MQM Translation Quality Report

Project: TranslateGemma_EN-HMN_Eval1

SCORE

62.5 %

FAILED

PER 1K TOKENS: 375.3

PER 1K WORDS: 1140.4

PENALTY TOTAL: **1129**

57	990	4751	3008	7	N/A
Total Errors	Total Words	Total Char.	Total Tokens	Segments	Duration

Severity Breakdown

Critical	43
Major	10
Minor	4

Top Categories

Accuracy/Mistranslation	36
Fluency/Grammar	8
Fluency/Inconsistency	7
Accuracy/Addition	5
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
Critical	Accuracy/Mistranslation *These 2 words together mean nothing intelligible*	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty Huab chaws kev siv lus hauv kev ntsuas lus hauv ntau yam lus: Ib txoj kev siv peev xwm uas muaj tus nqi qis, uas tuaj yeem kho kom haum rau qhov nyuaj ntawm kev ntsuas lus.	25	0.83%
Critical	Accuracy/Mistranslation *Completely wrong translation*	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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Critical	Accuracy/Addition *Repeat of error 2*	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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		(supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Critical	Accuracy/Mistranslation *This is pretty much gibberish resulting from an attempt at direct English to Hmong translation for each word. No Hmong person reading these words will understand what idea is being conveyed.*	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,	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Critical	Accuracy/Mistranslation *Wrong word for "hour"*	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	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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.		
		Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Critical	Fluency/Grammar *Again, this is a direct English to Hmong attempt that fails to convey the idea. It also reads like a run-on sentence in Hmong. Too many ideas trying to be captured in one sentence.*	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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Critical	Accuracy/Mistranslation *Wrong word for "advances"*	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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Critical	Accuracy/Mistranslation *Wrong word for "challenges"*	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	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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.		
		Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Critical	Accuracy/Mistranslation	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-	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		the-art (SOTA) performance. Moreover, imbalances in training data between high-resource and low-resource languages further limit the single-model solutions.		
		Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Critical	Accuracy/Mistranslation	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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Critical	Fluency/Grammar	Multilingual automatic speech recognition (ASR) models have	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
	Wrong word to use with the previous word to mean "Moreover"	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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Critical	Accuracy/Mistranslation *Long chain of gibberish resulting from direct English to Hmong translation. No Hmong person would be able to understand what idea is being conveyed here.*	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	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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.		
		Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Critical	Accuracy/Mistranslation *wrong word for "identification"*	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]. Ib txoj kev uas feem ntau siv los pab cov teeb meem no yog siv ib qho kev kos lus (language identification, LID) kom pom tias lus twg tau hais ua ntej, ces siv qhov kos lus uas zoo tshaj plaws (state-of-the-art, SOTA) kom sau lus, raws li qhia hauv Duab 1(b). Tab sis, txoj kev no muaj qee qhov tsis zoo. Ntau qhov	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		kos lus zoo tshaj plaws yog khoom lag luam [12] thiab muaj tus nqi siv, raws li qhov ntau npaum li cas koj siv, uas ua rau txoj kev no tsis muaj txiaj. Tsis tas li ntawd, yog hais tias LID pom lus tsis raug, nws yuav ua rau siv qhov kos lus tsis raug, uas yuav ua rau tsis zoo rau cov neeg siv.		
Critical	Accuracy/Mistranslation *Inaccurate translation of what's trying to be conveyed in the English text*	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]. Ib txoj kev uas feem ntau siv los pab cov teeb meem no yog siv ib qho kev kos lus (language identification, LID) kom pom tias lus twg tau hais ua ntej, ces siv qhov kos lus uas zoo tshaj plaws (state-of-the-art, SOTA) kom sau lus, raws li qhia hauv Duab 1(b). Tab sis, txoj kev no muaj qee qhov tsis zoo. Ntau qhov kos lus zoo tshaj plaws yog khoom lag luam [12] thiab muaj tus nqi siv, raws li qhov ntau npaum li cas koj siv, uas ua rau txoj kev no tsis muaj txiaj. Tsis tas li ntawd, yog hais tias LID pom lus tsis raug, nws yuav ua rau siv qhov kos lus tsis	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		raug, uas yuav ua rau tsis zoo rau cov neeg siv.		
Critical	Accuracy/Mistranslation *Inaccurate translation for "SOTA models"*	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]. Ib txoj kev uas feem ntau siv los pab cov teeb meem no yog siv ib qho kev kos lus (language identification, LID) kom pom tias lus twg tau hais ua ntej, ces siv qhov kos lus uas zoo tshaj plaws (state-of-the-art, SOTA) kom sau lus, raws li qhia hauv Duab 1(b). Tab sis, txoj kev no muaj qee qhov tsis zoo. Ntau qhov kos lus zoo tshaj plaws yog khoom lag luam [12] thiab muaj tus nqi siv, raws li qhov ntau npaum li cas koj siv, uas ua rau txoj kev no tsis muaj txiaj. Tsis tas li ntawd, yog hais tias LID pom lus tsis raug, nws yuav ua rau siv qhov kos lus tsis raug, uas yuav ua rau tsis zoo rau cov neeg siv.	25	0.83%
Critical	Fluency/Grammar *When placed at this part of the sentence, the entire sentence no longer makes any sense.*	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	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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]. Ib txoj kev uas feem ntau siv los pab cov teeb meem no yog siv ib qho kev kos lus (language identification, LID) kom pom tias lus twg tau hais ua ntej, ces siv qhov kos lus uas zoo tshaj plaws (state-of-the-art, SOTA) kom sau lus, raws li qhia hauv Duab 1(b). Tab sis, txoj kev no muaj qee qhov tsis zoo. Ntau qhov kos lus zoo tshaj plaws yog khoom lag luam [12] thiab muaj tus nqi siv, raws li qhov ntau npaum li cas koj siv, uas ua rau txoj kev no tsis muaj txiaj. Tsis tas li ntawd, yog hais tias LID pom lus tsis raug, nws yuav ua rau siv qhov kos lus tsis raug, uas yuav ua rau tsis zoo rau cov neeg siv.		
Critical	Fluency/Inconsistency *This is placed here because the translation is trying to replicate English sentence structures with Hmong. If you try to structure your Hmong sentences like English sentences, the meaning will be lost or extremely confusing.*	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].	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		Ib txoj kev uas feem ntau siv los pab cov teeb meem no yog siv ib qho kev kos lus (language identification, LID) kom pom tias lus twg tau hais ua ntej, ces siv qhov kos lus uas zoo tshaj plaws (state-of-the-art, SOTA) kom sau lus, raws li qhia hauv Duab 1(b). Tab sis, txoj kev no muaj qee qhov tsis zoo. Ntau qhov kos lus zoo tshaj plaws yog khoom lag luam [12] thiab muaj tus nqi siv, raws li qhov ntau npaum li cas koj siv, uas ua rau txoj kev no tsis muaj txiaj. Tsis tas li ntawd, yog hais tias LID pom lus tsis raug, nws yuav ua rau siv qhov kos lus tsis raug, uas yuav ua rau tsis zoo rau cov neeg siv.		
Critical	Accuracy/Mistranslation *means "not good", instead of "alternative"*	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? Yog vim lawv cov kev txwv txhim, peb tau pom ib txoj kev tsis zoo uas siv cov qauv sib txawv raws li qhov tseeb ntawm suab uas tau los. Hauv cov haujlwm kuaj	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		suab, qhov nyuaj ntawm kev kuaj suab muaj qhov sib txawv ntau. Thaum tsis muaj suab thiab siv cov lus qab zib me, cov qauv zoo tshaj plaws thiab cov qauv siv tau feem ntau ua kom muaj qhov tsis raug ntawm cov lus (WER) qis. Tab sis, thaum muaj suab thiab nyuaj rau kuaj suab, qhov WER nce [14, 15, 16, 17], thiab cov qauv zoo tshaj plaws uas muaj zog ua tau zoo dua [6]. Qhov no tau coj mus rau ib lo lus nug tseem ceeb: Peb puas tuaj yeem sib txawv ntawm cov suab uas yooj yim thiab nyuaj, thiab siv peb cov tshuab kuaj suab kom raug?		
Critical	Accuracy/Mistranslation *gibberish*	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? Yog vim lawv cov kev txwv txhim, peb tau pom ib txoj kev tsis zoo uas siv cov qauv sib txawv raws li qhov tseeb ntawm suab uas tau los. Hauv cov haujlwm kuaj suab, qhov nyuaj ntawm kev kuaj suab muaj qhov sib	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		txawv ntau. Thaum tsis muaj suab thiab siv cov lus qab zib me, cov qauv zoo tshaj plaws thiab cov qauv siv tau feem ntau ua kom muaj qhov tsis raug ntawm cov lus (WER) qis. Tab sis, thaum muaj suab thiab nyuaj rau kuaj suab, qhov WER nce [14, 15, 16, 17], thiab cov qauv zoo tshaj plaws uas muaj zog ua tau zoo dua [6]. Qhov no tau coj mus rau ib lo lus nug tseem ceeb: Peb puas tuaj yeem sib txawv ntawm cov suab uas yooj yim thiab nyuaj, thiab siv peb cov tshuab kuaj suab kom raug?		
Critical	Accuracy/Mistranslation	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? Yog vim lawv cov kev txwv txhim, peb tau pom ib txoj kev tsis zoo uas siv cov qauv sib txawv raws li qhov tseeb ntawm suab uas tau los. Hauv cov haujlwm kuaj suab, qhov nyuaj ntawm kev kuaj suab muaj qhov sib txawv ntau. Thaum tsis muaj suab thiab siv cov lus	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		qab zib me, cov qauv zoo tshaj plaws thiab cov qauv siv tau feem ntau ua kom muaj qhov tsis raug ntawm cov lus (WER) qis. Tab sis, thaum muaj suab thiab nyuaj rau kuaj suab, qhov WER nce [14, 15, 16, 17], thiab cov qauv zoo tshaj plaws uas muaj zog ua tau zoo dua [6]. Qhov no tau coj mus rau ib lo lus nug tseem ceeb: Peb puas tuaj yeem sib txawv ntawm cov suab uas yooj yim thiab nyuaj, thiab siv peb cov tshuab kuaj suab kom raug?		
Critical	Accuracy/Mistranslation *gibberish*	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? Yog vim lawv cov kev txwv txhim, peb tau pom ib txoj kev tsis zoo uas siv cov qauv sib txawv raws li qhov tseeb ntawm suab uas tau los. Hauv cov haujlwm kuaj suab, qhov nyuaj ntawm kev kuaj suab muaj qhov sib txawv ntau. Thaum tsis muaj suab thiab siv cov lus qab zib me, cov qauv zoo tshaj plaws thiab cov qauv	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		siv tau feem ntau ua kom muaj qhov tsis raug ntawm cov lus (WER) qis. Tab sis, thaum muaj suab thiab nyuaj rau kuaj suab, qhov WER nce [14, 15, 16, 17], thiab cov qauv zoo tshaj plaws uas muaj zog ua tau zoo dua [6]. Qhov no tau coj mus rau ib lo lus nug tseem ceeb: Peb puas tuaj yeem sib txawv ntawm cov suab uas yooj yim thiab nyuaj, thiab siv peb cov tshuab kuaj suab kom raug?		
Critical	Accuracy/Mistranslation *wrong translation for "simple vocabulary"*	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? Yog vim lawv cov kev txwv txhim, peb tau pom ib txoj kev tsis zoo uas siv cov qauv sib txawv raws li qhov tseeb ntawm suab uas tau los. Hauv cov haujlwm kuaj suab, qhov nyuaj ntawm kev kuaj suab muaj qhov sib txawv ntau. Thaum tsis muaj suab thiab siv cov lus qab zib me, cov qauv zoo tshaj plaws thiab cov qauv siv tau feem ntau ua kom muaj qhov tsis raug ntawm	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		cov lus (WER) qis. Tab sis, thaum muaj suab thiab nyuaj rau kuaj suab, qhov WER nce [14, 15, 16, 17], thiab cov qauv zoo tshaj plaws uas muaj zog ua tau zoo dua [6]. Qhov no tau coj mus rau ib lo lus nug tseem ceeb: Peb puas tuaj yeem sib txawv ntawm cov suab uas yooj yim thiab nyuaj, thiab siv peb cov tshuab kuaj suab kom raug?		
Critical	Fluency/Grammar	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? Yog vim lawv cov kev txwv txhim, peb tau pom ib txoj kev tsis zoo uas siv cov qauv sib txawv raws li qhov tseeb ntawm suab uas tau los. Hauv cov haujlwm kuaj suab, qhov nyuaj ntawm kev kuaj suab muaj qhov sib txawv ntau. Thaum tsis muaj suab thiab siv cov lus qab zib me, cov qauv zoo tshaj plaws thiab cov qauv siv tau feem ntau ua kom muaj qhov tsis raug ntawm cov lus (WER) qis. Tab sis, thaum muaj suab thiab	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		nyuaj rau kuaj suab, qhov WER nce [14, 15, 16, 17], thiab cov qauv zoo tshaj plaws uas muaj zog ua tau zoo dua [6]. Qhov no tau coj mus rau ib lo lus nug tseem ceeb: Peb puas tuaj yeem sib txawv ntawm cov suab uas yooj yim thiab nyuaj, thiab siv peb cov tshuab kuaj suab kom raug?		
Critical	Accuracy/Mistranslation *doesn't mean anything*	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. Cov yeeb yis qhia tias, vim tias siv cov qauv zoo tshaj plaws (SOTA) kom raug, SIMA model muaj kev txhim kho loj heev hauv kev ntsuas lus (WER), yog 18.6%, 9.3%, thiab 28.2% ntau dua qhov qauv pib ntawm peb qhov datasets. Tsis tas li ntawd, piv txwv nrog kev siv qauv los xaus, SIMA ib txwm muaj qhov WER qis dua, nrog kev txhim kho ntawm 6.6%, 4.2%, thiab 16.8%. Qhov tseem ceeb, kev txhim kho ntawm FLEURS dataset yog	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		qhov tseem ceeb tshwj xeeb, vim tias nws tsis yog qhov uas qauv pib siv tau, tab sis qauv LID-Top siv tau. Cov lus teb no qhia meej tias SIMA muaj peev xwm zoo heev los txiav txim tias yuav siv qauv zoo tshaj plaws (SOTA) thaum twg, uas pab txhim kho kev ua haujlwm ntawm kev ntsuas lus (ASR) tag nrho.		
Critical	Accuracy/Mistranslation *gibberish*	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. Cov yeeb yis qhia tias, vim tias siv cov qauv zoo tshaj plaws (SOTA) kom raug, SIMA model muaj kev txhim kho loj heev hauv kev ntsuas lus (WER), yog 18.6%, 9.3%, thiab 28.2% ntau dua qhov qauv pib ntawm peb qhov datasets. Tsis tas li ntawd, piv txwv nrog kev siv qauv los xaus, SIMA ib txwm muaj qhov WER qis dua, nrog kev txhim kho ntawm 6.6%, 4.2%, thiab 16.8%. Qhov tseem ceeb, kev txhim kho	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		ntawm FLEURS dataset yog qhov tseem ceeb tshwj xeeb, vim tias nws tsis yog qhov uas qauv pib siv tau, tab sis qauv LID-Top siv tau. Cov lus teb no qhia meej tias SIMA muaj peev xwm zoo heev los txiav txim tias yuav siv qauv zoo tshaj plaws (SOTA) thaum twg, uas pab txhim kho kev ua haujlwm ntawm kev ntsuas lus (ASR) tag nrho.		
Critical	Accuracy/Mistranslation *odd translation for "base". No hmong person would understand this to mean "base".*	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. Cov yeeb yis qhia tias, vim tias siv cov qauv zoo tshaj plaws (SOTA) kom raug, SIMA model muaj kev txhim kho loj heev hauv kev ntsuas lus (WER), yog 18.6%, 9.3%, thiab 28.2% ntau dua qhov qauv pib ntawm peb qhov datasets. Tsis tas li ntawd, piv txwv nrog kev siv qauv los xaus, SIMA ib txwm muaj qhov WER qis dua, nrog kev txhim kho ntawm 6.6%, 4.2%, thiab 16.8%. Qhov	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		tseem ceeb, kev txhim kho ntawm FLEURS dataset yog qhov tseem ceeb tshwj xeeb, vim tias nws tsis yog qhov uas qauv pib siv tau, tab sis qauv LID-Top siv tau. Cov lus teb no qhia meej tias SIMA muaj peev xwm zoo heev los txiav txim tias yuav siv qauv zoo tshaj plaws (SOTA) thaum twg, uas pab txhim kho kev ua haujlwm ntawm kev ntsuas lus (ASR) tag nrho.		
Critical	Accuracy/Mistranslation *odd translation for "base". No hmong person would understand this to mean "base".*	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. Cov yeeb yis qhia tias, vim tias siv cov qauv zoo tshaj plaws (SOTA) kom raug, SIMA model muaj kev txhim kho loj heev hauv kev ntsuas lus (WER), yog 18.6%, 9.3%, thiab 28.2% ntau dua qhov qauv pib ntawm peb qhov datasets. Tsis tas li ntawd, piv txwv nrog kev siv qauv los xaus, SIMA ib txwm muaj qhov WER qis dua, nrog kev txhim kho ntawm 6.6%,	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		4.2%, thiab 16.8%. Qhov tseem ceeb, kev txhim kho ntawm FLEURS dataset yog qhov tseem ceeb tshwj xeeb, vim tias nws tsis yog qhov uas qauv pib siv tau, tab sis qauv LID-Top siv tau. Cov lus teb no qhia meej tias SIMA muaj peev xwm zoo heev los txiav txim tias yuav siv qauv zoo tshaj plaws (SOTA) thaum twg, uas pab txhim kho kev ua haujlwm ntawm kev ntsuas lus (ASR) tag nrho.		
Critical	Accuracy/Addition *unnecessary and changes the meaning of the phrase*	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. Cov yeeb yis qhia tias, vim tias siv cov qauv zoo tshaj plaws (SOTA) kom raug, SIMA model muaj kev txhim kho loj heev hauv kev ntsuas lus (WER), yog 18.6%, 9.3%, thiab 28.2% ntau dua qhov qauv pib ntawm peb qhov datasets. Tsis tas li ntawd, piv txwv nrog kev siv qauv los xaus, SIMA ib txwm muaj qhov WER qis dua, nrog kev	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		txhim kho ntawm 6.6%, 4.2%, thiab 16.8%. Qhov tseem ceeb, kev txhim kho ntawm FLEURS dataset yog qhov tseem ceeb tshwj xeeb, vim tias nws tsis yog qhov uas qauv pib siv tau, tab sis qauv LID-Top siv tau. Cov lus teb no qhia meej tias SIMA muaj peev xwm zoo heev los txiav txim tias yuav siv qauv zoo tshaj plaws (SOTA) thaum twg, uas pab txhim kho kev ua haujlwm ntawm kev ntsuas lus (ASR) tag nrho.		
Critical	Accuracy/Mistranslation *gibberish*	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. Cov yeeb yis qhia tias, vim tias siv cov qauv zoo tshaj plaws (SOTA) kom raug, SIMA model muaj kev txhim kho loj heev hauv kev ntsuas lus (WER), yog 18.6%, 9.3%, thiab 28.2% ntau dua qhov qauv pib ntawm peb qhov datasets. Tsis tas li ntawd, piv txwv nrog kev siv qauv los xaus, SIMA ib txwm muaj qhov	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		WER qis dua, nrog kev txhim kho ntawm 6.6%, 4.2%, thiab 16.8%. Qhov tseem ceeb, kev txhim kho ntawm FLEURS dataset yog qhov tseem ceeb tshwj xeeb, vim tias nws tsis yog qhov uas qauv pib siv tau, tab sis qauv LID-Top siv tau. Cov lus teb no qhia meej tias SIMA muaj peev xwm zoo heev los txiav txim tias yuav siv qauv zoo tshaj plaws (SOTA) thaum twg, uas pab txhim kho kev ua haujlwm ntawm kev ntsuas lus (ASR) tag nrho.		
Critical	Accuracy/Mistranslation *gibberish*	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. Cov yeeb yis qhia tias, vim tias siv cov qauv zoo tshaj plaws (SOTA) kom raug, SIMA model muaj kev txhim kho loj heev hauv kev ntsuas lus (WER), yog 18.6%, 9.3%, thiab 28.2% ntau dua qhov qauv pib ntawm peb qhov datasets. Tsis tas li ntawd, piv txwv nrog kev siv qauv los xaus,	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		SIMA ib txwm muaj qhov WER qis dua, nrog kev txhim kho ntawm 6.6%, 4.2%, thiab 16.8%. Qhov tseem ceeb, kev txhim kho ntawm FLEURS dataset yog qhov tseem ceeb tshwj xeeb, vim tias nws tsis yog qhov uas qauv pib siv tau, tab sis qauv LID-Top siv tau. Cov lus teb no qhia meej tias SIMA muaj peev xwm zoo heev los txiav txim tias yuav siv qauv zoo tshaj plaws (SOTA) thaum twg, uas pab txhim kho kev ua haujlwm ntawm kev ntsuas lus (ASR) tag nrho.		
Critical	Accuracy/Mistranslation *Wrong translation for "These findings"*	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. Cov yeeb yis qhia tias, vim tias siv cov qauv zoo tshaj plaws (SOTA) kom raug, SIMA model muaj kev txhim kho loj heev hauv kev ntsuas lus (WER), yog 18.6%, 9.3%, thiab 28.2% ntau dua qhov qauv pib ntawm peb qhov datasets. Tsis tas li ntawd, piv txwv	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		nrog kev siv qauv los xaus, SIMA ib txwm muaj qhov WER qis dua, nrog kev txhim kho ntawm 6.6%, 4.2%, thiab 16.8%. Qhov tseem ceeb, kev txhim kho ntawm FLEURS dataset yog qhov tseem ceeb tshwj xeeb, vim tias nws tsis yog qhov uas qauv pib siv tau, tab sis qauv LID-Top siv tau. Cov lus teb no qhia meej tias SIMA muaj peev xwm zoo heev los txiav txim tias yuav siv qauv zoo tshaj plaws (SOTA) thaum twg, uas pab txhim kho kev ua haujlwm ntawm kev ntsuas lus (ASR) tag nrho.		
Critical	Fluency/Grammar *Another case of trying to use English sentence structure with Hmong, leading to the sentence making no sense*	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. Cov yeeb yis qhia tias, vim tias siv cov qauv zoo tshaj plaws (SOTA) kom raug, SIMA model muaj kev txhim kho loj heev hauv kev ntsuas lus (WER), yog 18.6%, 9.3%, thiab 28.2% ntau dua qhov qauv pib ntawm peb qhov datasets.	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		Tsis tas li ntawd, piv txwv nrog kev siv qauv los xaus, SIMA ib txwm muaj qhov WER qis dua, nrog kev txhim kho ntawm 6.6%, 4.2%, thiab 16.8%. Qhov tseem ceeb, kev txhim kho ntawm FLEURS dataset yog qhov tseem ceeb tshwj xeeb, vim tias nws tsis yog qhov uas qauv pib siv tau, tab sis qauv LID-Top siv tau. Cov lus teb no qhia meej tias SIMA muaj peev xwm zoo heev los txiav txim tias yuav siv qauv zoo tshaj plaws (SOTA) thaum twg, uas pab txhim kho kev ua haujlwm ntawm kev ntsuas lus (ASR) tag nrho.		
Critical	Accuracy/Mistranslation	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. Cov ntawv ceeb toom txog kev txiav txim (ACC) thiab cov ntawv sau F1 yog ze li 70%, uas qhia tias peb pom zoo tias cov tshuaj yeeb SLLM tuaj yeem paub zoo txog qhov sib txawv ntawm cov lus hais raws li qhov tseeb. Txawm hais tias SIMA muaj qhov sib txawv me ntsis hais txog qhov tseeb ntawm kev paub lus (WER) piv rau LID-Top, tab sis nws txo qhov ntau ntxiv rau kev siv (invocation costs) los ntawm ze li 0.51 zaug hauv peb qhov	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		datasets, uas txo qhov ntau ntxiv rau kev siv nyob rau hauv kev ua haujlwm.		
Critical	Accuracy/Mistranslation *gibberish*	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. Cov ntawv ceeb toom txog kev txiav txim (ACC) thiab cov ntawv sau F1 yog ze li 70%, uas qhia tias peb pom zoo tias cov tshuaj yeeb SLLM tuaj yeem paub zoo txog qhov sib txawv ntawm cov lus hais raws li qhov tseeb. Txawm hais tias SIMA muaj qhov sib txawv me ntsis hais txog qhov tseeb ntawm kev paub lus (WER) piv rau LID-Top, tab sis nws txo qhov ntau ntxiv rau kev siv (invocation costs) los ntawm ze li 0.51 zaug hauv peb qhov datasets, uas txo qhov ntau ntxiv rau kev siv nyob rau hauv kev ua haujlwm.	25	0.83%
Critical	Accuracy/Mistranslation *gibberish, unintelligible*	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. Cov ntawv ceeb toom txog kev txiav txim (ACC) thiab	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		cov ntawv sau F1 yog ze li 70%, uas qhia tias peb pom zoo tias cov tshuaj yeeb SLLM tuaj yeem paub zoo txog qhov sib txawv ntawm cov lus hais raws li qhov tseeb. Txawm hais tias SIMA muaj qhov sib txawv me ntsis hais txog qhov tseeb ntawm kev paub lus (WER) piv rau LID-Top, tab sis nws txo qhov ntau ntxiv rau kev siv (invocation costs) los ntawm ze li 0.51 zaug hauv peb qhov datasets, uas txo qhov ntau ntxiv rau kev siv nyob rau hauv kev ua haujlwm.		
Critical	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. Txawm hais tias qhov kev kos SIMA tam sim no zoo dua ntau, tab sis tseem tsis zoo ib yam nrog Whisper [6] rau cov ntaub ntawv uas tsis yog qhov tseem ceeb, FLEURS [28]. Qhov kev txwv no ntsig los ntawm peb lub tswv yim uas peb pib tias qhov kev kos SLLM uas peb siv thaum pua tawm tuaj yeem ua haujlwm zoo. Qhov kev kos SLLM uas peb siv thaum pua	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		tawm [29] yog qhov tsis muaj zog ntau dua li cov kos tshwj xeeb xws li Whisper vim tias muaj kev txwv ntawm cov ntaub ntawv uas siv rau kev kos. Hauv lub sijhawm ntxiv, peb npaj yuav siv Whisper [6] ua qhov kos uas peb siv thaum pua tawm thiab txuas ntxiv kho qhov kev kos SIMA kom zoo dua qhov kev kos uas muaj peb kos zoo tshwj xeeb.		
Critical	Fluency/Grammar *unnecessary preposition*	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. Txawm hais tias qhov kev kos SIMA tam sim no zoo dua ntau, tab sis tseem tsis zoo ib yam nrog Whisper [6] rau cov ntaub ntawv uas tsis yog qhov tseem ceeb, FLEURS [28]. Qhov kev txwv no ntsig los ntawm peb lub tswv yim uas peb pib tias qhov kev kos SLLM uas peb siv thaum pua tawm tuaj yeem ua haujlwm zoo. Qhov kev kos SLLM uas peb siv thaum pua tawm [29] yog qhov tsis muaj zog ntau dua li cov kos tshwj xeeb xws li Whisper vim tias muaj kev	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		txwv ntawm cov ntaub ntawv uas siv rau kev kos. Hauv lub sijhawm ntxiv, peb npaj yuav siv Whisper [6] ua qhov kos uas peb siv thaum pua tawm thiab txuas ntxiv kho qhov kev kos SIMA kom zoo dua qhov kev kos uas muaj peb kos zoo tshwj xeeb.		
Critical	Accuracy/Mistranslation *jibberish*	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. Txawm hais tias qhov kev kos SIMA tam sim no zoo dua ntau, tab sis tseem tsis zoo ib yam nrog Whisper [6] rau cov ntaub ntawv uas tsis yog qhov tseem ceeb, FLEURS [28]. Qhov kev txwv no ntsig los ntawm peb lub tswv yim uas peb pib tias qhov kev kos SLLM uas peb siv thaum pua tawm tuaj yeem ua haujlwm zoo. Qhov kev kos SLLM uas peb siv thaum pua tawm [29] yog qhov tsis muaj zog ntau dua li cov kos tshwj xeeb xws li Whisper vim tias muaj kev txwv ntawm cov ntaub ntawv uas siv rau kev kos. Hauv lub sijhawm ntxiv, peb npaj yuav siv Whisper [6]	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		ua qhov kos uas peb siv thaum pua tawm thiab txuas ntxiv kho qhov kev kos SIMA kom zoo dua qhov kev kos uas muaj peb kos zoo tshwj xeeb.		
Critical	Accuracy/Mistranslation *gibberish*	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. Txawm hais tias qhov kev kos SIMA tam sim no zoo dua ntau, tab sis tseem tsis zoo ib yam nrog Whisper [6] rau cov ntaub ntawv uas tsis yog qhov tseem ceeb, FLEURS [28]. Qhov kev txwv no ntsig los ntawm peb lub tswv yim uas peb pib tias qhov kev kos SLLM uas peb siv thaum pua tawm tuaj yeem ua haujlwm zoo. Qhov kev kos SLLM uas peb siv thaum pua tawm [29] yog qhov tsis muaj zog ntau dua li cov kos tshwj xeeb xws li Whisper vim tias muaj kev txwv ntawm cov ntaub ntawv uas siv rau kev kos. Hauv lub sijhawm ntxiv, peb npaj yuav siv Whisper [6] ua qhov kos uas peb siv thaum pua tawm thiab txuas ntxiv kho qhov kev kos SIMA kom zoo dua	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		qhov kev kos uas muaj peb kos zoo tshwj xeeb.		
Critical	Accuracy/Mistranslation *gibberish*	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. Txawm hais tias qhov kev kos SIMA tam sim no zoo dua ntau, tab sis tseem tsis zoo ib yam nrog Whisper [6] rau cov ntaub ntawv uas tsis yog qhov tseem ceeb, FLEURS [28]. Qhov kev txwv no ntsig los ntawm peb lub tswv yim uas peb pib tias qhov kev kos SLLM uas peb siv thaum pua tawm tuaj yeem ua haujlwm zoo. Qhov kev kos SLLM uas peb siv thaum pua tawm [29] yog qhov tsis muaj zog ntau dua li cov kos tshwj xeeb xws li Whisper vim tias muaj kev txwv ntawm cov ntaub ntawv uas siv rau kev kos. Hauv lub sijhawm ntxiv, peb npaj yuav siv Whisper [6] ua qhov kos uas peb siv thaum pua tawm thiab txuas ntxiv kho qhov kev kos SIMA kom zoo dua qhov kev kos uas muaj peb kos zoo tshwj xeeb.	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Critical	Accuracy/Mistranslation *Wrong way to say "In future work"*	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. Txawm hais tias qhov kev kos SIMA tam sim no zoo dua ntau, tab sis tseem tsis zoo ib yam nrog Whisper [6] rau cov ntaub ntawv uas tsis yog qhov tseem ceeb, FLEURS [28]. Qhov kev txwv no ntsig los ntawm peb lub tswv yim uas peb pib tias qhov kev kos SLLM uas peb siv thaum pua tawm tuaj yeem ua haujlwm zoo. Qhov kev kos SLLM uas peb siv thaum pua tawm [29] yog qhov tsis muaj zog ntau dua li cov kos tshwj xeeb xws li Whisper vim tias muaj kev txwv ntawm cov ntaub ntawv uas siv rau kev kos. Hauv lub sijhawm ntxiv, peb npaj yuav siv Whisper [6] ua qhov kos uas peb siv thaum pua tawm thiab txuas ntxiv kho qhov kev kos SIMA kom zoo dua qhov kev kos uas muaj peb kos zoo tshwj xeeb.	25	0.83%
Critical	Accuracy/Mistranslation *gibberish*	Although the current SIMA model significantly improves WER, it still lags behind Whisper [6] on out-of-domain data, FLEURS [28]. This	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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. Txawm hais tias qhov kev kos SIMA tam sim no zoo dua ntau, tab sis tseem tsis zoo ib yam nrog Whisper [6] rau cov ntaub ntawv uas tsis yog qhov tseem ceeb, FLEURS [28]. Qhov kev txwv no ntsig los ntawm peb lub tswv yim uas peb pib tias qhov kev kos SLLM uas peb siv thaum pua tawm tuaj yeem ua haujlwm zoo. Qhov kev kos SLLM uas peb siv thaum pua tawm [29] yog qhov tsis muaj zog ntau dua li cov kos tshwj xeeb xws li Whisper vim tias muaj kev txwv ntawm cov ntaub ntawv uas siv rau kev kos. Hauv lub sijhawm ntxiv, peb npaj yuav siv Whisper [6] ua qhov kos uas peb siv thaum pua tawm thiab txuas ntxiv kho qhov kev kos SIMA kom zoo dua qhov kev kos uas muaj peb kos zoo tshwj xeeb.		
Critical	Accuracy/Mistranslation *gibberish*	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	25	0.83%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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. Txawm hais tias qhov kev kos SIMA tam sim no zoo dua ntau, tab sis tseem tsis zoo ib yam nrog Whisper [6] rau cov ntaub ntawv uas tsis yog qhov tseem ceeb, FLEURS [28]. Qhov kev txwv no ntsig los ntawm peb lub tswv yim uas peb pib tias qhov kev kos SLLM uas peb siv thaum pua tawm tuaj yeem ua haujlwm zoo. Qhov kev kos SLLM uas peb siv thaum pua tawm [29] yog qhov tsis muaj zog ntau dua li cov kos tshwj xeeb xws li Whisper vim tias muaj kev txwv ntawm cov ntaub ntawv uas siv rau kev kos. Hauv lub sijhawm ntxiv, peb npaj yuav siv Whisper [6] u a qhov kos uas peb siv thaum pua tawm thiab txuas ntxiv kho qhov kev kos SIMA kom zoo dua qhov kev kos uas muaj peb kos zoo tshwj xeeb.		
Major	Accuracy/Mistranslation *unnatural use of words*	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty Huab chaws kev siv lus hauv kev ntsuas lus hauv ntau yam lus: Ib txoj kev siv peev xwm uas muaj tus nqi qis, uas tuaj yeem kho kom haum rau qhov nyuaj ntawm kev ntsuas lus.	5	0.17%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Major	Accuracy/Addition *Unnecessary*	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty Huab chaws kev siv lus hauv kev ntsuas lus hauv ntau yam lus: Ib txoj kev siv peev xwm uas muaj tus nqi qis, uas tuaj yeem kho kom haum rau qhov nyuaj ntawm kev ntsuas lus.	5	0.17%
Major	Accuracy/Mistranslation *result of literal English to Hmong translation that completely disregards actual meaning of these words in Hmong*	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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho	5	0.17%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
Major	Accuracy/Addition *unnecessary*	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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-	5	0.17%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Major	Fluency/Grammar *wrong use of the article "qho"*	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	5	0.17%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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.		
		Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Major	Accuracy/Addition *Unnecessary*	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.	5	0.17%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT

Moreover, imbalances in training data between high-resource and low-resource languages further limit the single-model solutions.

Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Major	Fluency/Inconsistency *Unnatural use of words for "state of the art"*	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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov	5	0.17%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Major	Fluency/Grammar *I understand what's being said but the translation is*	Motivated by these limitations, we propose an alternative strategy that selectively invokes models based on the	5	0.17%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
	not natural	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? Yog vim lawv cov kev txwv txhim, peb tau pom ib txoj kev tsis zoo uas siv cov qauv sib txawv raws li qhov tseeb ntawm suab uas tau los. Hauv cov haujlwm kuaj suab, qhov nyuaj ntawm kev kuaj suab muaj qhov sib txawv ntau. Thaum tsis muaj suab thiab siv cov lus qab zib me, cov qauv zoo tshaj plaws thiab cov qauv siv tau feem ntau ua kom muaj qhov tsis raug ntawm cov lus (WER) qis. Tab sis, thaum muaj suab thiab nyuaj rau kuaj suab, qhov WER nce [14, 15, 16, 17], thiab cov qauv zoo tshaj plaws uas muaj zog ua tau zoo dua [6]. Qhov no tau coj mus rau ib lo lus nug tseem ceeb: Peb puas tuaj yeem sib txawv ntawm cov suab uas yooj yim thiab nyuaj, thiab siv peb cov tshuab kuaj suab kom raug?		
Major	Fluency/Inconsistency *Extremely difficult to understand*	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.	5	0.17%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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? Yog vim lawv cov kev txwv txhim, peb tau pom ib txoj kev tsis zoo uas siv cov qauv sib txawv raws li qhov tseeb ntawm suab uas tau los. Hauv cov haujlwm kuaj suab, qhov nyuaj ntawm kev kuaj suab muaj qhov sib txawv ntau. Thaum tsis muaj suab thiab siv cov lus qab zib me, cov qauv zoo tshaj plaws thiab cov qauv siv tau feem ntau ua kom muaj qhov tsis raug ntawm cov lus (WER) qis. Tab sis, thaum muaj suab thiab nyuaj rau kuaj suab, qhov WER nce [14, 15, 16, 17], thiab cov qauv zoo tshaj plaws uas muaj zog ua tau zoo dua [6]. Qhov no tau coj mus rau ib lo lus nug tseem ceeb: Peb puas tuaj yeem sib txawv ntawm cov suab uas yooj yim thiab nyuaj, thiab siv peb cov tshuab kuaj suab kom raug?		
Major	Fluency/Inconsistency *unnecessary*	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,	5	0.17%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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. Cov yeeb yis qhia tias, vim tias siv cov qauv zoo tshaj plaws (SOTA) kom raug, SIMA model muaj kev txhim kho loj heev hauv kev ntsuas lus (WER), yog 18.6%, 9.3%, thiab 28.2% ntau dua qhov qauv pib ntawm peb qhov datasets. Tsis tas li ntawd, piv txwv nrog kev siv qauv los xaus, SIMA ib txwm muaj qhov WER qis dua, nrog kev txhim kho ntawm 6.6%, 4.2%, thiab 16.8%. Qhov tseem ceeb, kev txhim kho ntawm FLEURS dataset yog qhov tseem ceeb tshwj xeeb, vim tias nws tsis yog qhov uas qauv pib siv tau, tab sis qauv LID-Top siv tau. Cov lus teb no qhia meej tias SIMA muaj peev xwm zoo heev los txiav txim tias yuav siv qauv zoo tshaj plaws (SOTA) thaum twg, uas pab txhim kho kev ua haujlwm ntawm kev ntsuas lus (ASR) tag nrho.		
Minor	Fluency/Inconsistency *Awkward use of this phrase for "multilingual"*	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty Huab chaws kev siv lus hauv kev ntsuas lus hauv ntau yam lus: Ib txoj kev	1	0.03%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		siv peev xwm uas muaj tus nqi qis, uas tuaj yeem kho kom haum rau qhov nyuaj ntawm kev ntsuas lus.		
Minor	Terminology *Not fitting for use with "Duab"*	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. Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los	1	0.03%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Minor	Fluency/Inconsistency *Not the typical way to say "for example" in Hmong.*	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	1	0.03%

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		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.		
		Cov qhia lus uas siv tshuaj yeeb (automatic speech recognition, ASR) uas tuaj yeem qhia ntau yam lus tau txais kev pom zoo ntau, vim tias lawv tuaj yeem qhia ntau yam lus siv ib qho qauv xwb [1, 2, 3, 4], raws li qhia hauv Duab 1(a). Cov kev txhim kho tshiab no tau ua kom muaj kev ua haujlwm zoo ntau hauv ntau yam lus los ntawm kev kawm ntawv loj ntau los ntawm kev kawm ntawv uas muaj kev saub npe (supervised) lossis kev kawm ntawv uas tsis muaj kev saub npe (self-supervised) [3, 5, 6, 7, 8, 9, 10, 11, 12]. Ua ib qho piv txwv, Whisper [6] tau kawm ntawv los ntawm 680,000 teej ntawm cov ntaub ntawv uas muaj ntau yam lus, uas ua rau nws tuaj yeem siv tau zoo hauv ntau		

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		yam kev sim ASR, thaum USM [9] siv 12 lab teej ntawm cov ntaub ntawv uas tsis muaj kev saub npe los ua kom muaj kev ua haujlwm zoo hauv ntau yam lus. Txawm hais tias muaj cov kev txhim kho no, kev siv cov tshuaj yeeb ASR uas muaj ntau yam lus thiab siv ib qho qauv xwb tseem muaj qhov tseem ceeb. Cov kev txawj ntseeg sib txawv, cov qauv lus sib txawv, thiab cov lus siv sib txawv hauv ntau yam lus ua rau nws nyuaj los ua kom muaj kev ua haujlwm zoo tshwj xeeb rau txhua yam lus. Ntxiv rau, qhov sib txawv ntawm cov ntaub ntawv kawm ntawv ntawm cov lus uas muaj ntau cov ntaub ntawv thiab cov lus uas muaj me ntsis cov ntaub ntawv ntxiv ua rau qhov kev daws teb siv ib qho qauv xwb nyuaj dua.		
Minor	Fluency/Inconsistency *Not the appropriate verb to use with "Duab"*	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]. Ib txoj kev uas feem ntau siv los pab cov teeb meem no yog siv ib qho kev kos lus (language identification,	1	`0.03%`

SEVERITY	CATEGORY	SOURCE / TARGET	PENALTY	IMPACT
		LID) kom pom tias lus twg tau hais ua ntej, ces siv qhov kos lus uas zoo tshaj plaws (state-of-the-art, SOTA) kom sau lus, raws li qhia hauv Duab 1(b). Tab sis, txoj kev no muaj qee qhov tsis zoo. Ntau qhov kos lus zoo tshaj plaws yog khoom lag luam [12] thiab muaj tus nqi siv, raws li qhov ntau npaum li cas koj siv, uas ua rau txoj kev no tsis muaj txiaj. Tsis tas li ntawd, yog hais tias LID pom lus tsis raug, nws yuav ua rau siv qhov kos lus tsis raug, uas yuav ua rau tsis zoo rau cov neeg siv.		