

MQM Translation Quality Report

Project: TranslateGemma_EN-KO_Eval1

SCORE

96.4%

⚠️

FAILED

PER 1K TOKENS

35.7

PER 1K WORDS

31.8

PENALTY

TOTAL

70

⚠️

22

Total Errors

▒▒

2204

Total Words

▒▒

2204

Total Char.

▒▒

1963

Total Tokens

✓

7

Segments

🕒

N/A

Duration

▒▒ Severity Breakdown

●

Critical

0

●

Major

12

●

Minor

10

▒▒ Top Categories

Accuracy/Mistranslation

6

Style

6

Accuracy/Omission

3

Other

3

Fluency/Grammar

2

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 = $\sum (\text{Error count} \times \text{Error weight})$

Score (%) = $(1 - \text{Total Penalty} \div \text{Total tokens}) \times 100$

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
Major	Accuracy/Mistranslation <i>The source and translation mismatch</i>	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty 다국어 음성 인식 시스템에서 효율적인 방법: 음성 인식 난이도에 맞춰 선택적으로 사용하는 방식.	5	0.25%
Major	Accuracy/Mistranslation <i>'Selective invocation' and 'cost-effective approach' are swapped in translation.</i>	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty 다국어 음성 인식 시스템에서 효율적인 방법: 음성 인식 난이도에 맞춰 선택적으로 사용하는 방식.	5	0.25%
Major	Accuracy/Mistranslation <i>"as illustrated in Figure 1(a) is a part of the first sentence, but translation uses it as a part of second sentence. It results in serious change in meaning.</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 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	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>languages further limit the single-model solutions.</p> <p>다국어 자동 음성 인식(ASR) 모델은 단일 모델을 사용하여 여러 언어를 인식할 수 있다는 점에서 상당한 주목을 받고 있습니다 [1, 2, 3, 4]. 그림 1(a)에서 볼 수 있듯이, 최근의 발전으로 인해 대규모 지도 학습 또는 자기 지도 학습을 통해 다양한 언어에서 뛰어난 성능을 달성하고 있습니다 [3, 5, 6, 10, 11, 12]. 예를 들어, Whisper [6]는 68만 시간 분량의 다국어 데이터를 사용하여 학습되었으며, 이를 통해 표준 ASR 평가 지표에서 효과적으로 성능을 일반화할 수 있습니다. 반면, USM [9]는 1200만 시간 분량의 레이블이 없는 데이터를 활용하여 강력한 다국어 성능을 달성합니다. 이러한 발전에도 불구하고, 단일 모델을 사용하는 다국어 ASR 시스템의 적용은 여전히 상당한 어려움에 직면해 있습니다. 언어 간의 음운학적 차이, 구문학적 변동, 어휘의 차이로 인해 일관된 최고 수준의 성능을 달성하기 어렵습니다. 또한, 풍부한 데이터를 가진 언어와 데이터가 부족한 언어 간의 학습 데이터 불균형은 단일 모델 솔루션의 한계를 더욱 심화시킵니다.</p>		
Major	Accuracy/Omission	<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>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>이러한 문제점을 해결하기 위한 일반적인 방법은 입력 음성의 언어를 먼저 식별하는 언어 식별 (LID) 모델을 사용한 후, 해당 모델을 호출하여 음성을 텍스트로 변환하는 것입니다(그림 1(b) 참조). 그러나 이러한 두 단계 접근 방식은 몇 가지 단점이 있습니다. 많은 최첨단(SOTA) 모델들이 상업용이며, 처리량에 따라 사용료가 부과되므로 이 방법은 비용이 많이 듭니다. 또한, 부정확한 언어 식별 예측은 잘못된 모델을 활성화시켜 사용자 경험에 부정적인 영향을 미칠 수 있습니다 [13].</p>		
Major	Fluency/Grammar	<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>이러한 문제점을 해결하기 위한 일반적인 방법은 입력 음성의 언어를 먼저 식별하는 언어 식별 (LID) 모델을 사용한 후, 해당 모델을 호출하여 음성을 텍스트로 변환하는 것입니다(그림 1(b) 참조). 그러나 이러한 두 단계 접근 방식은 몇 가지 단점이 있습니다. 많은 최첨단(SOTA) 모델들이 상업용이며, 처리량에 따라 사용료가 부과되므로 이 방법은 비용이 많이 듭니다. 또한, 부정확한 언어 식별 예측은 잘못된 모델을 활성화시켜 사용자 경험에 부정적인 영향을 미칠 수 있습니다 [13].</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Major	Accuracy/Omission [12]	<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>이러한 문제점을 해결하기 위한 일반적인 방법은 입력 음성의 언어를 먼저 식별하는 언어 식별 (LID) 모델을 사용한 후, 해당 모델을 호출하여 음성을 텍스트로 변환하는 것입니다(그림 1(b) 참조). 그러나 이러한 두 단계 접근 방식은 몇 가지 단점이 있습니다. 많은 최첨단(SOTA) 모델들이 상업용이며, 처리량에 따라 사용료가 부과되므로 이 방법은 비용이 많이 듭니다. 또한, 부정확한 언어 식별 예측은 잘못된 모델을 활성화시켜 사용자 경험에 부정적인 영향을 미칠 수 있습니다 [13].</p>	5	0.25%
Major	Style	<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</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>speech inputs and adapt our ASR system accordingly?</p> <p>이러한 한계점을 고려하여, 우리는 입력 음성의 복잡도에 따라 모델을 선택적으로 사용하는 대안적인 전략을 제안합니다. 음성 인식(ASR) 작업에서 인식의 난이도는 매우 다양합니다. 깨끗한 음향 환경에서 간단한 어휘를 사용하는 경우, 최첨단(SOTA) 모델과 일반 모델 모두 일반적으로 낮은 단어 오류율(WER)을 보입니다. 그러나 소음이 심하거나 음향적으로 어려운 환경에서는 WER이 증가합니다 [14, 15, 16, 17]. 이때, 강력한 SOTA 모델이 일반적으로 더 나은 성능을 보입니다 [6]. 이러한 관찰은 중요한 질문을 제기합니다. 즉, 우리는 간단한 음성과 복잡한 음성을 구별할 수 있으며, 이에 따라 ASR 시스템을 적응시킬 수 있을까요?</p>		
Major	Accuracy/Mistranslation	<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>연구 결과에 따르면, SIMA 모델은 최첨단(SOTA) 모델을 선택적으로 활용함으로써, 기존 모델에 비해 세 가지 데이터셋에서 각각 18.6%, 9.3%, 28.2%의 상당한 단어 오류율(WER) 감소를 달성했습니다. 또한, 무작위 모델 호</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>출 전략과 비교했을 때, SIMA는 일관되게 더 낮은 WER을 제공하며, 각각 6.6%, 4.2%, 16.8%의 성능 향상을 보였습니다. 특히, FLEURS 데이터셋에서의 성능 향상은 매우 두드러지는데, 이는 기준 모델에게는 해당 분야 (in-domain)가 아니지만, LID-Top 모델에게는 해당 분야이기 때문입니다. 이러한 결과는 SIMA가 최첨단 모델을 언제 활용해야 하는지 정확하게 판단하는 놀라운 능력을 입증하며, 이를 통해 전체 음성 인식(ASR) 성능을 최적화한다는 것을 명확하게 보여줍니다.</p>		
Major	Accuracy/Mistranslation	<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>연구 결과에 따르면, SIMA 모델은 최첨단(SOTA) 모델을 선택적으로 활용함으로써, 기준 모델에 비해 세 가지 데이터셋에서 각각 18.6%, 9.3%, 28.2%의 상당한 단어 오류율(WER) 감소를 달성했습니다. 또한, 무작위 모델 호출 전략과 비교했을 때, SIMA는 일관되게 더 낮은 WER을 제공하며, 각각 6.6%, 4.2%, 16.8%의 성능 향상을 보였습니다. 특히, FLEURS 데이터셋에서의 성능 향상은 매우 두드러지는데, 이는</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		기준 모델에게는 해당 분야(in-domain)가 아니지만, LID-Top 모델에게는 해당 분야이기 때문 입니다. 이러한 결과는 SIMA가 최첨단 모델을 언제 활용해야 하는지 정확하게 판단하는 놀라운 능력을 입증하며, 이를 통해 전체 음성 인식(ASR) 성능을 최적화 한다는 것을 명확하게 보여줍니다.		
Major	Accuracy/Omission	<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>음성 명령 인식 정확도(ACC)와 F1 점수는 약 70%로, 저희의 가설인 즉, SLLM(대규모 언어 모델)이 음성의 복잡도에 따라 효과적으로 구분할 수 있다는 점을 뒷받침합니다. SIMA는 LID-Top과 비교했을 때 약간의 단어 오류율(WER) 차이를 보이지만, 세 가지 데이터셋에서 약 0.51배만큼 음성 명령 호출 비용을 줄여, 관련 비용을 크게 낮춥니다.</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 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	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>현재 SIMA 모델은 음성 인식률 (WER)을 크게 향상시키지만, 여전히 Whisper [6]에 비해 외부 데이터셋(out-of-domain data)인 FLEURS [28]에서 성능이 떨어집니다. 이러한 제한은 초기 가설, 즉 기본 SLLM 모델이 "invoke" 작업을 효과적으로 수행할 수 있다는 가정에서 비롯됩니다. 우리의 기본 SLLM 모델 [29]은 훈련 데이터의 한계로 인해 Whisper와 같은 전문 모델에 비해 본질적으로 성능이 약합니다. 향후 연구에서는 Whisper [6]을 기본 모델로 채택하고, SIMA 시스템을 더욱 개선하여 최첨단(SOTA) 모델의 음성 인식 성능을 향상시킬 계획입니다.</p>		
Major	Accuracy/Untranslated	<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>현재 SIMA 모델은 음성 인식률 (WER)을 크게 향상시키지만, 여전히 Whisper [6]에 비해 외부 데이터셋(out-of-domain data)인 FLEURS [28]에서 성능이 떨어집니다. 이러한 제한은 초기 가설, 즉 기본 SLLM 모델이 "invoke" 작업을 효과적으로 수행할 수 있다는 가정에서 비롯됩니다. 우리의 기본 SLLM 모델 [29]은 훈련 데이터의 한계로 인해 Whisper와 같은 전문 모델에 비해 본질적으로 성능이 약합니다. 향후 연구에서는 Whisper</p>	5	0.25%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		[6]을 기본 모델로 채택하고, SIMA 시스템을 더욱 개선하여 최첨단(SOTA) 모델의 음성 인식 성능을 향상시킬 계획입니다.		
Minor	Style <i>unnatural sentence flow</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>이러한 문제점을 해결하기 위한 일반적인 방법은 입력 음성의 언어를 먼저 식별하는 언어 식별 (LID) 모델을 사용한 후, 해당 모델을 호출하여 음성을 텍스트로 변환하는 것입니다(그림 1(b) 참조). 그러나 이러한 두 단계 접근 방식은 몇 가지 단점이 있습니다. 많은 최첨단(SOTA) 모델들이 상업용이며, 처리량에 따라 사용료가 부과되므로 이 방법은 비용이 많이 듭니다. 또한, 부정확한 언어 식별 예측은 잘못된 모델을 활성화시켜 사용자 경험에 부정적인 영향을 미칠 수 있습니다 [13].</p>	1	0.05%
Minor	Style	<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</p>	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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>이러한 한계점을 고려하여, 우리 는 입력 음성의 복잡도에 따라 모델을 선택적으로 사용하는 대안적인 전략을 제안합니다. 음성 인식 (ASR) 작업에서 인식의 난이도는 매우 다양합니다. 깨끗한 음향 환경에서 간단한 어휘를 사용하는 경우, 최첨단(SOTA) 모델과 일반 모델 모두 일반적으로 낮은 단어 오류율(WER)을 보입니다. 그러나 소음이 심하거나 음향적으로 어려운 환경에서는 WER이 증가합니다 [14, 15, 16, 17]. 이때, 강력한 SOTA 모델이 일반적으로 더 나은 성능을 보입니다 [6]. 이러한 관찰은 중요한 질문을 제기합니다. 즉, 우리는 간단한 음성과 복잡한 음성을 구별할 수 있으며, 이에 따라 ASR 시스템을 적응시킬 수 있을까요?</p>		
Minor	Other <i>unnatural sentence flow</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>이러한 한계점을 고려하여, 우리 는 입력 음성의 복잡도에 따라 모델을 선택적으로 사용하는 대안적</p>	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		인 전략을 제안합니다. 음성 인식 (ASR) 작업에서 인식의 난이도는 매우 다양합니다. 깨끗한 음향 환경에서 간단한 어휘를 사용하는 경우, 최첨단(SOTA) 모델과 일반 모델 모두 일반적으로 낮은 단어 오류율(WER)을 보입니다. 그러나 소음이 심하거나 음향적으로 어려운 환경에서는 WER이 증가합니다 [14, 15, 16, 17]. 이때, 강력한 SOTA 모델이 일반적으로 더 나은 성능을 보입니다 [6]. 이러한 관찰은 중요한 질문을 제기합니다. 즉, 우리는 간단한 음성과 복잡한 음성을 구별할 수 있으며, 이에 따라 ASR 시스템을 적응시킬 수 있을까요?		
Minor	Other <i>unnatural sentence flow</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>이러한 한계점을 고려하여, 우리는 입력 음성의 복잡도에 따라 모델을 선택적으로 사용하는 대안적인 전략을 제안합니다. 음성 인식 (ASR) 작업에서 인식의 난이도는 매우 다양합니다. 깨끗한 음향 환경에서 간단한 어휘를 사용하는 경우, 최첨단(SOTA) 모델과 일반 모델 모두 일반적으로 낮은 단어 오류율(WER)을 보입니다. 그러나 소음이 심하거나 음향적으로 어려운 환경에서는 WER이 증가</p>	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		합니다 [14, 15, 16, 17]. 이때, 강력한 SOTA 모델이 일반적으로 더 나은 성능을 보입니다 [6]. 이러한 관찰은 중요한 질문을 제기합니다. 즉, 우리는 간단한 음성과 복잡한 음성을 구별할 수 있으며, 이에 따라 ASR 시스템을 적응시킬 수 있을까요?		
Minor	Style <i>translationese</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>연구 결과에 따르면, SIMA 모델은 최첨단(SOTA) 모델을 선택적으로 활용함으로써, 기존 모델에 비해 세 가지 데이터셋에서 각각 18.6%, 9.3%, 28.2%의 상당한 단어 오류율(WER) 감소를 달성했습니다. 또한, 무작위 모델 호출 전략과 비교했을 때, SIMA는 일관되게 더 낮은 WER을 제공하며, 각각 6.6%, 4.2%, 16.8%의 성능 향상을 보였습니다. 특히, FLEURS 데이터셋에서의 성능 향상은 매우 두드러지는데, 이는 기존 모델에게는 해당 분야(in-domain)가 아니지만, LID-Top 모델에게는 해당 분야이기 때문입니다. 이러한 결과는 SIMA가 최첨단 모델을 언제 활용해야 하는지 정확하게 판단하는 놀라운 능력을 입증하며, 이를 통해 전체 음</p>	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		성 인식(ASR) 성능을 최적화한 다는 것을 명확하게 보여줍니다.		
Minor	Other <i>translationese</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>연구 결과에 따르면, SIMA 모델은 최첨단(SOTA) 모델을 선택적으로 활용함으로써, 기존 모델에 비해 세 가지 데이터셋에서 각각 18.6%, 9.3%, 28.2%의 상당한 단어 오류율(WER) 감소를 달성했습니다. 또한, 무작위 모델 호출 전략과 비교했을 때, SIMA는 일관되게 더 낮은 WER을 제공하며, 각각 6.6%, 4.2%, 16.8%의 성능 향상을 보였습니다. 특히, FLEURS 데이터셋에서의 성능 향상은 매우 두드러지는데, 이는 기존 모델에게는 해당 분야(in-domain)가 아니지만, LID-Top 모델에게는 해당 분야이기 때문입니다. 이러한 결과는 SIMA가 최첨단 모델을 언제 활용해야 하는지 정확하게 판단하는 놀라운 능력을 입증하며, 이를 통해 전체 음성 인식(ASR) 성능을 최적화한다는 것을 명확하게 보여줍니다.</p>	1	0.05%
Minor	Fluency/Register	The invocation decision accuracy (ACC) and F1 scores are approximately 70%,	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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>음성 명령 인식 정확도(ACC)와 F1 점수는 약 70%로, 저희의 가설인 즉, SLLM(대규모 언어 모델)이 음성의 복잡도에 따라 효과적으로 구분할 수 있다는 점을 뒷받침합니다. SIMA는 LID-Top과 비교했을 때 약간의 단어 오류율(WER) 차이를 보이지만, 세 가지 데이터셋에서 약 0.51배만큼 음성 명령 호출 비용을 줄여, 관련 비용을 크게 낮춥니다.</p>		
Minor	Style	<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>음성 명령 인식 정확도(ACC)와 F1 점수는 약 70%로, 저희의 가설인 즉, SLLM(대규모 언어 모델)이 음성의 복잡도에 따라 효과적으로 구분할 수 있다는 점을 뒷받침합니다. SIMA는 LID-Top과 비교했을 때 약간의 단어 오류율(WER) 차이를 보이지만, 세 가지 데이터셋에서 약 0.51배만큼 음성 명령 호출 비용을 줄여, 관련 비용을 크게 낮춥니다.</p>	1	0.05%
Minor	Style	<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</p>	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		<p>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>현재 SIMA 모델은 음성 인식을 (WER)을 크게 향상시키지만, 여전히 Whisper [6]에 비해 외부 데이터셋(out-of-domain data)인 FLEURS [28]에서 성능이 떨어집니다. 이러한 제한은 초기 가설, 즉 기본 SLLM 모델이 "invoke" 작업을 효과적으로 수행할 수 있다는 가정에서 비롯됩니다. 우리의 기본 SLLM 모델 [29]은 훈련 데이터의 한계로 인해 Whisper와 같은 전문 모델에 비해 본질적으로 성능이 약합니다. 향후 연구에서는 Whisper [6]을 기본 모델로 채택하고, SIMA 시스템을 더욱 개선하여 최첨단(SOTA) 모델의 음성 인식 성능을 향상시킬 계획입니다.</p>		
Minor	Fluency/Grammar	<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>현재 SIMA 모델은 음성 인식을 (WER)을 크게 향상시키지만, 여전히 Whisper [6]에 비해 외부 데이터셋(out-of-domain data)인 FLEURS [28]에서 성</p>	1	0.05%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		능이 떨어집니다. 이러한 제한은 초기 가설, 즉 기본 SLLM 모델이 "invoke" 작업을 효과적으로 수행할 수 있다는 가정에서 비롯됩니다. 우리의 기본 SLLM 모델 [29]은 훈련 데이터의 한계로 인해 Whisper와 같은 전문 모델에 비해 본질적으로 성능이 약합니다. 향후 연구에서는 Whisper [6]을 기본 모델로 채택하고, SIMA 시스템을 더욱 개선하여 최첨단(SOTA) 모델의 음성 인식 성능을 향상시킬 계획입니다.		

MQM Annotation Tool by 

This tool uses the [Multidimensional Quality Metrics \(MQM\)](#) framework, licensed under [CC BY 4.0](#) by [The MQM Council](#).