

MQM Translation Quality Report

Project: TranslateGemma_EN-JA_Eval1

SCORE

87.8 %

⚠️ FAILED

PER 1K TOKENS

122.1

PER 1K WORDS

125.8

PENALTY

TOTAL

244

⚠️

28

Total Errors

📊

1940

Total Words

📊

1940

Total Char.

📊

1999

Total Tokens

✅

7

Segments

🕒

N/A

Duration

📊 Severity Breakdown

🔴 Critical

7

🟠 Major

12

🟡 Minor

9

📊 Top Categories

Accuracy/Mistranslation

10

Accuracy/Omission

6

Fluency/Register

5

Fluency/Grammar

4

Terminology

3

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.

$$Total\ Penalty = \sum (Error\ count \times Error\ weight)$$
$$Score\ (\%) = (1 - Total\ Penalty \div Total\ tokens) \times 100$$

Error Weights

Penalties are subtracted from the score based on error severity:

🔴 Critical

25

pts

🟠 Major

5

pts

🟡 Minor

1

pts

Pass if Score ≥ 99.0 % , otherwise Fail.

Detailed Error Log

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Critical	Accuracy/Omission "[1, 2, 3, 4]" after this phrase has been omitted.	<p>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.</p> <p>多言語自動音声認識（ASR）モデルは、単一のモデルで複数の言語を認識できるという能力から、近年注目を集めています（図1(a)を参照）。最近の進歩により、大規模な教師あり学習や自己教師あり学習による事前学習を通じて、様々な言語で優れた性能が実現されています [3, 5, 6, 7, 10, 11, 12]。例えば、Whisper[6]は、68万時間の多言語データで学習されており、標準的なASRベンチマークにおいて高い汎化性能を発揮</p>	25	1.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		します。一方、USM[9]は、1200万時間のラベルなしデータを利用することで、堅牢な多言語性能を実現しています。しかしながら、単一モデルを用いた多言語ASRシステムの応用は、依然として大きな課題を抱えています。言語間の音声的差異、構文の多様性、語彙の相違などが、一貫した最高の性能（SOTA）を達成することを困難にしています。さらに、リソースが豊富な言語と少ない言語間の学習データにおける不均衡が、単一モデルの解決策を制限する要因となっています。		
Critical	Accuracy/Omission <i>8 and 9 have been omitted.</i>	<p>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.</p> <p>多言語自動音声認識（ASR）モデルは、単一のモデルで複数</p>	25	1.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>の言語を認識できるという能力から、近年注目を集めています（図1(a)を参照）。最近の進歩により、大規模な教師あり学習や自己教師あり学習による事前学習を通じて、様々な言語で優れた性能が実現されています [3, 5, 6, 7, 10, 11, 12]。例えば、Whisper[6]は、68万時間の多言語データで学習されており、標準的なASRベンチマークにおいて高い汎化性能を発揮します。一方、USM[9]は、1200万時間のラベルなしデータを利用することで、堅牢な多言語性能を実現しています。しかしながら、単一モデルを用いた多言語ASRシステムの応用は、依然として大きな課題を抱えています。言語間の音声的差異、構文の多様性、語彙の相違などが、一貫した最高の性能（SOTA）を達成することを困難にしています。さらに、リソースが豊富な言語と少ない言語間の学習データにおける不均衡が、単一モデルの解決策を制限する要因となっています。</p>		
Critical	<p>Accuracy/Mistranslation</p> <p><i>"ASR" is an abbreviation for "Automated Speech Recognition." Since it does not include the word "モデル (model)," it should not arbitrarily change the word order. And, "Automated" has been omitted.</i></p>	<p>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].</p> <p>これらの課題に対処するための一般的な手法として、まず入力音声の言語を識別する言語識別モデル（LID）を使用し、その</p>	25	1.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		後、対応する最先端の音声認識モデル（ASR）を呼び出して文字起こしを行う方法があります（図1(b)を参照）。しかし、この二段階のアプローチにはいくつかの欠点があります。多くの最先端モデルは商用であり[12]、処理量に応じて利用料金が発生するため、この方法はコストがかかります。さらに、言語識別の予測が誤っている場合、誤ったモデルが選択され、ユーザーエクスペリエンスに悪影響を及ぼす可能性があります[13]。		
Critical	Accuracy/Mistranslation <i>"invocation decision" is translated as "呼び出しの判断," but this has been mistakenly rendered as "音声認識 (Speech Recognition)." This is a critical error.</i>	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. 音声認識の精度（ACC）とF1スコアは、いずれも約70%であり、これは、大規模言語モデル（SLLM）が、音声の複雑さに基づいて効果的に区別できるという仮説を支持するものです。SIMAは、LID-Topと比較してわずかに単語誤り率（WER）が高いものの、3つのデータセット全体で、呼び出しにかかるコストを約0.51倍に削減し、関連する費用を大幅に低減します。	25	1.25%
Critical	Accuracy/Mistranslation <i>"に" means "to," not "by." This may lead to misunderstanding.</i>	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	25	1.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>approximately 0.51× across the three datasets, significantly lowering associated expenses.</p> <p>音声認識の精度（ACC）とF1スコアは、いずれも約70%であり、これは、大規模言語モデル（SLLM）が、音声の複雑さに基づいて効果的に区別できるという仮説を支持するものです。SIMAは、LID-Topと比較してわずかに単語誤り率（WER）が高いものの、3つのデータセット全体で、呼び出しにかかるコストを約0.51倍に削減し、関連する費用を大幅に低減します。</p>		
Critical	Accuracy/Mistranslation <i>This translation engine arbitrarily interprets it as "高い (high)," but the original text only states "わずかな差がある (exhibits a slight gap)". "This may lead to misunderstanding.</i>	<p>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.</p> <p>音声認識の精度（ACC）とF1スコアは、いずれも約70%であり、これは、大規模言語モデル（SLLM）が、音声の複雑さに基づいて効果的に区別できるという仮説を支持するものです。SIMAは、LID-Topと比較してわずかに単語誤り率（WER）が高いものの、3つのデータセット全体で、呼び出しにかかるコストを約0.51倍に削減し、関連する費用を大幅に低減します。</p>	25	1.25%
Critical	Accuracy/Mistranslation <i>The relationship between "out-of-domain data" and "FLEURS" is not a parallel connection linked by "や (and)," but rather "～である～ (as)."</i>	<p>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</p>	25	1.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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.</p> <p>現在のSIMAモデルは、単語認識精度（WER）において大幅な改善が見られますが、依然としてWhisper [6]に比べて、学習データとは異なるデータセット（アウトオブドメインデータ）やFLEURS [28]においては性能が劣ります。この制限は、当初の仮説である「ベースのSLLMモデルが、特定のタスクを効果的に実行できる」という考え方に基づいています。弊社のベースとなるSLLMモデル[29]は、学習データの制約から、Whisperのような専門的なモデルに比べて、本来的に性能が劣ります。今後の研究では、Whisper [6]をベースモデルとして採用し、SIMAシステムをさらに改良することで、最先端モデル（SOTA）の音声認識性能を向上させることを計画しています。</p>		
Major	Accuracy/Omission <i>"Automated" has been omitted. "多言語自動音声認識" is more accurate.</i>	<p>Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty</p> <p>多言語音声認識における選択的呼び出し：音声認識の難易度に合わせて最適化する、費用対効果の高い手法。</p>	5	0.25%
Major	Terminology <i>In this context, "統語的バリエーション" is the appropriate translation.</i>	<p>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</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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.</p> <p>多言語自動音声認識（ASR）モデルは、単一のモデルで複数の言語を認識できるという能力から、近年注目を集めています（図1(a)を参照）。最近の進歩により、大規模な教師あり学習や自己教師あり学習による事前学習を通じて、様々な言語で優れた性能が実現されています [3, 5, 6, 7, 10, 11, 12]。例えば、Whisper[6]は、68万時間の多言語データで学習されており、標準的なASRベンチマークにおいて高い汎化性能を発揮します。一方、USM[9]は、1200万時間のラベルなしデータを利用することで、堅牢な多言語性能を実現しています。しかしながら、単一モデルを用いた多言語ASRシステムの応用は、依然として大きな課題を抱えています。言語間の音声的差異、構文の多様性、語彙の相違などが、一貫した最高の性能（SOTA）を達成することを困難にしています。さらに、リソースが豊富な言語と少ない言語</p>		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		間の学習データにおける不均衡が、単一モデルの解決策を制限する要因となっています。		
Major	Accuracy/Mistranslation <i>"LID" is an abbreviation for "language identification." Since it does not include the word "model," it should not arbitrarily change the word order.</i>	<p>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].</p> <p>これらの課題に対処するための一般的な手法として、まず入力音声の言語を識別する言語識別モデル (LID) を使用し、その後、対応する最先端の音声認識モデル (ASR) を呼び出して文字起こしを行う方法があります (図1(b)を参照)。しかし、この二段階のアプローチにはいくつかの欠点があります。多くの最先端モデルは商用であり [12]、処理量に応じて利用料金が発生するため、この方法はコストがかかります。さらに、言語識別の予測が誤っている場合、誤ったモデルが選択され、ユーザーエクスペリエンスに悪影響を及ぼす可能性があります [13]。</p>	5	0.25%
Major	Accuracy/Mistranslation <i>"trigger" should be translated as "トリガーする".</i>	<p>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</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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].</p> <p>これらの課題に対処するための一般的な手法として、まず入力音声の言語を識別する言語識別モデル（LID）を使用し、その後、対応する最先端の音声認識モデル（ASR）を呼び出して文字起こしを行う方法があります（図1(b)を参照）。しかし、この二段階のアプローチにはいくつかの欠点があります。多くの最先端モデルは商用であり[12]、処理量に応じて利用料金が発生するため、この方法はコストがかかります。さらに、言語識別の予測が誤っている場合、誤ったモデルが選択され、ユーザーエクスペリエンスに悪影響を及ぼす可能性があります[13]。</p>		
Major	Accuracy/Omission <i>"Automated" has been omitted.</i>	<p>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?</p> <p>これらの制約から、私たちは、入力音声の複雑さに応じてモデルを適切に選択する、別の戦略を提案します。音声認識</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>(ASR) のタスクにおいて、認識の難易度は大きく異なります。音声環境がクリアで、語彙が単純な場合、最先端 (SOTA) モデルと一般的なモデルの両方で、通常は低い単語誤り率 (WER) が得られます。しかし、騒音が多い環境や、音響的に困難な環境では、WERが増加します[14, 15, 16, 17]。そのような状況では、堅牢な最先端モデルの方が一般的に優れた性能を発揮します[6]。このことから、重要な疑問が生まれます。私たちは、単純な音声と複雑な音声の区別をつけ、それに応じてASRシステムを適応させることができるでしょうか？</p>		
Major	<p>Terminology</p> <p><i>This is not an industry-standard expression.</i></p>	<p>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 SOTA model, thereby optimizing overall ASR performance.</p> <p>その結果から、SIMAモデルは、最先端モデル (SOTAモデル) を適切に選択的に利用することで、ベースモデルと比較して、3つのデータセットでそれぞれ18.6%、9.3%、28.2%という大幅な単語認識エラー率 (WER) の削減を達成しました。さらに、ランダムなモデル</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>選択戦略と比較して、SIMAモデルは一貫して低いWERを示し、それぞれ6.6%、4.2%、16.8%の改善が見られました。特に、FLEURSデータセットにおける改善は顕著であり、これはベースモデルにとっては未知の領域であるのに対し、LID-Topモデルにとっては学習対象の領域であるためです。これらの結果は、SIMAモデルが、最先端モデルをいつ利用すべきかを正確に判断する優れた能力を持っていることを明確に示しており、これにより、全体的な音声認識性能が最適化されます。</p>		
Major	<p>Accuracy/Mistranslation</p> <p><i>The correct translation is "専門外の," not "未知の."</i></p>	<p>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 SOTA model, thereby optimizing overall ASR performance.</p> <p>その結果から、SIMAモデルは、最先端モデル（SOTAモデル）を適切に選択的に利用することで、ベースモデルと比較して、3つのデータセットでそれぞれ18.6%、9.3%、28.2%という大幅な単語認識エラー率（WER）の削減を達成しました。さらに、ランダムなモデル選択戦略と比較して、SIMAモデルは一貫して低いWERを示</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		し、それぞれ6.6%、4.2%、16.8%の改善が見られました。特に、FLEURSデータセットにおける改善は顕著であり、これはベースモデルにとっては未知の領域であるのに対し、LID-Topモデルにとっては学習対象の領域であるためです。これらの結果は、SIMAモデルが、最先端モデルをいつ利用すべきかを正確に判断する優れた能力を持っていることを明確に示しており、これにより、全体的な音声認識性能が最適化されます。		
Major	Accuracy/Omission <i>"Automated" has been omitted.</i>	<p>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 SOTA model, thereby optimizing overall ASR performance.</p> <p>その結果から、SIMAモデルは、最先端モデル（SOTAモデル）を適切に選択的に利用することで、ベースモデルと比較して、3つのデータセットでそれぞれ18.6%、9.3%、28.2%という大幅な単語認識エラー率（WER）の削減を達成しました。さらに、ランダムなモデル選択戦略と比較して、SIMAモデルは一貫して低いWERを示し、それぞれ6.6%、4.2%、16.8%の改善が見られまし</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>た。特に、FLEURSデータセットにおける改善は顕著であり、これはベースモデルにとっては未知の領域であるのに対し、LID-Topモデルにとっては学習対象の領域であるためです。これらの結果は、SIMAモデルが、最先端モデルをいつ利用すべきかを正確に判断する優れた能力を持っていることを明確に示しており、これにより、全体的な 音声認識 性能が最適化されます。</p>		
Major	Terminology <i>This is not an industry-standard expression.</i>	<p>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.</p> <p>現在のSIMAモデルは、単語認識精度（WER）において大幅な改善が見られますが、依然としてWhisper [6]に比べて、学習データとは異なるデータセット（アウトオブドメインデータ）やFLEURS [28]においては性能が劣ります。この制限は、当初の仮説である「ベースのSLLMモデルが、特定のタスクを効果的に実行できる」という考え方に基づいています。弊社のベースとなるSLLMモデル [29] は、学習データの制約から、Whisperのような専門的なモデルに比べて、本来的に性能が劣ります。今後の研究では、Whisper [6]をベースモデルとして採用し、SIMAシステ</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		ムをさらに改良することで、最先端モデル（SOTA）の音声認識性能を向上させることを計画しています。		
Major	Accuracy/Mistranslation	<p>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.</p> <p>現在のSIMAモデルは、単語認識精度（WER）において大幅な改善が見られますが、依然としてWhisper [6]に比べて、学習データとは異なるデータセット（アウトオブドメインデータ）やFLEURS [28]においては性能が劣ります。この制限は、当初の仮説である「ベースのSLLMモデルが、特定のタスクを効果的に実行できる」という考え方に基づいています。弊社のベースとなるSLLMモデル [29] は、学習データの制約から、Whisperのような専門的なモデルに比べて、本来的に性能が劣ります。今後の研究では、Whisper [6]をベースモデルとして採用し、SIMAシステムをさらに改良することで、最先端モデル（SOTA）の音声認識性能を向上させることを計画しています。</p>	5	0.25%
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	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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.</p> <p>現在のSIMAモデルは、単語認識精度（WER）において大幅な改善が見られますが、依然としてWhisper [6]に比べて、学習データとは異なるデータセット（アウトオブドメインデータ）やFLEURS [28]においては性能が劣ります。この制限は、当初の仮説である「ベースのSLLMモデルが、特定のタスクを効果的に実行できる」という考え方に基づいています。弊社のベースとなるSLLMモデル[29]は、学習データの制約から、Whisperのような専門的なモデルに比べて、本来的に性能が劣ります。今後の研究では、Whisper [6]をベースモデルとして採用し、SIMAシステムをさらに改良することで、最先端モデル（SOTA）の音声認識性能を向上させることを計画しています。</p>		
Major	Accuracy/Omission	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	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>performance of the SOTA model.</p> <p>現在のSIMAモデルは、単語認識精度（WER）において大幅な改善が見られますが、依然としてWhisper [6]に比べて、学習データとは異なるデータセット（アウトオブドメインデータ）やFLEURS [28]においては性能が劣ります。この制限は、当初の仮説である「ベースのSLLMモデルが、特定のタスクを効果的に実行できる」という考え方に基づいています。弊社のベースとなるSLLMモデル [29] は、学習データの制約から、Whisperのような専門的なモデルに比べて、本来的に性能が劣ります。今後の研究では、Whisper [6]をベースモデルとして採用し、SIMAシステムをさらに改良することで、最先端モデル（SOTA）の音声認識性能を向上させることを計画しています。</p>		
Minor	Fluency/Grammar <i>In this case, "される" is more natural.</i>	<p>Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty</p> <p>多言語音声認識における選択的呼び出し：音声認識の難易度に合わせて最適化する、費用対効果の高い手法。</p>	1	0.05%
Minor	Fluency/Register <i>Translating "performance" as "性能" is unnatural in this context.</i>	<p>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,</p>	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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.</p> <p>多言語自動音声認識（ASR）モデルは、単一のモデルで複数の言語を認識できるという能力から、近年注目を集めています（図1(a)を参照）。最近の進歩により、大規模な教師あり学習や自己教師あり学習による事前学習を通じて、様々な言語で優れた性能が実現されています [3, 5, 6, 7, 10, 11, 12]。例えば、Whisper[6]は、68万時間の多言語データで学習されており、標準的なASRベンチマークにおいて高い汎化性能を発揮します。一方、USM[9]は、1200万時間のラベルなしデータを利用することで、堅牢な多言語性能を実現しています。しかしながら、単一モデルを用いた多言語ASRシステムの応用は、依然として大きな課題を抱えています。言語間の音声的差異、構文の多様性、語彙の相違などが、一貫した最高の性能（SOTA）を達成することを困難にしています。さらに、リソースが豊富な言語と少ない言語間の学習データにおける不均衡が、単一モデルの解決策を制限する要因となっています。</p>		
Minor	Fluency/Grammar <i>The expression "学習される" sounds unnatural in this</i>	Multilingual automatic speech recognition (ASR) models have gained significant attention for	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
	context.	<p>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.</p> <p>多言語自動音声認識（ASR）モデルは、単一のモデルで複数の言語を認識できるという能力から、近年注目を集めています（図1(a)を参照）。最近の進歩により、大規模な教師あり学習や自己教師あり学習による事前学習を通じて、様々な言語で優れた性能が実現されています [3, 5, 6, 7, 10, 11, 12]。例えば、Whisper[6]は、68万時間の多言語データで学習されており、標準的なASRベンチマークにおいて高い汎化性能を発揮します。一方、USM[9]は、1200万時間のラベルなしデータを利用することで、堅牢な多言語性能を実現しています。しかしながら、単一モデルを用いた多言語ASRシステムの応用は、依然として大きな課題を抱えています。言語間の音声的差</p>		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		異、構文の多様性、語彙の相違などが、一貫した最高の性能（SOTA）を達成することを困難にしています。さらに、リソースが豊富な言語と少ない言語間の学習データにおける不均衡が、単一モデルの解決策を制限する要因となっています。		
Minor	Fluency/Register <i>This is a literal translation and sounds a bit unnatural, so it needs to be rephrased.</i>	<p>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.</p> <p>多言語自動音声認識（ASR）モデルは、単一のモデルで複数の言語を認識できるという能力から、近年注目を集めています（図1(a)を参照）。最近の進歩により、大規模な教師あり学習や自己教師あり学習による事前学習を通じて、様々な言語で優れた性能が実現されています [3, 5, 6, 7, 10, 11, 12]。例え</p>	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>ば、Whisper[6]は、68万時間の多言語データで学習されており、標準的なASRベンチマークにおいて高い汎化性能を発揮します。一方、USM[9]は、1200万時間のラベルなしデータを利用することで、堅牢な多言語性能を実現しています。しかしながら、単一モデルを用いた多言語ASRシステムの応用は、依然として大きな課題を抱えています。言語間の音声的差異、構文の多様性、語彙の相違などが、一貫した最高の性能（SOTA）を達成することを困難にしています。さらに、リソースが豊富な言語と少ない言語間の学習データにおける不均衡が、単一モデルの解決策を制限する要因となっています。</p>		
Minor	<p>Fluency/Grammar</p> <p>"一貫した" is a literal translation and grammatically awkward expression. I think "一貫性のある最高性能" is a better translation.</p>	<p>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</p>	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>languages further limit the single-model solutions.</p> <p>多言語自動音声認識（ASR）モデルは、単一のモデルで複数の言語を認識できるという能力から、近年注目を集めています（図1(a)を参照）。最近の進歩により、大規模な教師あり学習や自己教師あり学習による事前学習を通じて、様々な言語で優れた性能が実現されています [3, 5, 6, 7, 10, 11, 12]。例えば、Whisper[6]は、68万時間の多言語データで学習されており、標準的なASRベンチマークにおいて高い汎化性能を発揮します。一方、USM[9]は、1200万時間のラベルなしデータを利用することで、堅牢な多言語性能を実現しています。しかしながら、単一モデルを用いた多言語ASRシステムの応用は、依然として大きな課題を抱えています。言語間の音声的差異、構文の多様性、語彙の相違などが、一貫した最高の性能（SOTA）を達成することを困難にしています。さらに、リソースが豊富な言語と少ない言語間の学習データにおける不均衡が、単一モデルの解決策を制限する要因となっています。</p>		
Minor	Fluency/Register <i>This is a literal and unnatural expression.</i>	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	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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.</p> <p>多言語自動音声認識（ASR）モデルは、単一のモデルで複数の言語を認識できるという能力から、近年注目を集めています（図1(a)を参照）。最近の進歩により、大規模な教師あり学習や自己教師あり学習による事前学習を通じて、様々な言語で優れた性能が実現されています [3, 5, 6, 7, 10, 11, 12]。例えば、Whisper[6]は、68万時間の多言語データで学習されており、標準的なASRベンチマークにおいて高い汎化性能を発揮します。一方、USM[9]は、1200万時間のラベルなしデータを利用することで、堅牢な多言語性能を実現しています。しかしながら、単一モデルを用いた多言語ASRシステムの応用は、依然として大きな課題を抱えています。言語間の音声的差異、構文の多様性、語彙の相違などが、一貫した最高の性能（SOTA）を達成することを困難にしています。さらに、リソースが豊富な言語と少ない言語間の学習データにおける不均衡が、単一モデルの解決策を制限する要因となっています。</p>		
Minor	Fluency/Register	Motivated by these limitations, we propose an alternative strategy that selectively invokes models based on the complexity of the input speech.	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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?</p> <p>これらの制約から、私たちは、入力音声の複雑さに応じてモデルを適切に選択する、別の戦略を提案します。音声認識（ASR）のタスクにおいて、認識の難易度は大きく異なります。音声環境がクリアで、語彙が単純な場合、最先端（SOTA）モデルと一般的なモデルの両方で、通常は低い単語誤り率（WER）が得られます。しかし、騒音が多い環境や、音響的に困難な環境では、WERが増加します[14, 15, 16, 17]。そのような状況では、堅牢な最先端モデルの方が一般的に優れた性能を発揮します[6]。このことから、重要な疑問が生まれます。私たちは、単純な音声と複雑な音声の区別をつけ、それに応じてASRシステムを適応させることができるでしょうか？</p>		
Minor	Fluency/Register <i>It should be "できるのでしょうか？".</i>	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	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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?</p> <p>これらの制約から、私たちは、入力音声の複雑さに応じてモデルを適切に選択する、別の戦略を提案します。音声認識（ASR）のタスクにおいて、認識の難易度は大きく異なります。音声環境がクリアで、語彙が単純な場合、最先端（SOTA）モデルと一般的なモデルの両方で、通常は低い単語誤り率（WER）が得られます。しかし、騒音が多い環境や、音響的に困難な環境では、WERが増加します[14, 15, 16, 17]。そのような状況では、堅牢な最先端モデルの方が一般的に優れた性能を発揮します[6]。このことから、重要な疑問が生まれます。私たちは、単純な音声と複雑な音声の区別をつけ、それに応じてASRシステムを適応させることができますでしょうか？</p>		
Minor	Fluency/Grammar	<p>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</p>	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>invoke the SOTA model, thereby optimizing overall ASR performance.</p> <p>その結果から、SIMAモデルは、最先端モデル（SOTAモデル）を適切に選択的に利用することで、ベースモデルと比較して、3つのデータセットでそれぞれ18.6%、9.3%、28.2%という大幅な単語認識エラー率（WER）の削減を達成しました。さらに、ランダムなモデル選択戦略と比較して、SIMAモデルは一貫して低いWERを示し、それぞれ6.6%、4.2%、16.8%の改善が見られました。特に、FLEURSデータセットにおける改善は顕著であり、これはベースモデルにとっては未知の領域であるのに対し、LID-Topモデルにとっては学習対象の領域であるためです。これらの結果は、SIMAモデルが、最先端モデルをいつ利用すべきかを正確に判断する優れた能力を持っていることを明確に示しており、これにより、全体的な音声認識性能が最適化されます。</p>		