod. 3

- **VADER:** è un sistema **rule base** che in base alla struttura **sintatti**, **grammaticale** e **parole chiavi**, resistuisce un valore **[-1,1]** rappresentati il sentiment

Note Mod. 3

- **BARK:** costituito da tre moduli fondamentali
  1. Primo modulo che comprende la **struttura semantica** e **sintattica** della frase osservando e **estraendo i token** (inizia a capire come potrebbe venire l'audio)
  2. Secondo modulo prende in input cosa fatto dal module precedente e **genera** i **token audio per singole parole**
  3. L'ultimo modulo, in base alle analisi fatte in precedenza (dai due moduli) **genera l'audio effettivo**, con dovute **intonazioni**, **pause**, **token paralinguistici** ecc