Think Carefully and Check Again! Meta-Generation Unlocking LLMs for Low-Resource Cross-Lingual Summarization

Anonymous ACL submission

Abstract

Cross-lingual summarization (CLS) aims to generate a summary for the source text in a different target language. Currently, instruction-tuned large language models (LLMs) excel at various English tasks. However, unlike languages such as English, Chinese or Spanish, for those relatively low-resource languages with limited usage or data, recent studies have shown that LLMs' performance on CLS tasks remains unsatisfactory even with few-shot settings. This raises the question: *Are LLMs capable of handling cross-lingual summarization tasks for low-resource languages?* To resolve this question, we fully explore the potential of large language models on cross-lingual summarization task for low-resource languages through our four-step zero-shot method: **SUMMARIZATION**, **IMPROVEMENT**, **TRANSLATION** and **REFINEMENT** (SITR) with correspondingly designed prompts. We test our proposed method with multiple LLMs on two well-known cross-lingual summarization datasets with various low-resource target languages. The results show that: i) GPT-3.5 and GPT-4 significantly and consistently outperform other baselines when using our zero-shot SITR methods. ii) By employing our proposed method, we unlock the potential of LLMs, enabling them to effectively handle cross-lingual summarization tasks for relatively low-resource languages.

1 Introduction

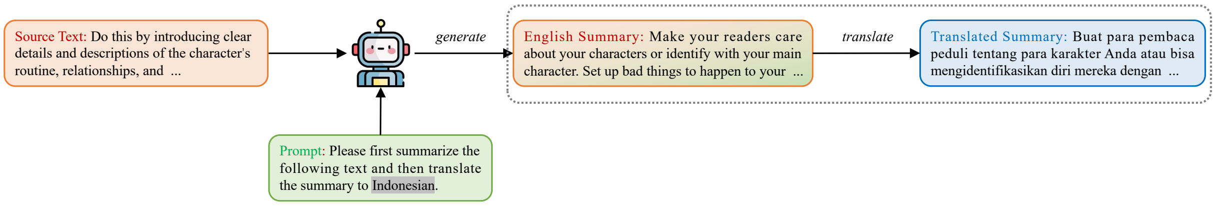
Cross-lingual summarization refers to summarizing the source text in another target language. Traditionally, CLS is approached through one of two methods: summarize-translate (see LLM implementation in Figure 1) or translate-summarize (Leuski et al., 2003; Orăsan and Chiorean, 2008). In the summarize-translate method, the text is first summarized in the source language and then translated into the target language. The translate-summarize method reverses this order. Both approaches, however, are prone to error accumulation during the two-step process, which can significantly degrade the final output quality.

With the advent of the Transformer architecture (Vaswani et al., 2017), end-to-end multilingual models like mBART (Liu et al., 2020), mBART-50 (Tang et al., 2020), and mT5 (Xue et al., 2020) have been developed and applied to CLS tasks. However, these models often require extensive fine-tuning, especially when applied to low-resource languages with limited pre-training data (Parnell et al., 2024).

In recent years, large language models (LLMs) such as GPT-2, InstructGPT, GPT-4, and Llama (Radford et al., 2019; Brown et al., 2020; Ouyang et al., 2022; OpenAI et al., 2024; Dubey et al., 2024) have shown significant potential for CLS tasks due to their extensive training on vast multilingual data. These models have achieved strong performance in high-resource languages like English, Chinese, and German (Wang et al., 2023) by implementing summarize-translate method. However, their effectiveness in low-resource languages remains limited, even when using few-shot learning techniques (Park et al., 2024).

This limitation underscores a critical area of research that has not yet been fully explored: whether LLMs can be effectively adapted for cross-lingual summarization tasks in low-resource languages, and if so, how effective they can be. Addressing this gap is crucial for extending the benefits of LLMs to a broader range of linguistic communities, making it an important area for further investigation.

To address these challenges, we propose a four-step zero-shot approach, *Summarization*, *Improvement*, *Translation*, and *Refinement* (SITR) — designed to unlock the full potential of LLMs for CLS tasks in low-resource languages. Our method mitigates the issues of traditional pipelines by incorporating meta-generation strategies, which allows LLMs to learn from feedback and use refiners to produce more accurate and reliable outputs



Source Text: Do this by introducing clear details and descriptions of the character's routine, relationships, and ...

Prompt: Please first summarize the following text and then translate the summary to Indonesian.

English Summary: Make your readers care about your characters or identify with your main character. Set up bad things to happen to your ...

Translated Summary: Buat para pembaca peduli tentang para karakter Anda atau bisa mengidentifikasikan diri mereka dengan ...

Figure 1: An example of single-step summarize-translate method for cross-lingual summarization.

(Welleck et al., 2024).

We test our method using GPT-3.5 and GPT-4 on two major cross-lingual summarization datasets, comparing them with fine-tuned models like mBART-50 (Tang et al., 2020), mT5 (Xue et al., 2020), and other LLM baselines employing few-shot and summarize-translate approaches. Results show that GPT-3.5 and GPT-4 significantly outperform other LLM baselines across four metrics and even surpass fine-tuned models on most low-resource languages, demonstrating LLMs' strong capability in cross-lingual summarization under our proposed method.

We also apply our method to mainstream LLMs such as Llama3 (Dubey et al., 2024), Gemma2 (Team et al., 2024), Mixtral (Jiang et al., 2024), and Qwen-1.5 (Bai et al., 2023), observing that most achieve impressive scores against the powerful GPT-4o model. This further validates the effectiveness of our approach and highlights that today's LLMs are capable of performing well on cross-lingual summarization tasks, even with low-resource languages.

In summary, this paper has the following contributions:

- Our proposed zero-shot SITR method enhances large language models' performance on cross-lingual summarization for low-resource languages, demonstrating strong robustness across different models, datasets, and target languages.

- Extensive experiments on two datasets and various low-resource languages reveal that our method significantly outperforms other LLM baselines and surpasses fine-tuned models.

- To the best of our knowledge, we are the first to evaluate various LLMs on cross-lingual summarization for low-resource languages, showing that they possess the capability to achieve impressive results in this domain.

2 Methodology

2.1 SITR (Two-Stage Meta-Generation)

In this paper, we propose a four-step zero-shot SITR method for cross-lingual summarization in low-resource languages (see Figure 2), comprising **SUMMARIZATION**, **IMPROVEMENT**, **TRANSLATION** and **REFINEMENT**. The **IMPROVEMENT** and **REFINEMENT** stages align with two-stage meta-generation, involving LLM strategies like feedback learning, and rethinking (Welleck et al., 2024). To maximize LLMs' potential, we design specific prompts for each step, guiding the models to generate reliable outputs and minimizing error accumulation.

[SUMMARIZATION]. LLMs should distill the long input source text (I) into concise summary (S). To counter their tendency to generate overly detailed summaries, we use a summarization prompt $(\mathrm{P_{sum}})$ (see Figure 6) to focus their output on the core essence of the text, ensuring the summary is both precise and relevant without unnecessary elaboration.

$$\mathrm{S} = \mathrm{LLM}(\mathrm{I};\ \mathrm{P_{sum}}) \tag{1}$$

[IMPROVEMENT]. The first stage of meta-generation, providing large language models with the input source text (I), the initial summary (S) from the **SUMMARIZATION** step, and the improvement prompt $(\mathrm{P_{imp}})$ (see Figure 7) to recheck and optimize the summary (S^*). This step reduces error accumulation by enabling self-improvement, preparing the more accurate summary for the next step of translation.

$$\mathrm{S}^* = \mathrm{LLM}(\mathrm{I};\ \mathrm{S};\ \mathrm{P_{imp}}) \tag{2}$$

[TRANSLATION]. Using the translation prompt $(\mathrm{P_{tra}})$ (see Figure 8), the optimized summary (S^*) after the **IMPROVEMENT** step is translated into the text (T) in low-resource target language. Due to limited training data and lack of confidence, LLMs often produce redundant and messy outputs in these

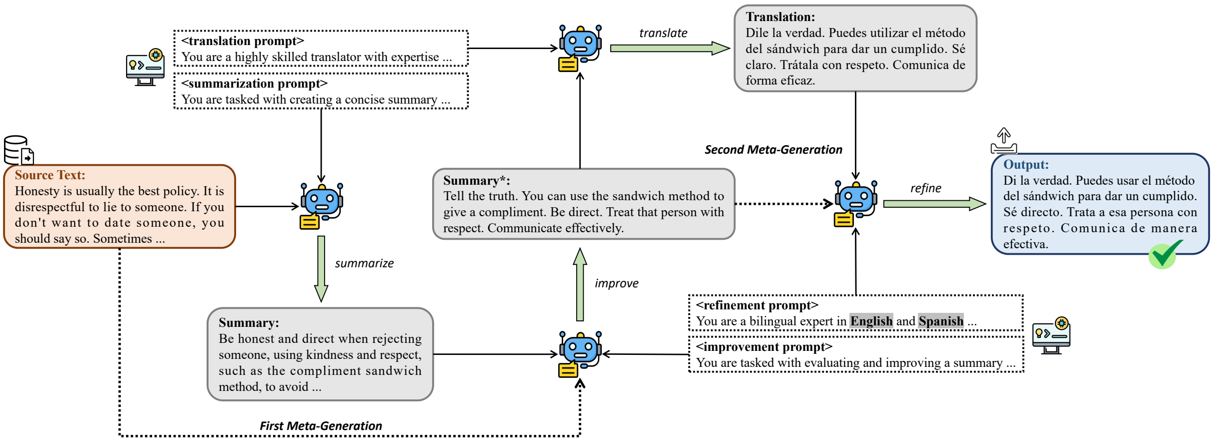

<translation prompt>
You are a highly skilled translator with expertise ...
<summarization prompt>
You are tasked with creating a concise summary ...
translate
Translation:
Dile la verdad. Puedes utilizar el método del sándwich para dar un cumplido. Sé claro. Trátala con respeto. Comunica de forma eficaz.
Second Meta-Generation
Source Text:
Honesty is usually the best policy. It is disrespectful to lie to someone. If you don't want to date someone, you should say so. Sometimes ...
Summary*:
Tell the truth. You can use the sandwich method to give a compliment. Be direct. Treat that person with respect. Communicate effectively.
refine
Output:
Di la verdad. Puedes usar el método del sándwich para dar un cumplido. Sé directo. Trata a esa persona con respeto. Comunica de manera efectiva.
summarize
improve
<refinement prompt>
You are a bilingual expert in English and Spanish ...
<improvement prompt>
You are tasked with evaluating and improving a summary ...
Summary:
Be honest and direct when rejecting someone, using kindness and respect, such as the compliment sandwich method, to avoid ...
First Meta-Generation

Figure 2: The architecture of our four-step zero-shot SITR method for cross-lingual summarization.

Dataset	Src Lang.	Trg Lang.	Domain	Train / Validation / Test
CrossSum	English	Uk & Bn & Id & Gu	News	1000 / 150 / 50
		Pa		769 / 100 / 50
WikiLingua	English	Id & Vi & Ar & Hi & Th	How-to Guide	1000 / 150 / 50

Table 1: The source and amount of experimental data. The abbreviations of the languages correspond to their full names: Uk(Ukarainian), Bn(Bengali), Id(Indonesian), Gu(Gujarati), Pa(Pashto), Vi(Vietnamese), Ar(Arabic), Hi(Hindi), Th(Thai).

languages. This step aims to produce more reliable translations to ease the subsequent process.

$$\mathrm{T} = \mathrm{LLM}(\mathrm{S}^{*};\, \mathrm{P}_{\mathrm{tra}}) \tag{3}$$

[REFINEMENT]. The optimized summary (S^{*}) after the **IMPROVEMENT** step, the initial translation (T) from the **TRANSLATION** step, and the refinement prompt $(\mathrm{P}_{\mathrm{ref}})$ (see Figure 9) are combined and input into the LLMs for self-correction to generate the final output (O). This process constitutes the second stage of meta-generation, enabling the LLMs to revise and produce a more accurate translation through re-evaluation.

$$\mathrm{O} = \mathrm{LLM}(\mathrm{S}^{*};\, \mathrm{T};\, \mathrm{P}_{\mathrm{ref}}) \tag{4}$$

Our proposed method generally involves four steps to leverage the large language model's inherent capabilities. For summarization or translation tasks where a perfect result cannot be achieved in a single attempt, we utilize meta-generation to enable the LLMs to self-reflect and improve their final output. Additionally, when the model lacks guidance or confidence, we use strategic prompts to prevent disorganized or unreliable results. This approach ensures that the large language model produces high-quality and coherent outputs through the implementation of two-stage meta-generation.

2.2 Large Language Models

In this paper, we conduct a thorough evaluation of various large language models using our proposed SITR method (Detail information in Appendix A).

Closed-Source Models. We utilize four different models developed by OpenAI, including the latest GPT-4O and GPT-4O-MINI.

Open-Source Models. We conduct our experiments on LLAMA3 and LLAMA3.1 (Touvron et al., 2023; Dubey et al., 2024) developed by MetaAI; QWEN-1.5 and QWEN2 trained by Alibaba Cloud (Bai et al., 2023); GEMMA and GEMMA2 created by Google (Team et al., 2024) and MIXTRAL from Mistral AI (Jiang et al., 2024).

3 Experiments

3.1 Datasets & Langueges

Datasets. In our research, we conduct experiments on two popular cross-lingual summarization datasets: **CrossSum** (Hasan et al., 2021) and **Wikilingua** (Ladhak et al., 2020).

For fine-tuning experiments, we randomly select a subset from the training split of each dataset. For evaluation, we consistently use 50 randomly chosen samples from the test split to assess different methods and large language models.

Languages. We consider the data ratio from the CommonCrawl corpus[1] and its intersection with two datasets, aligning with the languages used in the previous study (Park et al., 2024). Based on our research focus, we choose five challenging low-resource languages for each dataset to conduct our experiments. Detailed information about our experimental data is provided in Table 1 and more experimental languages are shown in Appendix C.

3.2 Metrics

In our experiments, we use **ROUGE-1/2/L** (Lin, 2004) and **BERTScore** (Zhang* et al., 2020) as four different metrics.

ROUGE metrics evaluate lexical overlap between the generated summaries and their references by considering unigrams, bigrams, and the longest common subsequence. BERTScore metric, however, focuses on measuring semantic similarity between two texts. We compute ROUGE scores with the multi-lingual ROUGE toolkit[2], and BERTScore is calculated using the bert-score toolkit[3].

3.3 Baselines

We select fine-tuned mBART-50, mT5-small, and mT5-base as baselines to demonstrate the capabilities of the fine-tuned encoder-decoder models on cross-lingual summarization tasks for low-resource languages.

For LLM-related baselines, we employ few-shot learning method with GPT-3.5 and GPT-4 following the prompt (see Figure 11) from a previous paper (Park et al., 2024). Besides, we also evaluate the single-step summarize-translate method (see Figure 6) as a baseline (Wang et al., 2023). (All implementation details are shown in Appendix B).

3.4 Experiment Results

The main experimental results on the CrossSum dataset are presented in Table 2. We compare our zero-shot SITR method with three types of baselines: fine-tuned encoder-decoder models, few-shot learning, and summarize-translate LLMs across various low-resource languages. Table 3 shows the main results for the WikiLingua dataset. (More experimental results are shown in Appendix C).

To further explore the potential of current large language models for cross-lingual summarization of low-resource languages and assess the robustness of our SITR architecture, we conduct extensive experiments with our method on various large language models. The results are presented in Table 4 and Table 5.

SITR vs Fine-tuned Models. Table 2 and Table 3 show that mT5-small and mT5-base both perform poorly on low-resource languages, even after fine-tuning with approximately 1,000 data points. While mBART-50 achieves better results, it still lags behind our zero-shot SITR method across almost all languages, except for *Pashto*, where fine-tuned mBART-50 has a slightly higher score. Notably, fine-tuning an encoder-decoder model for each low-resource language is significantly more costly than using large language models with our proposed SITR method.

SITR vs LLM Baselines. Table 2 and Table 3 demonstrate that under our approach, the outputs of the large language models significantly outperform other baselines in terms of both ROUGE and BERTScore metrics. This demonstrates that our outputs not only capture the key information of the text but also show notable improvements in word choice and semantic information.

On the CrossSum dataset, SITR improves the sum of ROUGE-1/2/L scores from 18.83 to 33.51 (a 78% increase) with GPT-3.5 and from 22.56 to 34.54 (a 53% increase) with GPT-4, compared to two-shot generation. The improvement over the summarize-translate method is even more notable, with the sum of ROUGE-1/2/L scores increasing by 103% (from 16.54 to 33.51) with GPT-3.5 and 98% (from 17.46 to 34.54) with GPT-4.

On the WikiLingua dataset, the sum of ROUGE-1/2/L scores improves by 44% (from 34.12 to 49.02) and 21% (from 37.34 to 45.38) when comparing two-shot generation to our zero-shot method. Additionally, the increases are 52% (from 32.34 to 49.02) and 42% (from 32.03 to 45.38) when compared to the summarize-translate method.

For BERTScore, our method shows an increase of 2 to 3 percentage points compared to two-shot generation, and 4 to 5 percentage points compared to the summarize-translate method. This significant improvement reflects a substantial enhancement in the semantic quality of the model's outputs.

Improvement via Language. According to Table

[1] http://commoncrawl.org

[2] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

[3] https://github.com/Tiiiger/bert_score

Model		Language Pair																				Average Score				
		English⇒Ukrainian				English⇒Bengali				English⇒Indonesian				English⇒Gujarati				English⇒Pashto ♠								
		R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	S-R	BS
mBART-50	0-shot	0.50	0.05	0.50	61.05	0.00	0.00	0.00	56.17	2.49	0.34	2.20	62.64	0.08	0.00	0.08	58.13	0.00	0.00	0.00	54.44	0.61	0.08	0.56	1.25	58.49
	1000-shot	13.46	3.17	10.03	67.64	9.74	1.74	8.86	67.60	18.20	3.23	14.36	68.87	11.23	1.56	10.26	67.97	**23.63**	**5.36**	**16.41**	**74.12**	15.25	3.01	11.98	30.24	69.24
mT5-small	1000-shot	1.39	0.00	1.21	49.72	1.46	0.00	1.42	56.19	2.89	0.00	2.61	54.21	0.35	0.00	0.33	49.58	5.58	0.02	5.03	61.43	2.33	0.00	2.12	4.45	54.23
mT5-base	1000-shot	1.72	0.00	1.57	49.39	1.76	0.07	1.67	60.43	3.72	0.00	3.47	56.73	1.09	0.00	1.09	50.74	6.11	0.05	5.50	61.66	2.88	0.02	2.66	5.56	55.79
GPT-3.5 (Park et al., 2024)	zero-shot	10.83	1.26	6.78	65.74	5.69	0.60	2.95	60.93	11.36	1.93	7.58	66.19	6.90	1.21	3.97	65.05	3.25	0.37	2.77	60.87	7.61	1.07	4.81	13.49	63.76
	one-shot	12.34	1.97	6.66	66.28	7.65	1.15	4.41	62.70	14.17	3.40	9.38	68.09	6.88	1.32	4.60	66.06	7.13	0.50	5.92	64.56	9.63	1.67	6.19	17.49	65.54
	two-shot	13.48	1.51	6.57	66.02	8.50	0.89	5.47	65.65	14.38	3.12	10.19	68.21	9.47	1.64	6.68	67.10	6.23	0.66	5.41	63.84	10.41	1.56	6.86	18.83	66.16
GPT-4 (Park et al., 2024)	zero-shot	8.75	1.91	5.75	65.35	8.51	1.31	5.74	65.21	8.94	1.84	6.14	65.70	8.14	1.09	6.00	66.85	10.10	2.06	7.29	68.27	8.89	1.64	6.18	16.71	66.28
	one-shot	13.74	2.47	8.70	67.76	10.25	1.25	6.04	66.34	10.55	1.80	6.19	66.84	9.94	1.60	5.94	67.61	13.48	2.79	9.83	69.08	11.59	1.98	7.34	20.91	67.53
	two-shot	13.40	2.25	8.42	67.87	10.97	1.96	7.23	67.09	12.60	2.34	8.57	67.41	10.20	1.80	6.23	68.67	14.63	2.92	9.30	68.82	12.36	2.25	7.95	22.56	67.97
GPT-3.5 w/ summarize-translate	zero-shot	13.32	2.40	9.03	67.54	9.81	0.85	6.62	64.81	12.72	2.33	7.69	67.40	7.17	0.76	5.68	65.71	2.14	0.38	1.82	58.60	9.03	1.34	6.17	16.54	64.81
GPT-4 w/ summarize-translate	zero-shot	9.79	2.12	6.45	65.54	9.17	1.61	5.95	65.45	9.14	1.59	6.11	65.84	7.77	1.19	5.72	66.46	10.55	2.58	7.55	66.29	9.28	1.82	6.36	17.46	65.92
GPT-3.5 w/ SITR (Ours)	zero-shot	**18.77**	**4.36**	**13.88**	**69.63**	14.28	2.74	10.16	**69.47**	20.65	4.44	**15.57**	**69.16**	**14.58**	**2.58**	**12.06**	**70.41**	17.06	2.56	13.84	71.29	17.07	3.34	**13.10**	33.51	69.99
GPT-4 w/ SITR (Ours)	zero-shot	17.04	4.24	11.74	68.51	**14.60**	**3.16**	**10.23**	69.39	**20.86**	**4.69**	14.59	69.05	14.24	2.45	11.95	70.28	21.74	5.23	15.96	73.43	**17.70**	**3.95**	12.89	**34.54**	**70.13**

Table 2: Experimental results on the CrossSum dataset. R-1, R-2, R-L, S-R and BS refer to ROUGE-1, ROUGE-2, ROUGE-L, sum of ROUGE-1/2/L and BERTScore respectively. The task with ♠ means training data less than 1000, where 1000-shot setting equals full fine-tuning, as the information shown in Table 1. The best result on every target language is highlighted in **bold** font, and the second best result is marked with an underline.

Model		Language Pair																				Average Score				
		English⇒Indonesian				English⇒Vietnamese				English⇒Arabic				English⇒Hindi				English⇒Thai								
		R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	S-R	BS
mBART-50	0-shot	2.48	0.22	1.93	63.74	0.71	0.08	0.66	63.43	0.23	0.05	0.21	61.77	1.46	0.34	1.37	58.95	11.14	1.14	10.09	59.27	3.20	0.37	2.85	6.42	61.43
	1000-shot	17.75	3.86	12.13	68.55	13.68	3.88	9.87	67.86	12.33	2.22	9.09	69.04	22.00	4.68	15.14	67.19	26.07	5.88	19.79	69.45	18.37	4.10	13.20	35.67	68.42
mT5-small	1000-shot	0.52	0.00	0.50	51.87	0.36	0.00	0.36	55.24	0.00	0.00	0.00	55.77	8.54	0.25	7.95	63.61	5.58	0.07	4.44	53.81	3.00	0.06	2.65	5.71	56.06
mT5-base	1000-shot	2.06	0.00	1.90	53.63	0.62	0.00	0.54	52.31	2.26	0.00	2.06	53.66	8.89	0.12	8.14	60.59	13.41	0.00	12.69	54.28	5.45	0.02	5.07	10.54	54.89
GPT-3.5 (Park et al., 2024)	zero-shot	12.90	2.21	9.51	68.06	18.22	6.43	12.72	69.42	7.45	1.52	4.50	66.42	16.63	4.74	10.95	66.59	13.65	4.74	10.54	68.71	13.77	3.93	9.64	27.34	67.84
	one-shot	16.26	3.47	10.90	69.08	22.77	8.61	16.62	71.14	9.94	1.84	6.77	68.20	17.53	4.35	11.03	67.45	14.18	4.61	10.79	68.74	16.14	4.58	11.22	31.94	68.92
	two-shot	17.01	3.54	11.79	68.16	23.94	9.07	15.58	71.65	10.79	2.49	7.07	68.24	17.24	5.11	12.72	68.56	18.12	4.73	11.40	69.77	17.42	4.99	11.71	34.12	69.28
GPT-4 (Park et al., 2024)	zero-shot	13.75	2.98	9.78	67.81	16.44	6.40	11.83	68.34	8.49	1.45	5.21	66.46	16.76	4.32	10.60	66.62	19.28	5.86	14.86	69.47	14.94	4.20	10.46	29.60	67.74
	one-shot	17.74	3.02	14.15	68.72	17.80	6.73	12.92	69.17	11.74	2.04	7.95	68.46	18.10	4.39	11.08	67.98	23.42	6.32	18.00	70.47	17.76	4.50	12.82	35.08	68.96
	two-shot	18.03	3.05	13.26	68.89	20.31	6.44	13.22	70.48	13.21	2.80	9.22	69.93	19.79	5.01	12.88	68.28	24.35	6.21	18.91	70.21	19.14	4.70	13.50	37.34	69.56
GPT-3.5 w/ summarize-translate	zero-shot	15.63	2.32	9.55	65.38	20.53	8.37	13.95	69.86	9.12	1.05	5.37	67.19	21.13	3.67	12.74	68.18	18.95	5.08	14.22	67.06	17.07	4.10	11.17	32.34	67.53
GPT-4 w/ summarize-translate	zero-shot	13.71	3.80	10.42	66.70	19.27	7.44	13.28	69.60	8.79	1.87	7.35	67.01	17.59	4.52	13.01	67.57	18.76	6.01	14.35	67.96	15.62	4.73	11.68	32.03	67.77
GPT-3.5 w/ SITR (Ours)	zero-shot	**20.40**	**4.65**	**15.74**	**69.98**	**30.85**	**12.36**	**22.26**	**72.60**	14.38	2.88	**12.22**	**71.98**	**24.66**	**5.49**	**18.16**	**70.90**	**30.28**	**7.60**	**23.19**	**71.77**	**24.11**	**6.60**	**18.31**	**49.02**	**71.45**
GPT-4 w/ SITR (Ours)	zero-shot	18.67	3.94	14.43	69.12	28.67	10.57	18.62	72.08	**15.32**	**3.67**	11.71	71.06	23.77	4.74	16.86	69.36	28.76	6.57	20.59	71.01	23.04	5.90	16.44	45.38	70.53

Table 3: Experimental results on the WikiLingua dataset. R-1, R-2, R-L, S-R and BS refer to ROUGE-1, ROUGE-2, ROUGE-L, sum of ROUGE-1/2/L and BERTScore respectively. The best result on every target language is highlighted in **bold** font, and the second best result is marked with an underline.

2 and Table 3, our proposed SITR method achieves a smaller percentage improvement on the WikiLingua dataset compared to CrossSum. This discrepancy may stem from the fact that while the target languages in WikiLingua are still low-resource, they are relatively more studied, providing greater resources for pre-trained LLMs. As a result, our method tends to yield more significant improvements in languages with fewer available resources. (Additional results can be found in Appendix C.)

Robustness of SITR and Capabilities of LLMs. Table 2 and Table 3 demonstrate the impressive performance of GPT-3.5 and GPT-4 using our SITR method. Additionally, Table 4 and Table 5 show how our SITR method effectively leverages large language models, allowing many of them to excel in cross-lingual summarization, even for low-resource languages. Notable examples in-

Model	Language Pair																				Average Score			
	English⇒Ukrainian				English⇒Bengali				English⇒Indonesian				English⇒Gujarati				English⇒Pashto							
	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
GPT-4o	17.48	4.38	12.85	68.56	**16.94**	**3.96**	**11.84**	**69.86**	22.42	5.03	16.37	69.64	**15.76**	**3.59**	**13.04**	**71.09**	**22.80**	4.38	**17.12**	**75.34**	**19.08**	**4.27**	**14.24**	**70.90**
GPT-4o-mini	17.41	4.38	12.20	68.73	15.34	2.90	10.26	69.01	18.54	3.04	13.59	69.78	13.49	2.60	11.16	70.58	21.99	**4.75**	15.68	73.30	17.35	3.53	12.58	70.28
Mixtral-8x22B-instruct	17.06	4.63	12.52	68.48	12.48	1.92	9.65	68.61	14.28	3.47	10.35	68.64	8.35	0.76	6.82	68.60	6.51	0.63	5.11	63.36	11.74	2.28	8.89	67.54
Mixtral-8x7B-32768	15.73	3.75	11.49	68.85	6.49	0.73	4.91	63.18	18.14	3.19	13.72	70.80	6.63	0.29	5.69	66.27	9.22	0.40	7.73	64.92	11.24	1.67	8.71	66.80
Qwen2-72B	**18.02**	4.28	12.02	69.73	14.80	2.50	10.78	69.29	20.32	5.18	14.87	71.19	9.44	1.16	7.51	68.68	13.31	0.90	10.19	67.34	15.18	2.80	11.07	69.25
Qwen1.5-110B-Chat	12.76	2.83	10.08	66.77	7.51	1.04	5.32	65.54	15.98	3.58	11.75	69.08	8.22	0.92	6.82	68.01	11.81	1.19	8.88	65.43	11.26	1.91	8.57	66.97
Qwen1.5-72B-Chat	17.70	3.34	12.47	68.86	9.84	1.30	7.45	67.42	21.79	4.51	15.35	71.53	10.96	1.16	9.01	69.02	11.40	1.33	9.31	65.91	14.34	2.33	10.72	68.55
Llama3-8B-8192	9.92	2.29	7.08	65.40	9.83	1.78	6.67	65.86	14.59	3.28	11.32	68.08	10.21	1.77	8.76	67.23	6.52	0.90	5.57	62.67	10.21	2.00	7.88	65.85
Llama-3.1-8B-instant	15.35	3.39	11.22	68.89	14.42	2.33	9.64	68.42	21.64	4.56	14.79	71.44	9.57	0.84	7.15	66.42	9.35	1.44	7.17	65.05	14.07	2.51	9.99	68.04
Llama3-70B-8192	9.16	1.87	7.12	64.20	6.59	1.71	4.34	65.53	16.46	4.37	12.10	69.09	8.34	1.59	6.62	65.88	13.68	2.78	10.54	69.12	10.85	2.46	8.14	66.76
Llama-3.1-70B-versatile	18.01	**4.74**	**13.51**	69.70	16.04	3.61	11.16	69.03	21.59	5.48	15.17	71.18	14.03	3.25	11.23	70.25	18.45	2.87	13.89	72.46	17.62	3.99	12.99	70.52
Gemma-7B-it	12.93	1.86	10.48	68.48	12.67	1.44	9.45	68.95	20.23	3.44	15.78	71.29	7.67	0.89	6.60	67.54	1.46	0.00	1.46	58.71	10.99	1.53	8.75	66.99
Gemma2-9B-it	16.93	3.84	13.24	69.37	13.47	1.82	9.57	69.27	21.62	4.89	16.43	71.98	14.11	3.10	11.60	70.39	13.13	1.19	9.98	68.47	15.85	2.97	12.16	69.90
Gemma2-27B	17.08	4.03	13.31	**69.87**	15.28	2.71	10.49	69.81	**23.28**	**5.58**	**17.84**	**72.50**	14.72	3.36	12.04	70.97	16.18	1.66	11.38	71.35	17.31	3.47	13.01	**70.90**

Table 4: Performance of various LLMs using our proposed SITR method on CrossSum dataset. R-1, R-2, R-L and BS refer to ROUGE-1, ROUGE-2, ROUGE-L and BERTScore respectively. Light blue and blue denotes models inference through llama-api and groq. The best result on every target language is highlighted in **bold** font, and the second best result is marked with an underline.

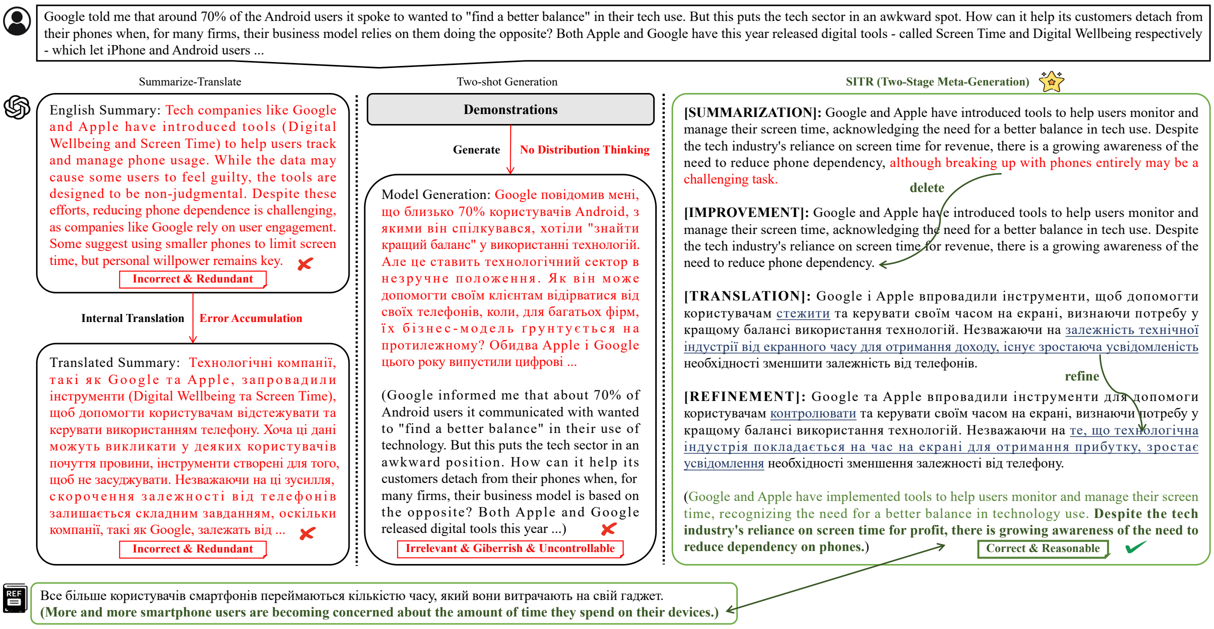
Model	Language Pair																				Average Score			
	English⇒Indonesian				English⇒Vietnamese				English⇒Arabic				English⇒Hindi				English⇒Thai							
	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
GPT-4o	**22.31**	4.63	**16.25**	**70.31**	**30.95**	12.44	**22.79**	**72.63**	**15.22**	**3.43**	**11.85**	**71.74**	**26.07**	4.06	17.80	69.78	**30.87**	**9.06**	**23.38**	**72.52**	**25.08**	6.72	**18.41**	**71.40**
GPT-4o-mini	21.62	4.32	16.05	69.94	28.12	11.56	19.20	71.74	13.44	2.96	10.07	70.94	23.74	3.82	15.50	68.01	29.11	6.04	21.05	72.39	23.21	5.74	16.37	70.60
Mixtral-8x22B-instruct	19.54	**5.70**	15.20	70.10	27.10	9.68	19.00	71.36	12.43	2.88	10.36	70.23	21.28	4.57	15.44	67.86	27.44	7.31	20.83	70.29	21.56	6.03	16.17	69.97
Mixtral-8x7B-32768	19.76	4.18	14.82	69.21	23.52	6.67	15.76	70.82	8.24	0.57	6.79	69.00	19.98	2.81	13.77	67.75	22.52	5.57	17.89	69.30	18.80	3.96	13.81	69.22
Qwen2-72B	17.54	3.84	13.21	68.68	30.09	11.40	21.08	72.22	13.69	2.06	9.58	71.21	22.96	3.52	15.08	67.74	30.25	8.49	22.40	71.42	22.91	5.86	16.27	70.25
Qwen1.5-110B-Chat	19.24	5.19	14.47	69.37	29.89	10.81	20.19	72.37	12.38	2.09	9.51	70.49	25.57	4.74	17.62	**70.12**	26.43	6.75	19.77	70.17	22.70	5.92	16.31	70.50
Qwen1.5-72B-Chat	19.76	3.81	14.37	69.67	30.66	12.23	21.38	72.60	12.41	2.50	10.16	70.20	21.89	2.26	14.71	68.69	28.82	7.46	22.36	70.34	22.71	5.65	16.60	70.30
Llama3-8B-8192	16.78	4.02	12.75	67.63	22.91	7.79	15.89	69.77	9.39	1.21	7.82	67.62	20.86	4.59	15.43	67.26	23.21	5.85	18.28	68.02	18.63	4.69	14.03	68.06
Llama-3.1-8B-instant	20.34	4.67	14.15	69.63	29.92	10.71	20.49	72.20	10.20	0.89	8.69	68.79	24.26	5.34	17.05	68.42	23.61	6.11	17.66	69.57	21.67	5.54	15.61	69.72
Llama3-70B-8192	9.43	1.61	7.12	64.93	11.10	4.62	7.76	66.32	5.68	1.26	4.57	66.61	9.77	1.91	7.09	61.34	20.29	6.02	15.39	68.98	11.25	3.08	8.39	65.64
Llama-3.1-70B-versatile	20.96	5.04	14.27	69.79	30.11	**12.65**	20.19	71.97	12.61	2.34	10.19	69.28	26.04	**6.52**	**19.33**	69.72	29.01	7.66	22.27	71.01	23.75	**6.84**	17.25	70.35
Gemma-7B-it	18.15	2.94	14.73	69.39	23.14	8.04	16.64	70.85	6.61	0.56	5.81	68.51	20.09	2.32	14.74	67.08	20.10	4.15	15.32	68.60	17.62	3.60	13.45	68.89
Gemma2-9B-it	20.64	4.71	15.50	69.77	27.52	10.94	19.74	71.95	11.82	1.73	9.04	69.60	25.68	3.91	17.49	68.91	26.41	5.69	19.99	70.57	22.41	5.40	16.35	70.16
Gemma2-27B	20.38	3.84	14.39	69.93	29.39	10.92	20.02	72.38	12.17	1.59	9.92	69.58	25.53	4.20	18.03	68.91	27.95	5.66	20.93	70.68	23.08	5.24	16.66	70.30

Table 5: Performance of various LLMs using our proposed SITR method on WikiLingua dataset. R-1, R-2, R-L and BS refer to ROUGE-1, ROUGE-2, ROUGE-L and BERTScore respectively. Light blue and blue denotes models inference through llama-api and groq. The best result on every target language is highlighted in **bold** font, and the second best result is marked with an underline.

clude high-performing open-source models like Llama3.1-70B and Gemma2-27B.

When comparing Table 2 and Table 4, it becomes evident that many open-source large language models, under our SITR method, significantly outperform GPT-4 using two-shot learning. Table 5 further reveals that while GPT-4o consistently leads in most metrics, the other models achieve second-best performances across various languages and metrics. Overall, many LLMs could deliver impressive results on average, which also demonstrates the robustness of our SITR method.

Parameter via Capability. From Table 4 and Table 5, we observe that a large language model's cross-lingual capabilities on low-resource languages are not solely dependent on the number of model parameters. For example, within the Qwen-1.5 series, Qwen-1.5-72B outperforms Qwen-1.5-110B in several low-resource languages, such as *Gujarati* and *Ukrainian*. Additionally, the Gemma2-9B and Gemma2-27B models demonstrate strong performance, with Gemma2-27B



Google told me that around 70% of the Android users it spoke to wanted to "find a better balance" in their tech use. But this puts the tech sector in an awkward spot. How can it help its customers detach from their phones when, for many firms, their business model relies on them doing the opposite? Both Apple and Google have this year released digital tools - called Screen Time and Digital Wellbeing respectively - which let iPhone and Android users ...

English Summary: Tech companies like Google and Apple have introduced tools (Digital Wellbeing and Screen Time) to help users track and manage phone usage. While the data may cause some users to feel guilty, the tools are designed to be non-judgmental. Despite these efforts, reducing phone dependence is challenging, as companies like Google rely on user engagement. Some suggest using smaller phones to limit screen time, but personal willpower remains key.

Translated Summary: Технологічні компанії, такі як Google та Apple, запровадили інструменти (Digital Wellbeing та Screen Time), щоб допомогти користувачам відстежувати та керувати використанням телефону. Хоча ці дані можуть викликати у деяких користувачів почуття провини, інструменти створені для того, щоб не засуджувати. Незважаючи на ці зусилля, скорочення залежності від телефонів залишається складним завданням, оскільки компанії, такі як Google, залежать від ...

Model Generation: Google повідомив мені, що близько 70% користувачів Android, з якими він спілкувався, хотіли "знайти кращий баланс" у використанні технологій. Але це ставить технологічний сектор в незручне положення. Як він може допомогти своїм клієнтам відірватися від своїх телефонів, коли, для багатьох фірм, їх бізнес-модель ґрунтується на протилежному? Обидва Apple і Google цього року випустили цифрові ...

(Google informed me that about 70% of Android users it communicated with wanted to "find a better balance" in their use of technology. But this puts the tech sector in an awkward position. How can it help its customers detach from their phones when, for many firms, their business model is based on the opposite? Both Apple and Google released digital tools this year ...)

[SUMMARIZATION]: Google and Apple have introduced tools to help users monitor and manage their screen time, acknowledging the need for a better balance in tech use. Despite the tech industry's reliance on screen time for revenue, there is a growing awareness of the need to reduce phone dependency, although breaking up with phones entirely may be a challenging task.

[IMPROVEMENT]: Google and Apple have introduced tools to help users monitor and manage their screen time, acknowledging the need for a better balance in tech use. Despite the tech industry's reliance on screen time for revenue, there is a growing awareness of the need to reduce phone dependency.

[TRANSLATION]: Google і Apple впровадили інструменти, щоб допомогти користувачам стежити та керувати своїм часом на екрані, визнаючи потребу у кращому балансі використання технологій. Незважаючи на залежність технічної індустрії від екранного часу для отримання доходу, існує зростаюча усвідомленість необхідності зменшити залежність від телефонів.

[REFINEMENT]: Google та Apple впровадили інструменти для допомоги користувачам контролювати та керувати своїм часом на екрані, визнаючи потребу у кращому балансі використання технологій. Незважаючи на те, що технологічна індустрія покладається на час на екрані для отримання прибутку, зростає усвідомлення необхідності зменшення залежності від телефону.

(Google and Apple have implemented tools to help users monitor and manage their screen time, recognizing the need for a better balance in technology use. **Despite the tech industry's reliance on screen time for profit, there is growing awareness of the need to reduce dependency on phones.**)

Все більше користувачів смартфонів переймаються кількістю часу, який вони витрачають на свій гаджет.
(More and more smartphone users are becoming concerned about the amount of time they spend on their devices.)

Figure 3: Comparison of three different LLM methods on one single test example to summarize English source text in Ukarainian. The English translation of each model output is shown in brackets.

achieving the best result in *Indonesian* and the second-best result in *Gujarati*, only slightly behind the GPT-4O-MINI model on average.

Best LLM Under Our Method. We evaluate the performance of various LLMs, with the results shown in Table 4 and Table 5. Overall, the GPT series models consistently deliver better performance across different low-resource languages, particularly GPT-4O, which is considered one of the most powerful LLMs available. While other open-source models do not surpass the overall performance of GPT-4O, some are able to match or even exceed its performance in specific languages.

3.5 Output Analysis

In Figure 3, we compare our SITR method with other two LLM baselines in summarizing English news into Ukrainian.

The outputs from the other two methods are suboptimal due to their lack of relevance to the main topic and the generation of nonsensical content. The single-step summarize-translate method, which lacks self-correction and crucial prompt guidance, translates inaccurate summaries directly into the target language, causing error accumulation. On the other hand, the two-shot generation method skips the distributed thinking process, leading to uncontrollable outputs when the model fails to learn effectively from the examples. Both approaches, therefore, exhibit significant limitations.

In contrast, our method leverages meta-generation with targeted guidance, ensuring the model produces controlled and coherent outputs. This approach also allows the model to engage in self-reflection and iterative improvement, leading to more reliable and accurate results. The improvement step streamlines the summary by removing unnecessary sentences, while the refinement step adjusts sentence structure to better match the style of news reporting. Compared to the dataset's reference, our method captures the essence of the source text even more effectively.

3.6 Ablation Studies

Our proposed method improves upon the traditional single-step summarize-translate approach by integrating tailored prompts and employing a two-stage meta-generation process, which involves enhancing the summary and refining the translation.

The two additional steps, **IMPROVEMENT** and **REFINEMENT**, utilize meta-generation to optimize output and minimize error accumulation. These distinctions are particularly critical for cross-lingual summarization tasks in low-resource languages. Thus, we pose the question: *How significantly do meta-generation steps impact the overall performance of LLMs on this task?*

Here, we carry out three sets of comparative experiments to demonstrate the importance of two meta-generation steps: (i) Delete the **IMPROVEMENT** step. (ii) Delete the **REFINEMENT** step.

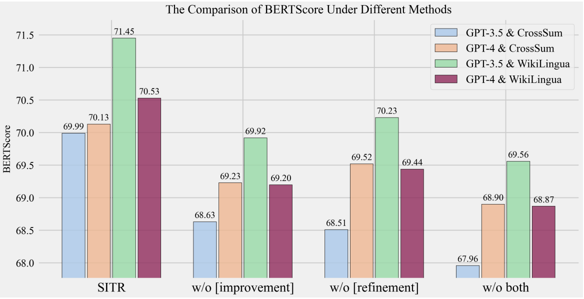


Figure 4: Comparison of the BERTScore after removing key meta-generation steps.

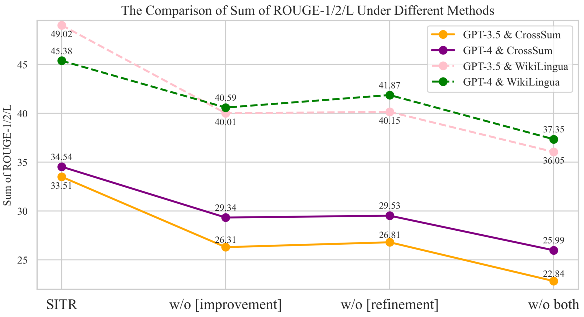


Figure 5: Comparison of the sum of ROUGE-1/2/L after removing key meta-generation steps.

(iii) Delete both the **IMPROVEMENT** and **REFINEMENT** step.

The ablation experimental results across different metrics are presented in Figure 4 and Figure 5, where we compare the performance of the complete SITR architecture with the three ablation experiments.

From these results, we could find that each step positively impacts the final outcomes. Deleting any step results in decreased ROUGE-1/2/L and BERTScore metrics, underscoring the overall significance of our proposed SITR architecture. Specifically, deleting the **IMPROVEMENT** step results in an approximate 18.4% drop in the sum of ROUGE-1/2/L scores (from 49.02 to 40.01) for GPT-3.5. Also, the BERTScore would decrease by 1-2 percentage points, which indicates a noticeable loss in semantic quality.

In summary, our proposed SITR method illustrates the cooperative and complementary nature of its architecture. This demonstrates the robustness of our SITR method: when model ouputs are less than ideal, the **IMPROVEMENT** and **REFINEMENT** steps allow the model to self-correct and reassess, mitigating the impact of error accumulation on the final output. (Further studies in Appendix D).

4 Related Works

Cross-lingual summarization is a critical task in natural language processing, involving the generation of a summary for text in one language based on a source text in another language (Wang et al., 2022; Zheng et al., 2022). The emergence of deep learning-based neural machine translation systems (Bahdanau et al., 2016; Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014) and text summarization approaches (Shini and Kumar, 2021; Nallapati et al., 2016; Shi et al., 2018), particularly those leveraging recurrent neural networks (Schuster and Paliwal, 1997; Chung et al., 2014; Hochreiter and Schmidhuber, 1997), enhanced model performance on CLS tasks.

Later, advances in neural network technologies, especially the Transformer architecture (Vaswani et al., 2017), have led to the development of end-to-end CLS models that integrate translation and summarization into a single framework, improving overall performance. Recent years, large language models have experienced a period of rapid development and widespread adoption (Ouyang et al., 2022; Brown et al., 2020; Touvron et al., 2023), and they have gained attention for their potential in cross-lingual summarization. Wang et al. (2023) showed their strong capabilities in high-resource languages like Chinese and German, while Park et al. (2024) found that LLMs using few-shot approaches still struggle with low-resource languages. This investigation is crucial for understanding and improving the models' ability to produce accurate and coherent summaries across various languages, thereby expanding the scope and applicability of LLMs in the CLS domain.

5 Conclusion and Future Work

In this paper, we introduce a four-step zero-shot SITR architecture, demonstrating the potential of LLMs for cross-lingual summarization in low-resource languages. Our approach enables LLMs to outperform three baseline types across various metrics, achieving notable performance in this domain.

We apply our SITR method to evaluate a wide range of LLMs, revealing their strong performance in cross-lingual summarization for low-resource languages and further demonstrating the robustness of our approach. For future research, we plan to investigate more effective methodologies to further unlock the potential of LLMs in this domain.

Limitations

While we evaluate the performance of LLMs in cross-lingual summarization on two datasets to showcase the effectiveness of both our zero-shot SITR method and the models, this study has several limitations: (i) The design of prompts can affect model performance, partly due to the models' limited confidence with low-resource languages. Future research could explore methods to enable large language models to generate reliable outputs without depending on manually designed prompts. (ii) We do not examine cross-lingual summarization tasks involving two low-resource languages. Future work could address this gap to fully explore the potential of LLMs in these more challenging scenarios.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, et al. 2023. Qwen technical report.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, et al. 2024. The llama 3 herd of models.

Tahmid Hasan, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yuan-Fang Li, Yong bin Kang, and Rifat Shahriyar. 2021. Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs. *CoRR*, abs/2112.08804.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003. Cross-lingual c*st*rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing*, 2(3):245–269.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report.

C Orăsan and OA Chiorean. 2008. Evaluation of a cross-lingual romanian-english multi-document summariser.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Gyutae Park, Seojin Hwang, and Hwanhee Lee. 2024. Low-resource cross-lingual summarization through few-shot learning with large language models.

Jacob Parnell, Inigo Jauregi Unanue, and Massimo Piccardi. 2024. SumTra: A differentiable pipeline for few-shot cross-lingual summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2399–2415, Mexico City, Mexico. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2018. Neural abstractive text summarization with sequence-to-sequence models. *CoRR*, abs/1812.02303.

R. Subha Shini and V.D. Ambeth Kumar. 2021. Recurrent neural network based text summarization techniques by word sequence generation. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 1224–1229.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, et al. 2024. Gemma: Open models based on gemini research and technology.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 12–23, Singapore. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. A Survey on Cross-Lingual Summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. 2024. From decoding to meta-generation: Inference-time algorithms for large language models.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Shaohui Zheng, Zhixu Li, Jiaan Wang, Jianfeng Qu, An Liu, Lei Zhao, and Zhigang Chen. 2022. Long-document cross-lingual summarization.

A Large Language Models

Here, we give the introduction of LLMs used in this paper.

- GPT-3.5: A powerful language model developed by OpenAI, designed to generate human-like text based on input prompts, which is highly effective for a variety of natural language processing (NLP) tasks.

- GPT-4: An advanced multi-modal LLM from OpenAI, which is able to accept both images and texts to do text-generation, and exhibits human-level performance on various NLP benchmarks.

- GPT-4o: A better multi-modal LLM compared with GPT-4 which was released on May 13, 2024 by OpenAI.

- GPT-4o-mini: A most cost-efficient multi-modal small model released on July 18, 2024 by OpenAI, which enables a broad range of tasks with low cost and latency.

- Llama3: A family of large language models includes two versions with 8B and 70B parameters developed by MetaAI, which were trained on 15 trillion tokens data.

- Llama3.1: The latest and strongest open-source LLM family released by MetaAI, contains three models with 8B, 70B and 405B parameters.

- Qwen-1.5: The improved version of Qwen, the LLM family developed by Alibaba Cloud. The whole family contains eight models: 0.5B, 1.8B, 4B, 7B, 14B, 32B, 72B, and 110B.

- Qwen2: Newest model series from the Qwen team with better performances.

- Gemma: Light-weight, text-to-text, decoder-only large language models trained by Google, which have two versions with 2B and 7B parameters.

- Gemma2: The next generation of open-source models and improved version of Gemma, released by Google on June 27, 2024, which contains three versions with 2B, 9B and 27B parameters.

- Mixtral: Mixture of Experts (MoE) models (Shazeer et al., 2017) with 8 experts trained by Mistral AI, now have two versions 8x7B and 8x22B.

B Experiments Details

We primarily use GPT-3.5, GPT-4, GPT-4o, and GPT-4o-mini models via OpenAI's official API[4]. Additionally, we utilize llama-api[5] to access two models from the Qwen-1.5 family, Qwen-2-72B, Mixtral-8x22B, and Gemma2-27B. For the remaining models, we conduct experiments using the groq platform[6].

For all LLM-related experiments, we set the *temperature* to 0.0 and *top-p* to 0.95 to minimize randomness and ensure consistent model outputs. To reproduce the few-shot results from the previous paper with GPT-3.5 and GPT-4, we use the provided prompt (Park et al., 2024) and access the OpenAI official API.

For baselines requiring further fine-tuning, we use three encoder-decoder transformer models: mBART-50[7], mT5-base[8], and mT5-small[9]. If a low-resource language's training data exceeds 1,000 samples, we randomly select 1,000 for the 1,000-shot experiments. If the data contains fewer than 1,000 samples, we use all available data for fine-tuning. To fine-tune mBART-50, mT5-base, and mT5-small, we perform all experiments on a single 24GB-VRAM A5000 GPU. We set the training epochs to 3, with learning rates of 1e-4, 3e-4, and 5e-4, respectively, and select the checkpoint with the highest sum of ROUGE-1/2/L scores.

At last, for OpenAI's GPT-3.5 and GPT-4 models, we use the *gpt-3.5-turbo-0125* and *gpt-4-0125-preview* versions, conducting all experiments between July 15th and August 1st.

C Additional Experiments

To better illustrate the effectiveness of our proposed SITR method in leveraging large language models for cross-lingual summarization in low-resource languages, we select eight additional low-resource languages from the CrossSum dataset (Igbo, Hausa, Nepali, Yoruba, Oromo, Welsh, Urdu, Swahili) and

[4] https://openai.com/

[5] https://www.llama-api.com/

[6] https://groq.com/

[7] https://huggingface.co/facebook/mbart-large-50

[8] https://huggingface.co/google/mt5-base

[9] https://huggingface.co/google/mt5-small

Model		Language Pair																Average Score				
		English⇒Igbo ♠				English⇒Hausa ♠				English⇒Nepali				English⇒Yoruba ♠								
		R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	S-R	BS
mBART-50	0-shot	3.38	0.00	2.46	59.65	3.20	0.31	2.93	55.76	0.00	0.00	0.00	60.07	2.86	0.05	1.82	57.34	2.36	0.09	1.80	4.25	58.21
	1000-shot	6.59	0.00	5.50	60.89	14.24	2.94	11.57	67.34	9.40	1.91	8.18	68.18	7.68	0.00	7.43	61.10	9.48	1.21	8.17	18.86	64.38
mT5-base	1000-shot	0.43	0.00	0.43	51.76	6.90	0.17	5.61	55.82	0.72	0.00	0.72	56.35	0.55	0.00	0.55	51.28	2.15	0.04	1.83	4.02	53.80
GPT-3.5 (Park et al., 2024)	zero-shot	3.44	0.35	2.61	58.23	4.95	0.67	4.12	57.71	4.30	0.56	3.27	63.87	2.34	0.32	1.95	58.48	3.76	0.48	2.99	7.23	59.57
	one-shot	4.35	0.59	3.27	58.45	7.82	0.88	6.21	61.78	5.26	0.40	4.74	66.24	2.87	0.15	2.01	58.72	5.08	0.51	4.06	9.65	61.30
	two-shot	6.85	1.46	5.22	60.13	10.12	2.05	7.16	62.59	5.56	0.70	4.46	65.52	4.07	1.10	2.81	58.27	6.65	1.33	4.91	12.89	61.63
GPT-4 (Park et al., 2024)	zero-shot	7.64	1.59	5.70	65.34	11.67	3.50	7.65	65.10	5.73	0.76	4.24	65.23	6.33	0.96	4.86	65.77	7.84	1.70	5.61	15.15	65.36
	one-shot	9.24	2.09	6.66	65.90	12.97	3.71	8.61	65.33	7.32	1.15	5.30	66.09	6.59	1.38	5.29	65.80	9.03	2.08	6.47	17.58	65.78
	two-shot	9.58	1.99	6.94	65.64	12.84	3.54	8.70	65.49	7.21	1.66	5.57	66.35	6.76	1.00	5.21	65.76	9.10	2.05	6.61	17.76	65.81
GPT-3.5 w/ summarize-translate	zero-shot	3.67	0.21	2.84	57.69	8.36	1.29	6.07	61.91	5.82	0.71	4.64	66.25	4.64	0.90	3.83	59.84	5.62	0.78	4.35	10.75	61.42
GPT-4 w/ summarize-translate	zero-shot	6.61	1.22	5.30	65.44	12.44	3.08	8.52	65.48	5.29	0.91	3.95	65.51	4.93	0.72	3.92	65.13	7.32	1.48	5.42	14.22	65.39
GPT-3.5 w/ SITR (Ours)	zero-shot	15.56	2.79	10.63	67.31	19.60	2.15	14.84	69.33	12.89	2.51	9.63	69.69	11.82	2.20	10.69	68.41	14.97	2.41	11.45	28.33	68.69
GPT-4 w/ SITR (Ours)	zero-shot	**20.47**	**4.02**	**14.90**	**70.17**	**23.02**	**4.30**	**14.22**	**70.04**	**15.81**	**3.17**	**11.09**	**70.19**	**15.39**	**2.15**	**9.81**	**70.90**	**18.67**	**3.41**	**12.51**	**34.59**	**70.33**

Table 6: Additional experimental results on the CrossSum dataset. R-1, R-2, R-L, S-R and BS refer to ROUGE-1, ROUGE-2, ROUGE-L, sum of ROUGE-1/2/L and BERTScore respectively. The task with ♠ means training data less than 1000, where 1000-shot setting equals full fine-tuning. The best result on every target language is highlighted in **bold** font, and the second best result is marked with an underline.

sample 30 test examples for each language. We evaluate our method against fine-tuned models and other LLM baselines, with the results presented in Table 6 and Table 7.

We observe that GPT-3.5 and GPT-4 still significantly and consistently outperform all other baselines when using our proposed SITR method.

The results from Table 6 and Table 7 show that, despite fine-tuning, the performance of the mT5-base model remains unsatisfactory, while the mBART-50 model performs significantly better. However, when comparing the scores of fine-tuned mBART-50 with our proposed zero-shot SITR method, it is clear that SITR still holds a significant advantage, showing notable improvements in both BERTScore and all three ROUGE metrics.

In Table 6, the sum of ROUGE-1/2/L scores for GPT-3.5 improves by 124% (from 12.89 to 28.83) and 168% (from 10.75 to 28.83) when comparing our SITR method with two-shot generation and the summarize-translate method. For GPT-4, the improvements are 95% (from 17.76 to 34.59) and 143% (from 14.22 to 34.59) respectively.

In Table 7, the sum of ROUGE-1/2/L scores for GPT-3.5 increases by 58% (from 25.59 to 40.35) and 84% (from 21.90 to 40.35) when comparing our SITR method with the best few-shot generation method (one-shot generation worked best for GPT-3.5) and the summarize-translate method. For GPT-4, the improvements are 51% (from 26.89 to 40.64) and 112% (from 19.20 to 40.64) respectively.

Additionally, our method shows significant improvement in the BERTScore metric, indicating a substantial semantic advantage for SITR outputs. Moreover, we find that the improvement of few-shot learning methods is significantly constrained when transitioning from one-shot to two-shot. In fact, Table 7 shows a slight performance decline with GPT-3.5, suggesting that few-shot learning may face certain limitations on this task.

These results, based on experiments with eight low-resource languages, further demonstrate that our SITR method effectively harnesses the capabilities of large language models in this domain.

D Prompt vs Meta-Generation

To further explore the impact of effective prompt guidance, we conduct three sets of comparative experiments for comprehensive ablation studies:

(i) Replace the designed summarization prompt with a simple summarization prompt (summarize the following text ...).

(ii) Replace the designed translation prompt with a simple translation prompt (translate the following text into ...).

(iii) Replace both summarization and translation prompts with their respective simple versions.

The ablation results for prompt replacements are shown in Table 8, comparing the performance of

Model		Language Pair																Average Score				
		English⇒Oromo ♠				English⇒Welsh				English⇒Urdu				English⇒Swahili								
		R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	S-R	BS
mBART-50	0-shot	1.90	0.24	1.22	55.16	1.77	0.08	1.77	60.84	0.09	0.00	0.09	60.98	1.65	0.20	1.65	59.82	1.35	0.13	1.18	2.66	59.20
	1000-shot	5.66	0.56	4.85	56.48	18.38	3.60	13.90	68.68	18.09	3.25	14.18	68.47	16.89	4.80	13.41	69.55	14.76	3.05	11.59	29.40	65.80
mT5-base	1000-shot	0.00	0.00	0.00	48.11	6.86	0.60	6.86	60.19	5.25	0.00	4.89	59.90	0.01	0.00	0.01	47.39	3.03	0.15	2.94	6.12	53.90
GPT-3.5 (Park et al., 2024)	zero-shot	3.56	0.92	3.33	55.12	9.78	2.07	6.62	65.87	14.36	3.39	9.36	66.21	14.66	4.23	9.97	67.56	10.59	2.65	7.32	20.56	63.69
	one-shot	4.10	0.36	3.52	57.91	14.23	4.21	10.54	67.13	13.84	2.25	9.92	65.01	19.77	5.49	14.11	69.38	12.99	3.08	9.52	25.59	64.86
	two-shot	3.02	0.89	2.97	54.22	15.19	4.18	10.87	67.09	14.91	2.79	10.41	66.12	18.06	4.97	12.79	68.87	12.80	3.21	9.26	25.27	64.08
GPT-4 (Park et al., 2024)	zero-shot	5.94	1.14	4.12	61.69	9.58	2.20	6.59	65.86	13.24	3.18	8.44	66.04	9.44	2.65	6.81	66.01	9.55	2.29	6.49	18.33	64.90
	one-shot	7.26	1.31	4.96	62.20	15.35	4.35	10.51	67.50	15.49	3.84	10.07	66.92	15.62	4.86	11.27	68.08	13.43	3.59	9.20	26.22	66.18
	two-shot	6.89	1.04	4.59	62.38	15.00	4.61	10.11	67.60	16.19	3.76	11.02	67.32	17.33	5.06	11.97	68.58	13.85	3.62	9.42	26.89	66.47
GPT-3.5 w/ summarize-translate	zero-shot	4.20	1.14	3.13	57.31	6.05	0.84	4.52	63.02	18.60	3.38	12.91	68.12	16.60	4.72	11.52	67.77	11.36	2.52	8.02	21.90	64.06
GPT-4 w/ summarize-translate	zero-shot	4.52	0.86	3.43	61.37	10.77	2.51	7.05	66.04	12.50	2.83	8.55	66.42	11.71	3.33	8.73	66.78	9.88	2.38	6.94	19.20	65.15
GPT-3.5 w/ SITR (Ours)	zero-shot	13.25	**3.92**	**10.35**	66.69	**26.11**	**4.96**	**16.83**	**70.42**	22.51	3.94	14.49	69.20	23.21	5.84	15.95	70.03	21.27	**4.67**	14.41	40.35	69.09
GPT-4 w/ SITR (Ours)	zero-shot	**15.74**	3.33	**11.00**	**67.50**	23.24	4.40	15.17	69.41	**23.09**	**4.53**	**15.47**	**69.29**	**23.91**	**6.01**	**16.63**	**70.39**	**21.50**	4.57	**14.57**	**40.64**	**69.15**

Table 7: Additional experimental results on the CrossSum dataset. R-1, R-2, R-L, S-R and BS refer to ROUGE-1, ROUGE-2, ROUGE-L, sum of ROUGE-1/2/L and BERTScore respectively. The task with ♠ means training data less than 1000, where 1000-shot setting equals full fine-tuning. The best result on every target language is highlighted in **bold** font, and the second best result is marked with an underline.

Model	Method	Dataset			
		CrossSum		WikiLingua	
		SUM-ROUGE	BERTScore	SUM-ROUGE	BERTScore
GPT-3.5	SITR	**32.15**	**69.58**	**49.02**	**71.45**
	W/O [SUMMARIZATION PROMPT]	30.17	69.01	46.71	70.52
	W/O [TRANSLATION PROMPT]	29.91	68.83	45.10	70.21
	W/O BOTH	27.45	68.76	43.29	70.03
GPT-4	SITR	**34.54**	**70.13**	**45.38**	**70.53**
	W/O [SUMMARIZATION PROMPT]	31.23	69.71	41.83	69.92
	W/O [TRANSLATION PROMPT]	32.09	69.59	42.37	69.88
	W/O BOTH	29.85	69.26	39.13	69.36

Table 8: Experimental results on two datasets for prompt replacement. SUM-ROUGE refers to the sum of ROUGE-1/2/L scores. The best result is highlighted in **bold** font.

the original SITR method with the three prompt variations. From these results, we can draw the following conclusions:

(1) Using appropriate prompts for both the **SUMMARIZATION** and **TRANSLATION** steps positively impacts the large language model's performance in cross-lingual summarization tasks for low-resource languages.

(2) The summarization prompt more significantly affects the quality of the final output than the translation prompt, given its role as the initial step in the architecture.

(3) Considering the results in Figure 4 and Figure 5, we could find that the decrease in performance from prompt replacement is much smaller than the decrease from removing key meta-generation steps.

In summary, manually designed prompts, along with the **IMPROVEMENT** and **REFINEMENT** steps, significantly enhance the performance of large language models in cross-lingual summarization tasks for low-resource languages. However, the two meta-generation steps contribute more significantly to improving model capabilities than the prompts themselves.

Prompt

Text Generation Prompt for Summarization Step:

You are tasked with creating a concise summary of a given text. The text to be summarized is provided below:

{{TEXT_TO_SUMMARIZE}}

To create an effective summary, follow these guidelines:

1. Read the entire text carefully to understand the main ideas and overall message.
2. Identify the key points, main arguments, or central themes of the text.
3. Focus on the most important information and avoid including minor details or examples.
4. Aim to capture the essence of the text in a concise manner.
5. The summary should be significantly shorter than the original text, ideally about 5-10% of its length.
6. Ensure that the summary flows logically and maintains coherence.
7. Do not include your own opinions or interpretations; stick to the information presented in the original text.

Write your summary within <summary> tags. The summary should be brief and to the point, covering only the main content without delving into excessive details. Aim for as few sentences as possible.

Figure 6: Text Generation Prompt for Summarization Step (Our SITR).

Prompt

Text Generation Prompt for Improvement Step:

You are tasked with evaluating and improving a summary of a given text. Your goal is to create a brief, concise summary that captures the main points without unnecessary details. Follow these steps:

1. First, read the original text:

{{SOURCE_TEXT}}

2. Now, read the current summary:

{{SUMMARY}}

3. Evaluate the current summary based on the following criteria:
   a. Accuracy: Does it correctly represent the main ideas of the original text?
   b. Conciseness: Is it brief and to the point?
   c. Clarity: Is it easy to understand?
4. Improve the summary by:
   a. Removing any unnecessary details or redundant information
   b. Limiting the length to an equal number or fewer sentences
5. Provide your improved summary within <improved_summary> tags.

Remember, the goal is to create a brief and accurate summary that captures the essence of the original text without going into details.

Figure 7: Text Generation Prompt for Improvement Step (Our SITR).

Prompt

Text Generation Prompt for Translation Step:
You are a highly skilled translator with expertise in various languages, including less commonly used ones. Your task is to translate an English text into a specified target language. Please follow these instructions carefully:
1. You will be provided with an English text to translate. The text is as follows:
{{SOURCE_TEXT}}
2. The target language for translation is:
{{TARGET_LANGUAGE}}
3. When translating, please consider the following:
   - Pay attention to cultural nuances and idiomatic expressions
   - Maintain the original tone and style of the text as much as possible
   - Ensure grammatical accuracy in the target language
   - If there are any terms or concepts that don't have a direct equivalent in the target language, provide the best possible translation and include a brief explanation in parentheses
4. Your output should adhere to these guidelines:
   - Do not repeat words or sentences unnecessarily
   - Avoid any gibberish or nonsensical text
   - Provide a fluent and coherent translation
   - If you're unsure about a particular word or phrase, provide your best translation and indicate your uncertainty with [?] after the word or phrase
5. Please provide your translation within <translation> tags. If you need to include any translator's notes or explanations, please add them after the translation within <notes> tags.
Now, please translate the given English text into the specified target language.

Figure 8: Text Generation Prompt for Translation Step (Our SITR).

Prompt

Text Generation Prompt for Refinement Step:
You are a bilingual expert in English and {{TARGET_LANGUAGE}}. Your task is to analyze and refine a translation from English to {{TARGET_LANGUAGE}}, focusing on fixing any duplicate content and gibberish. Follow these steps:
1. First, carefully read the original English text:
{{ENGLISH_TEXT}}
2. Now, examine the translation in {{TARGET_LANGUAGE}}:
{{TRANSLATED_TEXT}}
3. Analyze the translation for the following issues:
   a. Overall accuracy: Check if the translation accurately conveys the meaning of the original English text.
   b. Gibberish: Look for any parts of the translation that don't make sense or seem like nonsensical text.
4. Refine the translation by:
   a. Making minor adjustments to improve accuracy and fluency, while preserving the original style and tone
   b. Replacing gibberish with appropriate {{TARGET_LANGUAGE}} text that matches the meaning of the English original
5. Provide your refined translation inside <refined_translation> tags.

Figure 9: Text Generation Prompt for Refinement Step (Our SITR).

Prompt

Text Generation Prompt for Summarize-Translate Method:
Please first summarize the following text and then translate the summary into {{TARGET_LANGUAGE}}:
{{TEXT_TO_SUMMARIZE}}.
Return the final translated {{TARGET_LANGUAGE}} summary within <translated_summary> tags.

Figure 10: Text Generation Prompt for Single-Step Summarize-Translate Method.

Prompt

Text Generation Prompt for Few-Shot Method:
Please summarize the following text in {{TARGET_LANGUAGE}}.
Example 1
Text: {{EXAMPLE1_TEXT}}
Translated summary: {{EXAMPLE1_SUMMARY}}
Example 2
Text: {{EXAMPLE2_TEXT}}
Translated summary: {{EXAMPLE2_SUMMARY}}
Test Text
Text: {{TEST_TEXT}}
Translated summary:

Figure 11: Text Generation Prompt for Few-Shot Method.

Source Text

Her party and its allies won almost all of the 300 parliamentary seats contested, in its best ever result. But the opposition alliance, which won just seven seats, condemned the vote as "farcical", marred by violence, intimidation and vote rigging claims. Election authorities said there could be no re-run. Although the election commission earlier said it had heard vote-rigging allegations from "across the country" and would investigate, the commissioner said voting had been held in a peaceful manner and hence there would be no new vote. Bangladesh's parliament has 350 seats in total, 50 of which are reserved for women and allotted proportional to the overall vote. "We urge the election commission to void this farcical result immediately," opposition leader Kamal Hossain said on Sunday. "We are demanding that a fresh election is held under a neutral government as early as possible." At least 17 people have been killed in clashes between ruling party supporters and the opposition. What are the allegations? Sheikh Hasina's Awami League has run Bangladesh since 2009 and won 288 seats in this election. But one of the leading opposition parties has accused it of using stuffed ballot boxes. A spokesman ......

Reference

સોમવારે બાંગ્લાદેશના ચૂંટણી પંચે જણાવ્યું કે સતત ત્રીજી વખત શેખ હસિના વડાં પ્રધાન પદ પર જીત્યાં છે.
(On Monday, Bangladesh's Election Commission announced that Sheikh Hasina has won the Prime Minister's position for the third consecutive time.)

Summarize-Translate

તેની પાર્ટી અને તેના સહયોગીઓ ને તેની સૌથી સફળ પરિણામ માટે પ્રારંભમાં તેની સૌથી વધું પરિણામ મળ્યું હતું. પરંતુ વિરોધી ગઠબંધન, જેમાં માત્ર સાત સીટો જીતી, મતદાનને "વિચિત્ર" ઘોષિત કર્યો હતો, જેને હિંસા, ડરાવટ અને મતદાન રિગિંગ દાવોની રીતે નકાર્યું. ચૂંટણી અધિકારીઓ ને કોઇ પુનઃચાલન થવું સંભવ નથી. ચૂંટણી આયોગે પહેલાં કહ્યું હતું કે તે મતદાન રિગિંગ દાવોને "પૂરી દેશભરમાં" સાંભળ્યો હતો અને તે તપાસવા માટે કહ્યું હતું, પરંતુ આયોગને કહ્યું કે મતદાન શાંતિપૂર્વક થયું હતું અને તેથી નવો મતદાન થવો નથી. બાંગ્લાદેશની સભામંડળમાં 350 સીટો છે, જેમાંથી 50 સીટો મહિલાઓ માટે આપવામાં આવે છે અને તેની સમગ્ર મતને પ્રમાણાત્મક રીતે વહેંચવામાં આવે છે. "અમે ચાહીએ કે ચૂંટણી આયોગ આ વિચિત્ર પરિણામને તત્કાલ નિરાકરે," વિરોધી નેતા કમલ હોસેને રવિવારે કહ્યું. "અમે આવડી રીતે નિયમિત સરકાર હેઠળ તાજેતર ચૂંટણી યોગ્ય છે તેની માગ કરી રહ્યા છે." શાસક પાર્ટીના સમર્થકો અને વિરોધી વચ્ચે 17 લોકોની હત્યા થઈ છે. આ ચૂંટણી મહત્વપૂર્ણ કેમ હતી? ......

(His party and its allies initially achieved the highest outcome, securing the most successful result. However, the opposition coalition, which won only seven seats, declared the election 'strange,' rejecting it due to claims of violence, intimidation, and vote-rigging. Election officials stated that a rerun is unlikely. The Election Commission had earlier mentioned that it had heard allegations of vote-rigging 'across the country' and called for an investigation. However, the commission stated that the voting was peaceful and therefore a new election is not necessary. The Bangladeshi Parliament has 350 seats, of which 50 are reserved for women, distributed proportionally based on the overall vote. 'We want the Election Commission to immediately address this strange outcome,' opposition leader Kamal Hossain said on Sunday. 'We are demanding that the recent election under this regular government is not fair.' Seventeen people have been killed in clashes between supporters of the ruling party and the opposition. Why was this election significant? ......)

👎

Two-Shot Generation

બાંગ્લાદેશની ચૂંટણીમાં શેખ હસીનાની પાર્ટી અને તેના મિત્રોએ તેની સૌથી વધુ 300 સભાસદીય સીટોમાંથી પ્રાપ્ત કરી છે, જે તેની સૌથી ઉત્તમ પરિણામ છે. પરંતુ વિરોધી ગઠબંધન, જેમણે માત્ર સાત સીટો જીતી, આ મતદાનને "વિચારવહિત" ઘોષિત કરી છે, જેમણે હિંસા, ડરાવટ અને મતદાન રિગિંગની દાવોને લઈને મતદાનને "વિચારવહિત" ઘોષિત કરી છે.

(In the Bangladesh election, Sheikh Hasina's party and its allies secured the highest number of seats, winning the most out of the 300 parliamentary seats, marking their best result ever. However, the opposition coalition, which won only seven seats, has declared the election "questionable," citing claims of violence, intimidation, and vote-rigging.)

😐

SITR (Two-Stage Meta-Generation)

[SUMMARIZATION]: In the election, Sheikh Hasina' party and its allies won almost all of the 300 parliamentary seats, marking their best result ever. However, the opposition alliance won only 7 seats and condemned the election as "farcical," claiming it was marred by violence, intimidation, and allegations of vote rigging. At least 17 people were killed in clashes between supporters of the ruling party and the opposition.

[IMPROVEMENT]: Sheikh Hasina‘s Awami League and its allies secured a dominant victory in Bangladesh’s parliamentary election, while the opposition alliance criticized the vote as rigged and violent, calling for a fresh election under a neutral government.

[TRANSLATION]: શેખ હસિના ની અવામી લીગ અને તેના સાથીઓએ બાંગ્લાદેશની સંસદીય ચૂંટણીમાં એક વિશાળ વિજય મેળવ્યો, જ્યારે વિપક્ષી મોરચે મતદાનને ધાંધલભર્યું અને હિંસક ગણાવીને તટસ્થ સરકાર હેઠળ ફરી ચૂંટણીની માગણી કરી.

[REFINEMENT]: શેખ હસીનાની અવામી લીગ અને તેના સહયોગીઓ ને બાંગ્લાદેશના સંસદીય ચૂંટણીમાં એક પ્રમુખ વિજય મેળવી, જ્યારે વિપક્ષ ગઠબંધને ચૂંટણીને ફર્જી અને હિંસાત્મક તરીકે નિંદા કરી, નિયત્રિત સરકાર હેઠળ એક તાજું ચૂંટણી માટે આવાહન કર્યું.

(Sheikh Hasina's Awami League and its allies secured a decisive victory in Bangladesh's parliamentary elections, while the opposition coalition condemned the election as fraudulent and violent, calling for a fresh election under a neutral government.)

👍

Figure 12: Comparison of three different LLM methods on one single test example to summarize English source text in Gujarati. Our SITR method performs the best, the result is even better than the reference. The English translation of each model output is shown in brackets.