

Enhancing Factual Consistency in Cross-Lingual Dialogue Summarization via Self-Guidance Prompting

1st Zhongtian Bao

*Key Laboratory of DISec
College of Computer Science
Nankai University
Tianjin, China*

zhongtianbao@mail.nankai.edu.cn
ORCID: 0000-0001-6618-5325

2nd Wenjian Ding

*Key Laboratory of DISec
College of Computer Science
Nankai University
Tianjin, China*

wjding@mail.nankai.edu.cn
ORCID: 0009-0006-7498-5289

3rd Yao Zhang

*School of Statistics and
Data Science, LPMC,
KLMDASR & AAIS,
Nankai University
Tianjin, China*

yaozhang@nankai.edu.cn
ORCID: 0009-0003-3071-0513

4th Jun Wang

Ludong University

Shandong, China
junwang@mail.nankai.edu.cn
ORCID: 0000-0001-8932-6661

5th Zhe Sun

*Graduate School of Medicine
Faculty of Health Data Science
Juntendo University*

Japan
z.sun.kc@juntendo.ac.jp
ORCID: 0000-0002-6531-0769

6th Andrzej Cichocki

*Systems Research Institute
Polish Academy of Sciences and
UMK*

Poland
cichockiand@gmail.com
ORCID: 0000-0002-8364-7226

7th Zhenglu Yang

*Key Laboratory of DISec
College of Computer Science
Nankai University
Tianjin, China*

yangzl@nankai.edu.cn
ORCID: 0000-0001-9528-965X

Abstract—Recent breakthroughs in large language models have made it feasible to effectively summarize cross-lingual dialogue information, proving essential for the global communication context. However, existing methodologies encounter difficulties maintaining factual consistency across multiple dialogue exchanges and lack clear explanations of the summarization process. This paper presents a novel factually consistent prompting technology with large language models to address these challenges in cross-lingual dialogue summarization. First, we propose a factual replacement mechanism to enhance information analysis by incorporating noise information into summarization candidates. We adopt a self-guidance framework to enforce factual consistency, enhancing information flow tracking in cross-lingual hybrid dialogue scenarios with the assistance of GPT-based models. Furthermore, we introduce a view-aware chain-of-thought-driven architecture to improve the interpretability and transparency of the cross-lingual dialogue summarization process. Comprehensive experimental evaluations on cross-lingual summarization tasks, spanning English, French, Spanish, Russian, Chinese, and Arabic, and hybrid cross-lingual tasks substantiate that the proposed

model achieves superior performance relative to state-of-the-art baselines.

Index Terms—Cross-Lingual Dialogue Summarization, Factual Consistency, Large Language Model

I. INTRODUCTION

Cross-lingual summarization is an essential task in natural language processing that seeks to compress the content from a document context written in one language (e.g., Spanish) and generate a coherent summary in another language (e.g., French). Recently, research on multilingual summarization has expanded to more challenging scenarios, such as dialogues in which speakers interact using their preferred languages. We aim to define this task as Cross-Lingual Dialogue Summarization (CLDS). Accurately summarizing cross-lingual dialogues has grown increasingly crucial in our globalized society, where communication often transcends linguistic boundaries. CLDS is emerging as an essential tool for facilitating comprehension and collaboration in diverse contexts.

Corresponding author: Yao Zhang (yaozhang@nankai.edu.cn), Zhenglu Yang (yangzl@nankai.edu.cn)

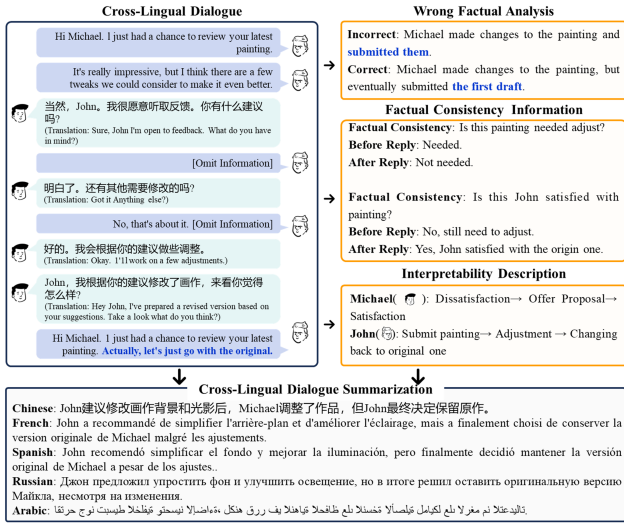


Fig. 1. An overview of the CLDS task. Different speakers communicate with each other using their native languages. Factual consistency and interpretability are crucial in generating accurate and appropriate summaries.

The demand for effective dialogue summarization has surged, especially in multinational conferences, where speakers interact using their preferred languages [1]–[6]. In such context, the specificity of feedback and the clarity of communication are essential. However, current research efforts struggle to effectively tackle the complexities that emerge when diverse participants utilize their native languages to interact, resulting in producing accurate and contextually relevant summaries. The specificity of CLDS is closely related to machine translation and mono-lingual summarization. Consequently, the CLDS domain has advanced through dedicated research initiatives that enhance frameworks and datasets adapted to cross-lingual contexts [7]–[12].

Despite the progress achieved, two fundamental challenges persist in CLDS: factual consistency and the requirement for explanations of summarization generation. In terms of factual consistency, it is a critical issue in summarization that has been extensively studied in various contexts, e.g., long document summarization [13], [14] and multilingual summarization [15]–[17]. As shown in Fig. 1, speakers’ perspectives evolve through ongoing interactions, and highlighting this dynamic nature of dialogue, primarily factual consistency, can significantly improve performance. However, it has received limited attention in multilingual dialogue summarization, essential for effectively addressing the unique characteristics of dialogue contexts, such as noise and fragmented information. To ensure factual accuracy, it is often vital to provide clear explanations during summarization. This paper presents a noise-tolerant self-reply module and a temporal information analysis module designed to enhance factual consistency and improve interpretability.

In addition to factual consistency, summarization interpretability also poses a challenge for current CLDS studies. Existing approaches often overlook the potential benefits of interpretability in the summarization process, which are crucial for validating source content and establishing the reliability of

the summaries.

To effectively capture and utilize complex cross-lingual information, we introduce a factual replacement mechanism that enhances information analysis by incorporating noise into CLDS candidates. Additionally, we present a self-guidance structure to ensure factual consistency, which improves the tracking of information flow in cross-lingual dialogue scenarios with the assistance of LLMs. We also introduce a view-aware CoT-based prompt model, which integrates diverse perspectives to facilitate self-guidance dialogue summarization.

Our main contributions are as follows:

(1) We introduce an integrated framework that follows the recent trend of using CoT approaches with LLMs. This framework incorporates a novel factual consistency-aware module and an interpretability module to more effectively address the challenges in CLDS.

(2) We propose a factual replacement mechanism alongside a self-reply analysis for factual consistency to capture the flow of dialogue effectively. This approach inputs noisy information and analyzes comments to enhance the understanding of relationships between different speakers.

(3) An extensive experimental evaluation validates our model’s exceptional performance across multiple benchmarks in the cross-lingual dialog summarization task. Moreover, our model surpasses existing methods in factual consistency and interpretability.

II. RELATED WORK

Cross-lingual dialogue summarization involves two core tasks, i.e., cross-lingual summarization and translation. Traditional approaches use pipeline architectures, i.e., either summarizing before translating (SumTran) or translating before summarizing (TranSum) [5], [18]. While interpretable, these methods suffer from error propagation between stages and lead to inferior performance.

To address this issue, recent work has shifted toward End-to-End (End2End) models that fine-tune LLMs in a sequence-to-sequence framework to directly generate summaries in the target language and can significantly improve results. Enhancements such as multi-task learning, specialized attention mechanisms, knowledge distillation, and variational autoencoders further boost performance [1].

However, full fine-tuning remains computationally expensive which requires extensive data and gradient-based optimization. Lightweight alternatives like prefix tuning that updates only a subset of parameters [19] and prompt-based methods have emerged. These approaches guide LLMs via carefully designed inputs to leverage pre-trained knowledge with minimal computation. Prompting and chain-of-thought strategies show strong promise in zero-shot and few-shot cross-lingual summarization [20], [21].

Fact consistency remains a major challenge in summarization and has been well-studied in long-document [13] and multilingual settings [15]. Yet it is largely overlooked in multilingual dialogue summarization. This gap is critical, as dialogues contain noise, interruptions, and fragmented

utterances that heighten the risk of factual inaccuracies in summaries.

III. THE PROPOSED FRAMEWORK

As illustrated in Fig. 2, we introduce an integrated CLDS framework that comprises three key components: the main stream of CoT with LLMs, a Factual Consistency Module, and an Interpretability Module. Given the superior reasoning capabilities of CoT approaches with LLMs, we incorporate a CoT prompting structure within the main framework to address the CLDS task. Specifically, the cross-lingual dialogue information coupled with contextual CoT prompts¹ is fed into the LLM to obtain the raw output. Subsequently, two optimization modules are incorporated into the process to refine the results, as explained below.

In the Factual Consistency Module (portrayed on the top of Fig. 2), we deliberately incorporate noise into the generated candidates, followed by a self-reply analysis process to evaluate factual consistency. The detail of the factual consistency module will be presented in Section III-A. Meanwhile, in the Interpretability Module (demonstrated at the bottom of Fig. 2), we compute the temporal weights of different sentences, enabling the detection of shifts in viewpoints among various participants throughout the communication process. The detail of the Interpretability Module will be depicted in Section III-B.

A. Factual Consistency Module

We adhere to factual consistency [22], which requires the text to be faithful to its grounding source, regardless of the “correctness” w.r.t. the “real world”. We assess faithfulness by merely considering the information present in the input text rather than external knowledge. Models may yield wrong results when generating a summary that differs from the detailed information in the given document. According to the research on this phenomenon [15], [22], [23], the reason may be attributed to the restriction of pre-trained data source, the lack of support from external knowledge source and insufficient treatment of uncertain information. To overcome this deficit, we propose the GPT self-analysis structure to capture the factual consistency information to reduce the influence of wrong inference progress.

In our proposed structure, we inject noise into the generated summary candidates and then use GPT as the pre-trained LLM to analyze the altered information. In particular, we inject noise into the summary candidates via the Sentence Shuffle Rearrangement Mechanism, Entity Replacement Mechanism, and Paraphrase Mechanism.

1) *Sentence Shuffle Rearrangement Mechanism*: The pre-trained task of the T5 Module inspires the main idea of our Sentence Shuffle Rearrangement Mechanism. We randomly shuffle the sentences in the dialogue and use the GPT API to reorder the shuffled dialogue with the guidance information provided by the prediction summary. We use the Sentence

Shuffle Rearrangement Mechanism to evaluate the factual consistency between summary and dialogue sentences.

The whole progress can be formulated as follows: the sequence of dialogue sentences could be represented as $S = (s_1, s_2, \dots, s_n)$, where s_i is the i -th sentence in the dialogue progress. We shuffle this information using a random permutation function π , which is a bijection mapping the indices $1, 2, \dots, n$ onto itself, as follows:

$$\pi : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\}, \quad (1)$$

where $\pi(i)$ indicates the new position of the i -th sentence after shuffling. With the assistance of π , the shuffled dialogue sequence S' can be defined as follows:

$$S' = [s_{\pi(1)}, s_{\pi(2)}, \dots, s_{\pi(n)}]. \quad (2)$$

We use the GPT API denoted as $\text{GPT}()$ with task description request T_{Reorder} to generate the reordered $\tilde{S}_{SSRM}$ results through the task mapping function as follows:

$$\text{GPT}(S', T_{\text{Reorder}}) \rightarrow \tilde{S}_{SSRM}. \quad (3)$$

Then, we use the absolute position loss Loss_{SSRM} to evaluate the factual consistency between S and $\tilde{S}_{SSRM}$:

$$\text{Loss}_{SSRM} = \sum_{i=1}^n |i - \text{Index}_{\tilde{S}_{SSRM}}(s_i)|, \quad (4)$$

where $\text{Index}_{\tilde{S}_{SSRM}}(s_i)$ is the index number of the sentence s_i in $\tilde{S}_{SSRM}$. The parameters of temperature and top-p nucleus sampling in generating progress are important for the learning performance of GPT, which controls the randomness and diversity of the model’s output, respectively. Their effects on our model will be evaluated in Section IV-C.

2) *Entity Replacement Mechanism*: This mechanism injects factual inconsistencies into dialogue sentences. In this process, we modify details in the dialogue and use GPT to analyze the factual consistency.

We focus on the rationales generated by GPT, as they describe its inference process. These rationales reveal how the model captures subtle relationships among objects, characters, and references.

Our data generation process consists of three main steps for introducing noise. First, we perform entity substitution by replacing entities with random words. Second, we conduct action substitution, replacing character behaviors with random actions. Third, we implement summary manipulation, altering summaries to contradict the dialogue. Finally, we prompt GPT to generate explanations for the factual misalignments in these noise-injected summaries. These explanations provide valuable insights into the model’s reasoning capabilities.

We use the GPT API with the task description request T_{ERM} to generate noise candidates $\tilde{S}_{\text{ERM}}$:

$$\text{GPT}(S, T_{\text{ERM}}) \rightarrow \tilde{S}_{\text{ERM}}. \quad (5)$$

Then, we instruct GPT through instruct prompting $T_{\text{FactEvaluate}}$, to analyze whether this altered information is

¹A representative prompt is designed as follows: “Focus solely on summarizing the key points and do not analyze or comment on the languages used in the conversation. Let’s think step by step”.

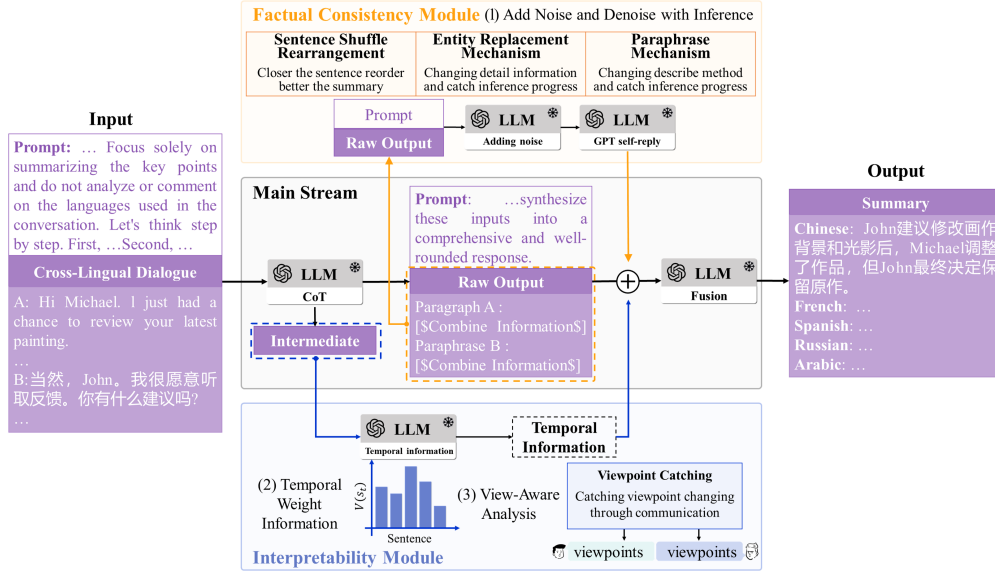


Fig. 2. The proposed framework consists of three components: the main stream of CoT with LLMs in the center, the Factual Consistency Module at the top, and the Interpretability Module at the bottom. In the Factual Consistency Module, we introduce noise to challenge the LLM and extract factual information. In the Interpretability Module, we utilize temporal information and conduct view-aware analysis.

correct with the summary candidate set C and its inference progress, and analyze the factual consistency value s_{fact} :

$$s_{fact} = \text{GPT}(\tilde{S}_{ERM}, C, T_{FactEvaluate}) \in [0, 1]. \quad (6)$$

We use the loss Loss_{ERM} which tries to instruct GPT to catch factual consistency information under the guidance of the ground-truth summary:

$$\text{Loss}_{ERM} = \max_{\text{GPT}(T_{Fact})} s_{fact}. \quad (7)$$

3) *Paraphrase Mechanism*: In this mechanism, we insert the noise information to maintain the idea of summary candidates. In specific, we paraphrase the summary by replacing the word with another word with a similar meaning. Moreover, we also change the sentence structure and tone, attempting to confuse the factual consistency module.

Similar to the Entity Replacement Mechanism, we use GPT with the task description request T_{PM} to create a paraphrase $\tilde{S}_{PM}$ through raw summary S . Then, we propose self-reply analysis, which evaluates the factual consistency score s_{fact} as follows:

$$\text{GPT}(S, T_{PM}) \rightarrow \tilde{S}_{PM}. \quad (8)$$

We minimize the loss Loss_{PM} to keep the crucial information without changing descriptions as follows:

$$\text{Loss}_{PM} = - \max_{\text{GPT}(T_{Fact})} \tilde{S}_{PM}. \quad (9)$$

B. Interpretability Module

One of the primary challenges researchers encounter is the black-box nature of LLMs. Within the CoT framework, the CLDS task is divided into summarization and translation, enabling a deeper understanding of the intermediate processes.

To improve interpretability and capture critical details, we introduce the temporal weight and view-aware analysis designed to extract meaningful insights from the model's decision-making process.

1) *Temporal Weight*: During interactions, participants' viewpoints may shift dynamically over time. To capture this temporal variation, we adopt the Temporal-Difference algorithm [24]. This method leverages current state values while incorporating future states in reinforcement learning, allowing for a more adaptive understanding of evolving perspectives.

Given the sentence value at time t as s_t , we employ the LLM to evaluate s_t , denoted as $V(\cdot)$. The temporal-difference value $V(s_t)$ is then updated using the following function:

$$V(s_t) \leftarrow V(s_t) + \alpha(r_{t-1} + V(s_{t-1}) - V(s_t)), \quad (10)$$

where α represents the learning rate, r_{t-1} denotes the weight of information gain at time $t-1$, as generated by GPT.

2) *View-Aware Analysis*: Since participants' attitudes and perspectives naturally evolve during communication, we introduce a view-aware prompting structure. This approach integrates CoT reasoning as additional context, allowing the GPT API to enhance self-guidance and generate more effective responses. Throughout this process, we constrain the viewpoint to a limited number of words, providing GPT with a clear framework to comprehend the overall structure of the communication. Subsequently, we utilize the task description T_{view} to extract the relevant information $\tilde{S}_{View}$ as

$$\text{GPT}(S, T_{View}) \rightarrow \tilde{S}_{View}. \quad (11)$$

IV. EXPERIMENTS

A. Dataset

We evaluate our approach using the MSAMSum [25] and ClidSum [26] datasets. For MSAMSum, we focus on the Mix-

To-Many task to assess performance in a multilingual setting. For ClidSum, we use the XSAMSum and XMediaSum40k subsets for our evaluation.

B. Experimental Settings

Our experiments are conducted using the GPT-3.5-turbo and GPT-4o API configurations. In addition to GPT-based methods, we also compare with the following baselines:

- **mT5**: We fine-tune the mT5 model on the MSAMSum dataset using an end-to-end approach for downstream tasks. This setup directly maps input sequences to the target outputs;
- **mBART**: We fine-tune mBART on the MSAMSum dataset using the same end-to-end methodology.

We fine-tune the model using the Adam optimizer with hyperparameters set to $\beta_1 = 0.9$, $\beta_2 = 0.998$, and $\epsilon = 10^{-9}$. For evaluation, we utilize beam search with a beam width of four and apply a repetition penalty of 5.0. The temperature is set to 0.75, and the top-p parameter in the GPT API is configured to 0.9. These experimental settings facilitate a comparative evaluation of the performance of state-of-the-art pre-trained LLMs. Under the CoT paradigm, the results produced by the summarization and translation modules are jointly consolidated and systematically integrated via a structured reasoning process.

C. Experimental Results

We evaluate all models using ROUGE-1/2/L F1 scores. For non-English languages, we employ language-specific tokenizers (e.g., Jieba for Chinese, SpaCy for Russian, and NLTK for Arabic) prior to evaluation.

As shown in Table I, on the MSAMSum dataset, our proposed model, which integrates factual consistency and interpretability, significantly outperforms the CoT baseline. It proves effective across all six evaluated languages, improving the ROUGE-1 score from 0.3266 to 0.3440 in the Mix2English CLDS task. Further analysis reveals that these two components are synergistic; an ablation study confirms that their integration boosts the ROUGE-1 score from 0.2587 to 0.3440.

To validate our method’s generalizability, we conduct experiments on the ClidSum dataset. The results demonstrate state-of-the-art performance in both monolingual and cross-lingual summarization. Specifically, for cross-lingual tasks from English, our model increases the ROUGE-1 score from 0.1653 to 0.2487 for German summaries and from 0.1554 to 0.2241 for Chinese summaries, underscoring its robustness.

1) **Interpretability Analysis**: To better capture the interpretability information in our structure, we introduce a case study for analysis. In Fig. 3, we visualize the intermediate results of a given sample, which exhibit superior interpretability features compared to other methods. We first extract view-aware information to serve as a skeleton for guiding GPT. Additionally, we introduce noise to the generated candidates.

As illustrated in the case study presented in Fig. 3, our proposed method effectively and intuitively captures the shifts in participants’ opinions throughout the dialogue, from the

Dialogue Sentence:

```
[en_XX] John: Hi Michael, I just had a chance to review your latest painting.
[en_XX] John: It' s really impressive, but I think there are a few tweaks we could consider to make it even better.
[zh_CN] Michael: 当然, John. 我很愿意听取反馈. 你有什么建议吗?
%Translation: Sure, John. I' m open to feedback. What do you have in mind?
[en_XX] John: Well, the background feels a little too busy.
[en_XX] John: Maybe simplifying it could help draw more attention to the main subject.
[en_XX] John: Also, the lighting on the figure could use a bit more contrast—it' s a bit flat right now.
[zh_CN] Michael: 明白了. 简化背景, 增强主体人物的光影对比. 还有其他需要修改的吗?
%Translation: Got it. Simplify the background and enhance the lighting contrast on the figure. Anything else?
[en_XX] John: No, that' s about it. I think these changes will make the painting pop more.
[zh_CN] Michael: 好的. 我会根据你的建议做些调整, 然后把更新版本给你看看.
%Translation: Okay. I' ll work on a few adjustments and share some updated versions with you.
[zh_CN] Michael: John, 我根据你的建议修改了画作, 来看一下吧——你觉得怎么样?
%Translation: Hey John, I' ve prepared a revised version based on your suggestions. Take a look—what do you think?
[en_XX] John: Hmm. Actually, let' s just go with the original.
[zh_CN] Michael: ..... 好的.
%Translation: ... Alright.
```

View-Aware Analysis:

```
John: Initial Feedback → Revised Decision
Christian: Acknowledgment of Feedback → Reporting the Revision → Acceptance of Final Decision
```

Predict:

```
[en_XX] John initially suggests refinements to Michael's painting, but after reviewing the revisions, he decides the original version is best, which Michael accepts.
[fr_XX] John suggère initialement des améliorations à la peinture de Michael, mais après avoir examiné la version révisée, il décide que la version originale est la meilleure.
[es_XX] John inicialmente sugiere mejoras en la pintura de Michael, pero después de revisar la versión revisada, decide que la versión original es la mejor, lo que Michael acepta.
[zh_CN] John 最初建议对 Michael 的画作进行改进, 但在审视修改版本后, 他决定还是原版最好.
```

Fig. 3. An example of the analysis visualization. In our proposed structure, we first attract the view-aware information and make the skeleton of the entire dialogue. Then, we inject noise into the generated summary, highlighted in red, to make our structure better understand the contextual information. Finally, we use cross-lingual dialogue to generate its summary in the target language based on the self-reply information.

initial need for project revisions to the eventual decision to retain the first draft. Furthermore, the intermediate outputs generated during this process enhance the interpretability of our approach.

V. CONCLUSION

In this paper, we introduce a factual replacement mechanism that incorporates noise information to improve cross-lingual summarization analysis. A self-guidance framework ensures factual consistency, aiding information flow tracking in cross-lingual hybrid dialogues using GPT-based models. A view-aware CoT-driven architecture enhances interpretability and transparency in the summarization process. The proposed CLDS framework integrates temporal dialogue characteristics and uses CoT prompts to boost factual consistency and interpretability. Experiments show the approach outperforms baseline methods in CLDS.

ACKNOWLEDGMENT

This project has received funding from the National Natural Science Foundation of China (Nos.62306156, 72571150, and 62106091);the Tianjin Municipal Science and Technology Bureau (No.25JCZDSN00020);and the Shandong Provincial Natural Science Foundation (No.ZR2025MS1078).

REFERENCES

- [1] Y. Cai and Y. Yuan, “CAR-Transformer: Cross-attention reinforcement transformer for cross-lingual summarization,” in *Proceedings of AAAI*, 2024, pp. 17 718–17 726.

TABLE I
ROUGE MEASURED CLDS PERFORMANCE ACROSS MSAMSUM AND CLIDSUM DATASETS LANGUAGE SUBSETS.

MSAMSum Dataset									
Models	English			French			Spanish		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
mT5 fine-tuning	.2587	.0674	.1843	.2453	.0598	.1820	.2395	.0571	.1836
mBART fine-tuning	.2498	.0721	.1932	.2532	.0701	.1654	.2201	.0432	.1763
GPT-4o End2End	.3284	.1110	.2492	.3245	.1304	.2455	.3285	.1291	.2519
CoT method	.3354	.1143	.2541	.3350	.1417	.2595	.3348	.1276	.2553
DeepSeek	.3451	.1158	.2534	.3387	.1356	.2523	.3386	.1305	.2589
Claude	.3425	.1105	.2496	.3371	.1346	.2512	.3349	.1295	.2548
Qwen	.3442	.1098	.2505	.3401	.1368	.2552	.3358	.1301	.2574
Ours	.3440	.1208	.2587	.3498	.1405	.2637	.3423	.1325	.2687
Models	Arabic			Russian			Chinese		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
mT5 fine-tuning	.1226	.0147	.0732	.2057	.0432	.1576	.2567	.1042	.2031
mBART fine-tuning	.1341	.0304	.0894	.2210	.0457	.1498	.2432	.1124	.2005
GPT-4o End2End	.1820	.0346	.1584	.2767	.0785	.2226	.3303	.1673	.2573
CoT method	.1905	.0358	.1594	.2876	.0809	.2375	.3507	.1810	.2742
DeepSeek	.1912	.0381	.1553	.2905	.0823	.2356	.3558	.1856	.2758
Claude	.1920	.0376	.1532	.2889	.0810	.2305	.3524	.1815	.2742
Qwen	.1956	.0389	.1568	.2920	.0843	.2325	.3564	.1863	.2761
Ours	.1983	.0405	.1601	.2940	.0876	.2381	.3598	.1888	.2789
ClidSum Dataset									
Models	English			German			Chinese		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
mT5 fine-tuning	.1841	.0658	.1341	.1653	.0597	.1263	.1554	.0631	.1005
mBART fine-tuning	.1876	.0623	.1317	.1607	.0554	.1301	.1589	.0674	.1076
GPT-4o End2End	.2336	.0800	.1754	.2070	.0733	.1560	.1866	.0915	.1390
CoT method	.2576	.0954	.1812	.2274	.0743	.1597	.2014	.0987	.1489
DeepSeek	.2558	.0984	.1825	.2371	.0823	.1652	.2156	.1005	.1534
Claude	.2518	.0910	.1805	.2351	.0805	.1625	.2143	.0995	.1520
Qwen	.2564	.0986	.1825	.2395	.0835	.1663	.2158	.1012	.1553
Ours	.2604	.1002	.1841	.2487	.0894	.1701	.2241	.1035	.1571

- [2] N. Liu, K. Wei, Y. Yang *et al.*, “Multimodal cross-lingual summarization for videos: A revisit in knowledge distillation induced triple-stage training method,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10697–10714, 2024.
- [3] X. Shi, “MCLS: A large-scale multimodal cross-lingual summarization dataset,” in *Proceedings of CCL*, 2023, pp. 273–288.
- [4] S. Takeshita, T. Green, N. Friedrich *et al.*, “Cross-lingual extreme summarization of scholarly documents,” *Int. J. Digit. Libr.*, vol. 25, no. 2, pp. 249–271, 2024.
- [5] J.-D. Wang, D. Chang, F.-Q. Meng *et al.*, “A comprehensive survey and prospect of cross-lingual summarization method research,” *J. Netw. Intel.*, vol. 9, no. 1, pp. 384–412, 2024.
- [6] J. Zhu, Q. Wang, Y. Wang *et al.*, “NCLS: Neural cross-lingual summarization,” in *Proceedings of EMNLP-IJCNLP*, 2019, pp. 3054–3064.
- [7] Y. Cao, H. Liu, and X. Wan, “Jointly learning to align and summarize for neural cross-lingual summarization,” in *Proceedings of ACL*, 2020, pp. 6220–6231.
- [8] Y. Cao, X. Wan, J. Yao *et al.*, “MultiSumm: Towards a unified model for multi-lingual abstractive summarization,” in *Proceedings of AAAI*, 2020, pp. 11–18.
- [9] N. Liu, K. Wei, X. Sun *et al.*, “Assist non-native viewers: Multimodal cross-lingual summarization for how2 videos,” in *Proceedings of EMNLP*, 2022, pp. 6959–6969.
- [10] S. Takeshita, T. Green, N. Friedrich *et al.*, “X-SCITLDR: Cross-lingual extreme summarization of scholarly documents,” in *Proceedings of JCDL*, 2022, pp. 1–12.
- [11] X. Wan, F. Luo, X. Sun *et al.*, “Cross-language document summarization via extraction and ranking of multiple summaries,” *Knowl. Inf. Syst.*, vol. 58, pp. 481–499, 2019.
- [12] Y. Zhang, S. Gao, Y. Huang *et al.*, “A cross-lingual summarization method based on cross-lingual fact-relationship graph generation,” *Pattern Recog.*, vol. 146, p. 109952, 2024.
- [13] J. A. Bishop, Q. Xie, and S. Ananiadou, “LongDocFACTScore: Evaluating the factuality of long document abstractive summarisation,” in *Proceedings of LREC-COLING*, 2024, pp. 10777–10789.
- [14] H. Y. Koh, J. Ju, M. Liu *et al.*, “An empirical survey on long document summarization: Datasets, models, and metrics,” *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–35, 2022.
- [15] R. Aharoni, S. Narayan, J. Maynez *et al.*, “Multilingual summarization with factual consistency evaluation,” in *Proceedings of ACL*, 2023, pp. 3562–3591.
- [16] M. Gao, W. Wang, X. Wan *et al.*, “Evaluating factuality in cross-lingual summarization,” in *Proceedings of Findings of ACL*, 2023, pp. 12415–12431.
- [17] J. Li, J. Chen, H. Chen *et al.*, “Multilingual generation in abstractive summarization: A comparative study,” in *Proceedings of LREC-COLING*, 2024, pp. 11827–11837.
- [18] Y. Lin, Z. Liu, and M. Sun, “Neural relation extraction with multi-lingual attention,” in *Proceedings of ACL*, 2017, pp. 34–43.
- [19] N. Bhatnagar, A. Urlana, V. Mujadia *et al.*, “Automatic data retrieval for cross lingual summarization,” in *Proceedings of ICNLP*, 2023, pp. 822–827.
- [20] J. Ma, Y. Huang, L. Wang *et al.*, “Augmenting low-resource cross-lingual summarization with progression-grounded training and prompting,” *ACM T. Asian and Low-Reso.*, vol. 23, no. 9, pp. 1–22, 2024.
- [21] J. Wei, X. Wang, D. Schuurmans *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” in *Proceedings of NeurIPS*, 2022, pp. 24824–24837.
- [22] O. Honovich, R. Aharoni, J. Herzig *et al.*, “TRUE: Re-evaluating factual consistency evaluation,” in *Proceedings of ACL*, 2022, pp. 161–175.
- [23] Y. Li, X. Hao, and L. Li, “A survey of factual consistency in summarization from 2021 to 2023,” in *Proceedings of CAIT*, 2023, pp. 84–89.
- [24] E. W. Myers, “An $O(ND)$ difference algorithm and its variations,” *Algorithmica*, vol. 1, no. 1, pp. 251–266, 1986.
- [25] X. Feng, X. Feng, and B. Qin, “MSAMSum: Towards benchmarking multi-lingual dialogue summarization,” in *Proceedings of ACL*, 2022, pp. 1–12.
- [26] J. Wang, F. Meng, Z. Lu *et al.*, “Clidsum: A benchmark dataset for cross-lingual dialogue summarization,” in *Proceedings of EMNLP*, 2022, pp. 7716–7729.