MQM Translation Quality Report

Project: TranslateGemma_EN-IT_Eval1

SCORE

98.9 %

FAILED

PER 1K TOKENS: 11.1

PER 1K WORDS: 32.5

PENALTY TOTAL: **23**

7	708	4986	2066	7	N/A
Total Errors	Total Words	Total Char.	Total Tokens	Segments	Duration

Severity Breakdown

Critical	0
Major	4
Minor	3

Top Categories

Accuracy/Mistranslation	3
Accuracy/Omission	2
Fluency/Grammar	1
Terminology	1

Scoring Formula

Penalties are normalized by translation length in **tokens** (XLM-R SentencePiece). This ensures better unification across languages (e.g. for CJK languages) compared to word counts.

Total Penalty = Σ (*Error count* × *Error weight*)

Score (%) = (1 – *Total Penalty* ÷ *Total tokens*) × 100

Pass if Score ≥ **99.0 %** , *otherwise Fail.*

Error Weights

Penalties are subtracted from the score based on error severity:

Critical	25	pts
Major	5	pts
Minor	1	pts

Detailed Error Log

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Major	Fluency/Grammar *"Invoca" is a verb form and is grammatically incorrect as a noun in Italian. Correct version is "Invocazione selettiva"*	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty Invoca selettiva per il riconoscimento vocale multilingue: un approccio economicamente vantaggioso che si adatta alla difficoltà del riconoscimento vocale.	5	0.24%
Major	Accuracy/Mistranslation *"Dati multilingue etichettati in modo parziale" is a semantic misinterpretation of weakly multilingual data. The correct concept is weak supervision, not partial labeling. Correct: "680.000 ore di dati multilingue debolmente supervisionati"*	Multilingual automatic speech recognition (ASR) models have gained significant attention for their ability to recognize multiple languages using a single model [1, 2, 3, 4], as illustrated in Figure 1(a). Recent advances have led to impressive performance in various languages through large-scale supervised or self-supervised pre-training [3, 5, 6, 7, 8, 9, 10, 11, 12]. For example, Whisper [6] is trained on 680,000 hours of weakly multilingual data, enabling it to generalize effectively across standard ASR benchmarks, while USM [9] leverages 12 million hours of unlabeled data to achieve robust cross-lingual performance. Despite these advances, the application of multilingual ASR systems with a single model still faces significant challenges. Phonetic differences, syntactic variations, and vocabulary disparities across languages make it difficult to achieve consistent universal state-of-the-art (SOTA) performance. Moreover, imbalances in training data between high-resource and low-resource languages further limit the single-model solutions. I modelli automatici di riconoscimento vocale multilingue (ASR) hanno	5	0.24%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		suscitato notevole interesse per la loro capacità di riconoscere più lingue utilizzando un unico modello [1, 2, 3, 4], come illustrato in Figura 1(a). I recenti progressi hanno portato a prestazioni notevoli in diverse lingue grazie all'utilizzo di tecniche di pre-addestramento supervisionato o auto-supervisionato su larga scala [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ad esempio, Whisper [6] è stato addestrato su 680.000 ore di dati multilingue etichettati in modo parziale, il che gli consente di generalizzare efficacemente su standard di riferimento per l'ASR, mentre USM [9] sfrutta 12 milioni di ore di dati non etichettati per ottenere prestazioni robuste tra diverse lingue. Nonostante questi progressi, l'applicazione di sistemi ASR multilingue con un unico modello presenta ancora sfide significative. Le differenze fonetiche, le variazioni sintattiche e le disparità nel vocabolario tra le lingue rendono difficile ottenere prestazioni universalmente all'avanguardia (SOTA) costanti. Inoltre, gli squilibri nei dati di addestramento tra le lingue con molte risorse e quelle con poche risorse limitano ulteriormente le soluzioni basate su un unico modello.		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Major	Accuracy/Mistranslation *“Riduce i costi di circa lo 0,51 volte” is mathematically and scientifically incorrect in Italian. The original meaning is a reduction to 0.51×, not by 0.51×. Correct: "riduce i costi di elaborazione a circa 0,51 volte su tutti e tre i dataset"*	The invocation decision accuracy (ACC) and F1 scores are approximately 70%, supporting our hypothesis that SLLMs can effectively differentiate speech inputs based on complexity. Although SIMA exhibits a slight WER gap compared to LID-Top, it reduces invocation costs by approximately 0.51× across the three datasets, significantly lowering associated expenses. L'accuratezza (ACC) e i punteggi F1 relativi alla capacità di identificazione sono di circa il 70%, confermando la nostra ipotesi che i modelli linguistici di grandi dimensioni (LLM) possano distinguere efficacemente gli input vocali in base alla loro complessità. Sebbene SIMA mostri una leggera differenza nel tasso di errore di riconoscimento (WER) rispetto a LID-Top, riduce i costi di elaborazione di circa lo 0,51 volte nei tre dataset, diminuendo significativamente le spese associate.	5	0.24%
Major	Accuracy/Mistranslation *The translation incorrectly states that the SLLM performs the speech recognition task. In the source text, it performs the invocation decision task. This is a major conceptual error. Correct: "Questa limitazione deriva dalla nostra ipotesi iniziale che il modello SLLM di base possa svolgere efficacemente il compito di invocazione."*	Although the current SIMA model significantly improves WER, it still lags behind Whisper [6] on out-of-domain data, FLEURS [28]. This limitation stems from our initial hypothesis that the base SLLM model can effectively perform the invoke task. Our base SLLM model [29] is inherently weaker than specialized models such as Whisper because of the limitation of training data. In future work, we plan to adopt Whisper [6] as the base model and further refine the SIMA system to improve the ASR performance of the SOTA model.	5	0.24%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		Sebbene il modello SIMA attuale migliori significativamente il tasso di riconoscimento vocale (WER), rimane comunque inferiore a Whisper [6] quando si tratta di dati provenienti da domini diversi, come dimostrato da FLEURS [28]. Questa limitazione deriva dalla nostra ipotesi iniziale che il modello SLLM di base possa svolgere efficacemente il compito di riconoscimento vocale. Il nostro modello SLLM di base [29] è intrinsecamente meno performante rispetto a modelli specializzati come Whisper a causa delle limitazioni dei dati di addestramento. In futuro, prevediamo di adottare Whisper [6] come modello di base e di perfezionare ulteriormente il sistema SIMA per migliorare le prestazioni del modello all'avanguardia (SOTA) nel riconoscimento vocale.		
Minor	Terminology *“segnale vocale” introduces an unnecessary technical narrowing. The source text refers to speech input, not explicitly to the signal level. Correct: "parlato"*	Motivated by these limitations, we propose an alternative strategy that selectively invokes models based on the complexity of the input speech. In ASR tasks, the recognition difficulty varies significantly. Under clean acoustic conditions with simple vocabulary, both the SOTA and regular models typically yield low word error rates (WER). However, in noisy or acoustically challenging environments, the WER increases [14, 15, 16, 17], where robust SOTA models generally perform better [6]. This observation raises a key question: Can we distinguish	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		between simple and complex speech inputs and adapt our ASR system accordingly? Motivati da queste limitazioni, proponiamo una strategia alternativa che seleziona i modelli in base alla complessità del segnale vocale in ingresso. Nelle attività di riconoscimento vocale automatico (ASR), il livello di difficoltà del riconoscimento varia notevolmente. In condizioni acustiche favorevoli e con un vocabolario semplice, sia i modelli all'avanguardia (SOTA) che quelli standard tendono a produrre bassi tassi di errore di parola (WER). Tuttavia, in ambienti rumorosi o con condizioni acustiche difficili, il WER aumenta [14, 15, 16, 17], mentre i modelli SOTA più robusti generalmente offrono prestazioni migliori [6]. Questa osservazione solleva una domanda fondamentale: possiamo distinguere tra segnali vocali semplici e complessi e adattare di conseguenza il nostro sistema ASR?		
Minor	Accuracy/Omission *Missing "specializzati" (S in SLLM)*	The invocation decision accuracy (ACC) and F1 scores are approximately 70%, supporting our hypothesis that SLLMs can effectively differentiate speech inputs based on complexity. Although SIMA exhibits a slight WER gap compared to LID-Top, it reduces invocation costs by approximately 0.51× across the three datasets, significantly lowering associated expenses. L'accuratezza (ACC) e i punteggi F1 relativi alla capacità di identificazione	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		sono di circa il 70%, confermando la nostra ipotesi che i modelli linguistici di grandi dimensioni (LLM) possano distinguere efficacemente gli input vocali in base alla loro complessità. Sebbene SIMA mostri una leggera differenza nel tasso di errore di riconoscimento (WER) rispetto a LID-Top, riduce i costi di elaborazione di circa lo 0,51 volte nei tre dataset, diminuendo significativamente le spese associate.		
Minor	Accuracy/Omission *missing "errore"*	Although the current SIMA model significantly improves WER, it still lags behind Whisper [6] on out-of-domain data, FLEURS [28]. This limitation stems from our initial hypothesis that the base SLLM model can effectively perform the invoke task. Our base SLLM model [29] is inherently weaker than specialized models such as Whisper because of the limitation of training data. In future work, we plan to adopt Whisper [6] as the base model and further refine the SIMA system to improve the ASR performance of the SOTA model. Sebbene il modello SIMA attuale migliori significativamente il tasso di riconoscimento vocale (WER), rimane comunque inferiore a Whisper [6] quando si tratta di dati provenienti da domini diversi, come dimostrato da FLEURS [28]. Questa limitazione deriva dalla nostra ipotesi iniziale che il modello SLLM di base possa svolgere	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		efficacemente il compito di riconoscimento vocale. Il nostro modello SLLM di base [29] è intrinsecamente meno performante rispetto a modelli specializzati come Whisper a causa delle limitazioni dei dati di addestramento. In futuro, prevediamo di adottare Whisper [6] come modello di base e di perfezionare ulteriormente il sistema SIMA per migliorare le prestazioni del modello all'avanguardia (SOTA) nel riconoscimento vocale.		