MQM Translation Quality Report

Project: TranslateGemma_EN-DE_Eval1

SCORE

98.9 %

FAILED

PER 1K TOKENS	PER 1K WORDS	PENALTY TOTAL
11.0	38.0	23

11	605	4902	2095	7	N/A
Total Errors	Total Words	Total Char.	Total Tokens	Segments	Duration

Severity Breakdown

Critical	0
Major	3
Minor	8

Top Categories

Accuracy/Mistranslation	8
Accuracy/Omission	2
Style	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	Accuracy/Omission	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. Mehrsprachige automatische Spracherkennungssysteme (ASR) haben aufgrund ihrer Fähigkeit, mehrere Sprachen mit einem einzigen Modell zu erkennen, große Aufmerksamkeit erlangt [1, 2, 3, 4], wie in Abbildung 1(a) dargestellt. Jüngste Fortschritte haben zu beeindruckenden Ergebnissen in verschiedenen Sprachen durch großflächiges, überwachtes oder selbstüberwachtes Vortraining geführt [3, 5, 6, 11, 12]. Beispielsweise wurde Whisper	5	0.24%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		[6] mit 680.000 Stunden schwach mehrsprachiger Daten trainiert, wodurch es effektiv über gängige ASR-Benchmarks generalisieren kann. USM [9] nutzt hingegen 12 Millionen Stunden unbeschrifteter Daten, um eine robuste, sprachübergreifende Leistung zu erzielen. Trotz dieser Fortschritte stehen die Anwendung von mehrsprachigen ASR-Systemen mit einem einzigen Modell weiterhin vor erheblichen Herausforderungen. Phonetische Unterschiede, syntaktische Variationen und Unterschiede im Vokabular zwischen den Sprachen erschweren es, eine konsistente, universelle Spitzenleistung (State-of-the-Art, SOTA) zu erreichen. Darüber hinaus schränken Ungleichgewichte in den Trainingsdaten zwischen Sprachen mit vielen und wenigen Ressourcen die Lösungen mit einem einzigen Modell weiter ein.		
Major	Accuracy/Mistranslation *des gezielten Aufrufs*	The results indicate that, due to the selective invocation of SOTA models, the SIMA model achieves significant WER reductions of 18.6%, 9.3%, and 28.2% relative to the base model on the three datasets. Furthermore, compared to the random invocation strategy, SIMA consistently delivers lower WER, with improvements of 6.6%, 4.2%, and 16.8%. Notably, the improvement on the FLEURS dataset is especially significant, as it is out-of-domain for the base model but in-domain for the LID-Top model. These findings convincingly demonstrate SIMA's remarkable ability to precisely determine when to invoke the	5	0.24%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		SOTA model, thereby optimizing overall ASR performance. Die Ergebnisse zeigen, dass das SIMA-Modell, aufgrund der gezielten Auswahl von modernsten Modellen, im Vergleich zum Basismodell signifikante Reduktionen der Wortfehlerrate (WER) von 18,6 %, 9,3 % und 28,2 % auf den drei Datensätzen erzielt. Darüber hinaus liefert SIMA im Vergleich zu einer zufälligen Auswahlstrategie durchgängig eine niedrigere WER, mit Verbesserungen von 6,6 %, 4,2 % und 16,8 %. Besonders bemerkenswert ist die Verbesserung auf dem FLEURS-Datensatz, da dieser für das Basismodell außerhalb des Anwendungsbereichs liegt, während er für das LID-Top-Modell innerhalb des Anwendungsbereichs liegt. Diese Ergebnisse belegen überzeugend die bemerkenswerte Fähigkeit von SIMA, präzise zu bestimmen, wann das modernste Modell eingesetzt werden soll, wodurch die Gesamtleistung der automatischen Spracherkennung optimiert wird.		
Major	Accuracy/Mistranslation	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
		Obwohl das aktuelle SIMA-Modell die Wort-Fehlerquote (WER) deutlich verbessert, liegt es bei Daten, die nicht zum Trainingsdatensatz gehören, immer noch hinter Whisper [6], insbesondere bei den FLEURS-Daten [28]. Diese Einschränkung ergibt sich aus unserer ursprünglichen Hypothese, dass das Basis-SLLM-Modell die Aufgabe der Spracherkennung effektiv ausführen kann. Unser Basis-SLLM-Modell [29] ist aufgrund der Beschränkungen der Trainingsdaten grundsätzlich schwächer als spezialisierte Modelle wie Whisper. In zukünftigen Arbeiten planen wir, Whisper [6] als Basismodell zu verwenden und das SIMA-System weiter zu verfeinern, um die Spracherkennungsleistung des aktuell besten Modells (State-of-the-Art) zu verbessern.		
Minor	Accuracy/Mistranslation	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty Selektive Aktivierung für mehrsprachige Spracherkennung: Ein kostengünstiger Ansatz zur Anpassung an unterschiedliche Schwierigkeitsgrade der Spracherkennung.	1	0.05%
Minor	Accuracy/Omission *Spracherkennungssysteme*	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty Selektive Aktivierung für mehrsprachige	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		Spracherkennung: Ein kostengünstiger Ansatz zur Anpassung an unterschiedliche Schwierigkeitsgrade der Spracherkennung.		
Minor	Accuracy/Mistranslation	A common strategy to address these challenges is to use a language identification (LID) model that first detects the language of the input speech before invoking the corresponding SOTA ASR model for transcription, as shown in Figure 1(b). However, this two-stage approach has its drawbacks. Many SOTA models are commercial [12] and incur usage fees based on the volume of processing, making this method costly. Additionally, an incorrect LID prediction may trigger the wrong model, further affecting the user experience [13]. Eine gängige Strategie zur Bewältigung dieser Herausforderungen besteht darin, ein Sprachidentifikationsmodell (LID) zu verwenden, das zunächst die Sprache der eingegebenen Sprache erkennt, bevor das entsprechende, modernste Spracherkennungssystem (ASR) für die Transkription aufgerufen wird, wie in Abbildung 1(b) dargestellt. Dieser zweistufige Ansatz hat jedoch seine Nachteile. Viele der modernsten Modelle sind kommerziell [12] und verursachen Nutzungsgebühren, die sich nach dem Verarbeitungsvolumen richten, was diese Methode kostspielig macht. Darüber hinaus kann eine falsche LID-Vorhersage dazu führen, dass das falsche Modell aktiviert wird, was die	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		Benutzererfahrung weiter beeinträchtigen kann [13].		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Minor	Accuracy/Mistranslation *Spracherkennungsaufgaben*	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 between simple and complex speech inputs and adapt our ASR system accordingly? Angesichts dieser Einschränkungen schlagen wir eine alternative Strategie vor, die Modelle selektiv einsetzt, basierend auf der Komplexität der eingegebenen Sprache. Bei Spracherkennungstätigkeiten variiert der Schwierigkeitsgrad der Erkennung erheblich. Unter sauberen akustischen Bedingungen mit einfachem Vokabular liefern sowohl die modernsten als auch die Standardmodelle typischerweise niedrige Wortfehlerraten (WER). Allerdings steigt die WER in lauten oder akustisch anspruchsvollen Umgebungen [14, 15, 16, 17], wobei robuste, modernste Modelle in der Regel besser abschneiden [6]. Diese Beobachtung wirft eine wichtige Frage auf: Können wir zwischen einfachen und komplexen Spracheingaben unterscheiden und unser Spracherkennungssystem entsprechend anpassen?	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Minor	Accuracy/Mistranslation *Eingangssprache*	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 between simple and complex speech inputs and adapt our ASR system accordingly? Angesichts dieser Einschränkungen schlagen wir eine alternative Strategie vor, die Modelle selektiv einsetzt, basierend auf der Komplexität der eingegebenen Sprache. Bei Spracherkennungstätigkeiten variiert der Schwierigkeitsgrad der Erkennung erheblich. Unter sauberen akustischen Bedingungen mit einfachem Vokabular liefern sowohl die modernsten als auch die Standardmodelle typischerweise niedrige Wortfehlerraten (WER). Allerdings steigt die WER in lauten oder akustisch anspruchsvollen Umgebungen [14, 15, 16, 17], wobei robuste, modernste Modelle in der Regel besser abschneiden [6]. Diese Beobachtung wirft eine wichtige Frage auf: Können wir zwischen einfachen und komplexen Spracheingaben unterscheiden und unser Spracherkennungssystem entsprechend anpassen?	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Minor	Accuracy/Mistranslation *Befehlsaufrufe*	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. Die Genauigkeit (ACC) und die F1-Werte bei der Erkennung von Sprachbefehlen liegen bei etwa 70 %, was unsere Hypothese unterstützt, dass große Sprachmodelle (SLLMs) Spracheingaben aufgrund ihrer Komplexität effektiv unterscheiden können. Obwohl SIMA im Vergleich zu LID-Top einen geringfügigen Unterschied in der Wortfehlerrate (WER) aufweist, reduziert es die Kosten für die Befehlserkennung um etwa das 0,51-fache über die drei Datensätze, was die damit verbundenen Ausgaben erheblich senkt.	1	0.05%
Minor	Accuracy/Mistranslation *Genauigkeit der Aufrufentscheidung*	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. Die Genauigkeit (ACC) und die F1-Werte bei der Erkennung von Sprachbefehlen liegen bei etwa 70 %, was unsere Hypothese unterstützt, dass große Sprachmodelle (SLLMs) Spracheingaben aufgrund	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		ihrer Komplexität effektiv unterscheiden können. Obwohl SIMA im Vergleich zu LID-Top einen geringfügigen Unterschied in der Wortfehlerrate (WER) aufweist, reduziert es die Kosten für die Befehlserkennung um etwa das 0,51-fache über die drei Datensätze, was die damit verbundenen Ausgaben erheblich senkt.		
Minor	Style *außerhalb des Trainingsdatensatzes*	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. Obwohl das aktuelle SIMA-Modell die Wort-Fehlerquote (WER) deutlich verbessert, liegt es bei Daten, die nicht zum Trainingsdatensatz gehören, immer noch hinter Whisper [6], insbesondere bei den FLEURS-Daten [28]. Diese Einschränkung ergibt sich aus unserer ursprünglichen Hypothese, dass das Basis-SLLM-Modell die Aufgabe der Spracherkennung effektiv ausführen kann. Unser Basis-SLLM-Modell [29] ist aufgrund der Beschränkungen der Trainingsdaten grundsätzlich schwächer als spezialisierte Modelle wie Whisper. In zukünftigen Arbeiten planen wir, Whisper	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		[6] als Basismodell zu verwenden und das SIMA-System weiter zu verfeinern, um die Spracherkennungsleistung des aktuell besten Modells (State-of-the-Art) zu verbessern.		