GLOBESUMM: A Challenging Benchmark Towards Unifying Multi-lingual, Cross-lingual and Multi-document News Summarization

Anonymous ACL submission

Abstract

News consumption in today's globalized world presents a challenge due to the overwhelming amount of multilingual information and diverse perspectives provided by different sources. However, there has been a lack of research addressing this issue. To address this gap, we aim to unify **M**ulti-lingual, **C**ross-lingual and **M**ulti-document **S**ummarization into a novel task, i.e., **MCMS**. Nevertheless, the absence of a benchmark dataset inhibits researchers from effectively studying this invaluable problem. To tackle this obstacle, we have meticulously constructed the GLOBESUMM dataset. Specifically, we start by collecting a vast amount of news data and retrieving multilingual news related to the same event. We then meticulously design a reference annotation method called protocol-guided prompting, which utilizes the domain expertise of human annotators and effectively captures news *redundancies*, *omissions*, and *conflicts* to guide large-scale language models (LLMs) in generating high-quality reference summaries. Through extensive experimental analysis, we validate the quality of our dataset and highlight the inherent challenges of the task. We firmly believe that the GLOBESUMM, given its challenging nature, will contribute to advancements in global news summarization and the evaluation of LLMs[1].

1 Introduction

Summarization is a long-standing task in natural language processing research (Paice, 1990). In recent years, significant advancements have been made in the field thanks to the rapid development of large language models (LLMs) (Zhao et al., 2023; Liu et al., 2023; Dong et al., 2023; Wei et al., 2022a,b; Shanahan, 2022). While LLMs have effectively addressed many traditional text summarization tasks (Adams et al., 2023; Goyal et al., 2022a; Pu et al., 2023; Zhang et al., 2023), the rapid globalization of information dissemination has created new demands for summarization techniques that can effectively summarize event-centric multilingual news articles worldwide.

Events involved with armed conflicts, international relations, and political elections have always fascinated people worldwide. However, relying solely on news articles in a single language to gain an in-depth understanding of such events can be limiting. This is because news reports from different countries are often influenced by their national standpoints and cultural biases, resulting in potential distortions (Boykoff and Boykoff, 2004; Baum and Groeling, 2009; Baumeister and Hastings, 2013). To obtain a more comprehensive insight into these events, it is crucial to explore news articles from various countries and languages, allowing us to consider diverse perspectives and access more objective information. Surprisingly, while advancements in LLMs have shown promising results in many natural language processing tasks, little research has been conducted on the specific task of summarizing event-centric multilingual news articles (Goyal et al., 2022b).

To this end, we present the task of MCMS that unifies **M**ulti-lingual, **C**ross-lingual and **M**ulti-document **S**ummarization into a more general setting, aiming to align better with the multifaceted requirements in real-world scenarios. The goal of MCMS is to succinctly capture the key information from a collection of documents written in various languages and present a cohesive summary in the target language. It is worth noting that the MCMS task has three distinctive features: (1) the input consists of multiple documents, (2) the multiple documents are in different languages, and (3) the multiple documents revolve around the same event. However, there is currently no dataset available that encompasses the aforementioned three features for researchers to study.

To close this gap, we meticulously construct the

[1]Dataset and codes will be released upon publication.

Error Type	Definition	Example (pair of sentences)	Strategy & Explanation
Redundancy	Same information or facts repeated in both news reports.	A: No se han registrado heridos en la explosión. B: Никто не пострадал при дистанционном взрыве первой бомбы рядом с автомобилем.	Remove the same information or facts repeated in both news reports, avoid unnecessary duplication in summary.
Omission	Additional information not present in the other.	/	Include the additional information that is not present in the other, avoid potentially leading to an incomplete understanding of the event.
Conflict	Conflicts arise when there are contradictory or incompatible details.		
-- Time Updates	The inconsistencies arising from the evolving updates over time between initial news reports and subsequently developing news.	A: أعلنت هيئة البث الرواندية، الأربعاء، عن صلخشأ ١٠٩ قافو نتيجة الانهيارات الأرضية والفيضانات في غرب البلاد. B: Liczba zabitych w wyniku powodzi i osuwisk ziemi na zachodzie Rwandy wzrosła do 129, według informacji podanych przez tamtejsze władze.	Overwrite the original information with subsequent information and refrain from mentioning the original information, avoid causing confusion with information from different times.
-- Perspectives	The contradictions among news articles regarding the same detail, arising from diverse standpoints or differing viewpoints.	A: Soldados israelíes realizaron una operación este jueves que terminó en la muerte de tres terroristas palestinos. B: Premier Mohammad Shtayyeh ha accusando il governo israeliano di essere "responsabile dei suoi orribili crimini e delle continue violazioni contro il popolo palestinese".	Coexist with these viewpoints and present them in an appropriate manner, while maintaining neutrality. (Because the essence of the summarization is to collect, organize and condense information, without delving into judgments of "right or wrong" values in this context)
-- Cultural Discrepancies	The misunderstandings that may arise from the multilingual nature of news reports or cultural discrepancies, especially concerning the same details when reported in different languages.	A: 지중해 난민 구조선을 운영하는 독일 구호단체 'SOS 휴머니티'는 23일 독일 정부에게서 약 11억4천만원 를 지원받기로 했다고 밝혔다. B: Az olasz kormány korábban fenntartásait fejezte ki azzal kapcsolatban, hogy a német kormány 790 ezer euró finanszírozást adott az SOS Humanity berlini szervezetnek.	Reconcile the conflicts with the expertise of LLMs, presenting them as reasonable statements from the perspectives of all the languages involved.
-- Inherent Error	The conflicts arising from the inherent errors within a specific news article itself.	A: L'ouragan Otis a touché terre avec une force de 5 sur l'échelle de Richter. B: Ураганът "Отис"бе оценен с категория 5 на скалата за сила на ураганите.	Correct the error with the potentially accurate information deduced from the news or the common sense knowledge already acquired.
-- Other	The conflict caused by some other unknown possible reasons.	/	Unify the conflicts with a general statement, minimizing the possibility of any misunderstanding or contradiction.

Table 1: The protocol, we formulated for MCMS task, includes definitions of all potential error types, along with corresponding real-life examples for each type and approximate resolution strategies.

GLOBESUMM dataset. **Regarding news collection**, we begin by collecting a massive amount of news data from the GDELT database[2], followed by a careful event-centric reranking and filtering process. **Regarding reference annotation**, we introduce a tailored method called protocol-guided prompting for high-quality and cost-effective reference annotation. Specifically, based on extensive manual observations, we first develop a protocol for creating a reference summary, which takes into account three main challenges of MCMS: *redundancies*, *omissions*, and *conflicts* (shown in Table 1). Surprisingly, this automatic reference annotation method combines the expertise of human annotators embedded in the protocol with the power of LLMs, aiming to guide the summarization process through human prior knowledge (high-quality) and reduce the burden of manual annotation (cost-effective). Notably, the GLOBESUMM presents new challenges for future models to concurrently address *redundancies*, *omissions*, and *conflicts* among multiple documents in a multilingual context, ultimately generating the final summary.

Building upon these foundations, we conduct extensive and comprehensive experiments and analyses based on LLMs. The results not only validate the high quality of the GLOBESUMM dataset that we have created but more importantly, highlight the challenges inherent in MCMS task and expose the limitations of current LLMs in this context. We firmly believe that the GLOBESUMM dataset will serve as an important benchmark for evaluating the multilingual capabilities and summarization performance of LLMs.

2 The GLOBESUMM Dataset

2.1 Data Collection

We gather news reports from countries and regions across 26 languages worldwide. This is accomplished by utilizing the news URLs provided in Google GDELT 2.0 project[3]. We exclusively utilize news data from May 2023 to October 2023 for

[2] https://www.gdeltproject.org/

[3] https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/

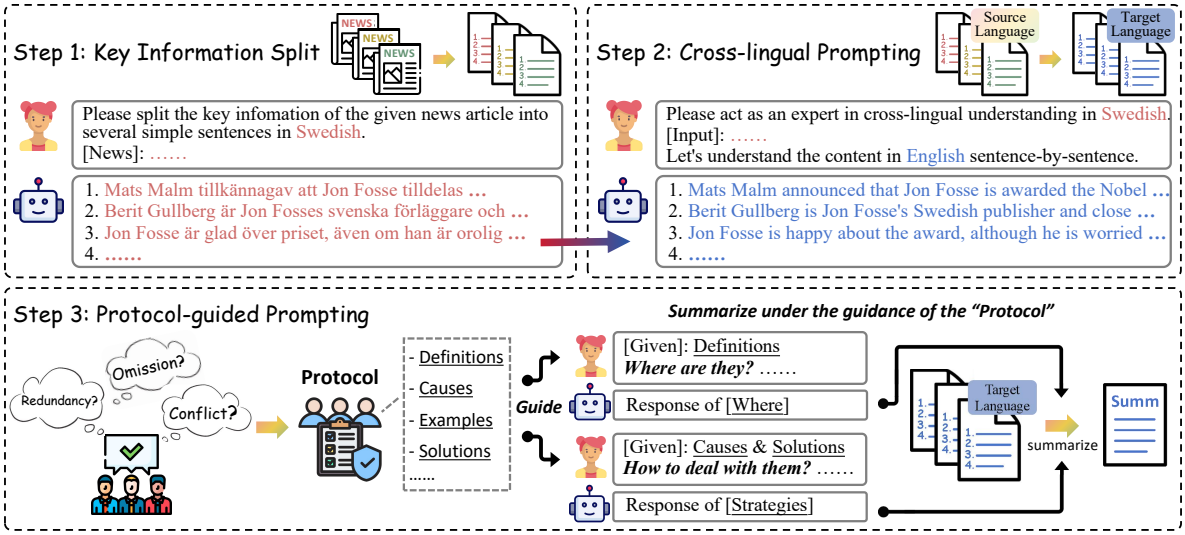

Figure 1: Overview of our reference annotation methodology. The method consists of key information split, cross-lingual prompting and protocol-guided prompting.

our benchmark to avoid introducing prior knowledge from web-crawled articles used in pre-training large language models[4].

2.2 Source Data Construction

One setting in MCMS is noteworthy that the multiple news reports within the same round input should be highly relevant to the same news event, rather than an open-domain task. To address this, we employ a method involving event retrieval and manual verification to group the news reports.

Event Retrieval To pinpoint news related to specific events, we leverage Wikipedia's current events portal[5] as a seed set. Each event in this set serves as a query input for our retrieval process. Our goal is to identify highly relevant news reports from the multilingual corpus. Initially, we translate the query event (originally in English) into multiple languages[6]. Subsequently, we employ the BM25 retriever in Lin et al. (2021) for retrieval in the respective language corpora, extracting the most query-relevant news articles.

Manual Verification The retrieved news articles in different languages are supposed to be highly relevant to the provided description, but *high relevance does not necessarily imply that they all present the same news event*. Hence, we incorporate a post-retrieval manual verification process (see Appendix 1.1).

2.3 Reference Annotation Methodology

Next, we will elaborate on how we seamlessly integrate the robust reasoning and language modeling capabilities of LLMs with the assistance of human expertise through a semi-automated reference annotation approach to craft our reference summaries.

Chronological Recurrent Summarization (CRS) Our reference annotation approach is conducted under the CRS schema, aiming to distill key information from news articles in chronological order.

Specifically, we begin by organizing these news documents in order of their respective timestamps. Then the summarization process is initiated by generating a concise summary for the first two articles. The obtained summary is then integrated with the subsequent article, and iteratively throughout the whole document set. CRS delivers a concise, timely summary by capturing the dynamic narrative and providing a comprehensive overview of the evolving information landscape in news articles.

Step 1: Key Information Split (KIS) The large input length of a whole document set, averaging nearly 12K tokens in GLOBESUMM, poses a great obstacle in MCMS. Therefore, we introduce the method of KIS to reduce the length of input by organizing key information from each document into several finely-grained sentences before summarizing the whole document set. The request is

[4]The cutoff of GPT-4-1106-preview is April 2023. https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

[5]https://en.wikipedia.org/wiki/Portal:Current_events

[6]We translate the queries by Google translation API

as follows (see more in Appendix 1.1):

> Please split the key information of the given news article into several simple sentences in {Source Language}.
> Request: {Given news article X}

Given the news article, we ask LLM to present the key information of the given news in sentence-level, rather than as a continuous block of text. This would be a more accessible form for identifying and addressing redundancy, omission, and conflict.

Step 2: Cross-lingual Prompting (CLP) Achieving cross-lingual alignment poses another fundamental challenge in multi- and cross-lingual tasks. To effectively capture alignment information from various input languages to our target language, we employ the cross-lingual alignment prompting method, which was first introduced in Qin et al. (2023). Specifically, the prompt is designed as:

> Please act as an expert in cross-lingual understanding in {Source Language}.
> Request: {Given content X}
> Let's understand the content in {Target Language} sentence-by-sentence.

Step 3: Protocol-guided Prompting (PGP) We first introduce the method of *protocol-guided prompting* (PGP) to achieve reference summary annotation. Based on our manual observation of diverse news articles across multiple languages and documents, we have concluded three primary hurdles in MCMS: *redundancies*, *omissions*, and *conflicts*. The details shown in Table 1, which constitute our protocol, will be incorporated as part of the prompt to assist the LLMs in more effectively identifying and handling these hurdles while summarizing the documents.

PGP does not simply rely on LLM's remarkable capabilities, but utilizes the human expertise embedded in the protocol, thereby achieving more reliable and high-quality task performance. More specifically, the procedure of how we address redundancy, omission and conflict can be broadly divided into two parts: (1) *where are they?* and (2) *how to deal with them?* (all the prompts can be found in Appendix 1.2)

(1) Where are they? (`[where]`) We furnish LLM with the definitions of redundancy, omission, and conflict (in Table 1) and request LLM to adeptly pinpoint the occurrences of these issues between documents based on provided definitions.

(2) How to deal with them? (`[strategies]`) As shown in our protocol (refer to Table 1), we have conducted a manual synthesis and conclusion for these issues, especially conflicts. Based on the various causes that may give rise to these problems, we have elegantly formulated different solutions. Then, we request LLM to delineate specific strategies for each conflict arising from different reasons in the actual scenarios, following the customized general solutions we have outlined.

With the assistance of the knowledge in our protocol, we effectively achieve the two subtasks of `[where]` and `[strategies]`. Consequently, LLM's responses to `[where]` and `[strategies]` are utilized to generate our reference summaries.

Dataset	# Events	# Docs	# Summs
Total			
Num	370	4687	4317
Num. Languages	-	26	-
Avg Token Length	11568.46	913.23	368.04
Train Set Size	222	2848	2626
Valid Set Size	74	897	823
Test Set Size	74	942	868

Table 2: Statistics of the dataset. The token length was calculated by *tiktoken*[7].

2.4 Statistics

Following the methodology described in Section 2.3, our reference summaries are generated with GPT-4 model[8] as the backbone.

A total of 370 news events, consisting of 4687 news articles, have been finally retained in GLOBESUMM. The entire dataset spans 26 languages and each news event is associated with a minimum of 10 news reports, adding to the complexity of our dataset. Due to the recurrent nature of CRS schema (Section 2.3), GLOBESUMM offers reference summaries for document subsets of any size within the whole document set, totaling 4317 in number.

And GPT-4's responses to `[where]` and `[strategies]` are also available in GLOBESUMM.

We split GLOBESUMM into train, validation and test sets (Table 2). Subsequent experiments (Section 4) are carried out on the test set. Our expenses can be found in Appendix D.

[7] https://github.com/openai/tiktoken

[8] All GPT-4 mentioned in this paper refer to GPT-4-1106-preview. In order to significantly reduce costs, only the PGP phase is handled by GPT-4, while KIS and CLP process are executed by GPT-3.5-turbo-16k.

Method (GPT-3.5-turbo)	AR	DE	EL	EN	ES	HI	RO	RU	TH	TR	VI	ZH	AVG
KIS vs. Summarize													
Summarize	34.7	49.4	37.7	53.5	49.1	34.4	49.2	33.7	38.7	42.0	43.1	51.4	43.1
Summarize-Extend	45.6	64.5	51.0	65.0	64.1	44.7	57.5	45.2	44.7	50.7	57.5	63.9	54.5
KIS	**50.9**	**66.9**	**55.0**	**74.5**	**69.4**	**50.9**	**65.8**	**50.9**	**50.5**	**51.9**	**62.5**	**67.7**	**59.8**
CLP vs. Translate													
Translate-En	51.3	62.9	60.4	-	62.6	**60.2**	62.9	54.7	**48.7**	60.2	58.7	49.5	57.5
CLP-En	**55.0**	**67.1**	**60.6**	-	**66.8**	56.6	**64.8**	**59.0**	47.0	**61.9**	**60.2**	**55.6**	**59.5**

Table 3: The Acc. performance of KIS vs. Summarize and CLP vs. Translate on XQuAD.

Aspect	P*	R*	F_1*
Redundancy	0.887	1.015	0.947
Omission	0.973	0.948	0.961
Conflict	0.845	1.938	1.177

Table 4: Precision*, recall* and F_1* scores of GPT-4's performance against human annotation. All values are calculated in micro-averaging.

3 GLOBESUMM Quality Assessment

We next examine the superiority of our reference annotation method.

3.1 Compete with Human Annotation

In this section, we evaluate how well GPT-4 addresses [where] and [strategies] under the guidance of our protocol by comparing its performance with human annotation.

We randomly select 50 pairs of documents in GLOBESUMM, with each pair focusing on the same event. Next, we invite human annotators to identify and mark the quantities, locations of redundancies, omissions, and conflicts between pairs of documents. This annotated data will serve as the golden standard for evaluating the performance of GPT-4.

Variations of precision, recall, and F_1 metrics (see Appendix 2.1) are utilized for evaluation.

The scoring results in Table 4 support that GPT-4 can effectively address [where] under the guidance of our protocol. GPT-4 performs comparably to human annotators in terms of **redundancy** and **omission**, with F_1 scores approaching human levels. In terms of **conflict**, we observe that GPT-4 outperforms human annotators in F_1 scores, and its recall value achieves nearly double that of humans (see Figure 3 in Appendix). The results present above strongly indicate that guided by our protocol, GPT-4 can effectively replace manual implementation in the subtask [where], even outperforming human capabilities.

Concerning the subtask [strategies], both addressing redundancies and omissions is straightforward through the removal of duplications and inclusion of additional information. We prioritize conflict resolution strategies, manually verifying those produced by GPT-4 for 50 sample pairs. The outcome reveals a 96.8% accuracy (90 out of 93), indicating that GPT-4 consistently generates correct, reasonable, and protocol-compliant strategies.

3.2 Component-wise Analysis

Next, we will explore where the advantages of KIS and CLP are specifically manifested. We conduct comparative experiments on XQuAD (Artetxe et al., 2020; Dumitrescu et al., 2021), exploring KIS versus Summarize and CLP versus Direct Translate (see detailed implementation in Appendix B).

(1) KIS results better condensing quality. As shown in Table 3, we find that KIS exhibits a remarkable superiority over Summarize across all languages (with 16.7% improvements on average accuracy), strongly indicating that the context after KIS is more comprehensible for LLMs compared to the summarized context. Recognizing the impact of compression ratios on the total information provided (KIS~343 tokens; Summarize~242 tokens), we introduce another control group named Summarize-Extend with a longer compressed context (averaging 579 tokens). Nevertheless, KIS still outperforms Summarize-Extend by 5.3% in accuracy, further illustrating that KIS is a better method for long-text compression.

(2) CLP outperforms direct translation in cross-lingual alignment. As depicted in Table 3, CLP demonstrates higher accuracy than Translate by averaging 2.0%, which illustrates that CLP can assist LLMs more effectively in achieving semantic alignment between languages, thereby enhancing cross-lingual comprehension. As verified in Qin et al. (2023), CLP is not a vanilla translation but utilizes the cross-lingual semantic alignment.

4 Experiments

4.1 Baselines

Our experiments are conducted with various baselines, each composed of a combination of schema and pipeline.

Schemas To validate the effectiveness of Chronological Recurrent Summarization (CRS), we investigate the two schemas for comparison: (a) *Single-turn Summarization* summarizes a document set within a single-turn generation; (b) *Chronological Recurrent Summarization* iteratively summarizes two documents at a time in a time-ordered manner (Section 2.3).

Pipelines To further validate the advantages of KIS and CLP in addressing excessively lengthy inputs and cross-language understanding, we conduct comparative tests with these commonly used methods: (a) *Translate-then-Summarize*; (b) *Summarize-then-Translate*; (c) *KIS-then-CLP*. Here, *Summarize* denotes the summarization of each document before summarizing the whole document set.

Similarly, we conduct experiments using two different approaches for summarization: (a) *Direct Summarization*; (b) *Protocol-guided Prompting*.

Models We select three representative LLMs that feature long context capability, each of which supports at least a 16k context window.

- **GPT-3.5-turbo-16k** is an advanced GPT-3.5 series model with a 16k context window. We evaluate its performance by requesting the official API provided by OpenAI with default settings.
- **Vicuna-7B-v1.5-16k** (Zheng et al., 2023a) is an open-source language model fine-tuned from Llama2, and supports a 16k context window.
- **ChatGLM3-6B-32k** (Du et al., 2022) is an open-source language model based on General Language Model (GLM) framework, and supports a 32K context window.

4.2 Metrics

We evaluate the quality of the generated summaries using following metrics (see Appendix 3.1 for detailed formulations):

- **ROUGE** (Lin, 2004) measures the overlap co-occurrence of n-grams between the candidate and reference summaries.
- **Red** (Chen et al., 2021) is a self-referenced metric for *redundancy* evaluation.
- **Normalized Inverse of Coverage (NIC)** captures *Omission*, as the inverse of a coverage of key information from reference summary.
- **Conflict Resolution Effectiveness (CRE)** metric evaluates how well a candidate summary addresses *conflict*.

4.3 Main Results

The main results are illustrated in Table 5 (see Table 10 in Appendix for full ROUGE results). From the results, we have the following observations:

(1) Omissions and Conflicts mitigated, yet Redundancies persist. As shown in Table 5, unlike omissions and conflicts, which can be mitigated with the introduction of our methodology (CRS, KIS, CLP and PGP), redundancies, on the contrary, tends to persist, even exacerbate. The results across all three models do not seem to reflect the effectiveness of our approach in addressing redundancy. This divergence in the impact on different types of issues emphasizes the nuanced nature of MCMS, highlighting its intricate and multifaceted characteristics.

(2) Preferential performance in CRS with Protocol-guided Prompting. From the results on GPT-3.5-turbo-16k and Vicuna-7b-v1.5-16k as illustrated in Table 5, we find that protocol-guided prompting outperforms `Direct` only under the CRS schema, while its superiority is not evident under STS. This is within our expectations, as STS requires LLMs to simultaneously identify and coordinate redundancies, omissions, and conflicts across all news documents. While the recurrent nature of CRS, processing only two news documents at a time, makes it a more reasonable summarization schema. Nevertheless, protocol-guided prompting still improves the performance, particularly when aligned with the CRS schema, supporting our conclusions in Section 3.1.

(3) LLM's Sensitivity to Protocol-guided Prompting. As illustrated in Table 5, protocol-guided prompting demonstrates certain advantages on both GPT-3.5-turbo-16k and Vicuna-7b-v1.5-16k. However, in the experiments with the Chatglm3-6b-32k model, regardless of the schema being STS or CRS, protocol-guided prompting performs worse than direct summarization. This indicates that the actual effectiveness of protocol-guided prompting is influenced by the model's performance. The implementation of protocol-guided

Schema & Pipeline	GPT-3.5-turbo-16k				Vicuna-7b-v1.5-16k				Chatglm3-6b-32k			
	R-L ↑	Red ↓	NIC ↓	CRE ↑	R-L ↑	Red ↓	NIC ↓	CRE ↑	R-L ↑	Red ↓	NIC ↓	CRE ↑
Single-turn Summaization (STS)												
Translate-then-Summarize + Direct	19.86	**30.06**	84.77	56.33	18.49	**29.38**	87.12	55.20	18.98	32.49	85.17	**58.43**
Summarize-then-Translate + Direct	20.07	30.57	84.59	56.18	19.08	30.70	87.44	54.62	19.45	32.82	86.87	56.05
KIS-then-CLP + Direct	21.25	30.91	82.05	59.02	18.96	30.73	87.19	53.37	19.27	35.06	85.08	57.11
Translate-then-Summarize + Protocol	19.41	31.20	89.34	54.29	19.07	30.09	87.82	53.06	18.76	30.33	87.20	55.51
Summarize-then-Translate + Protocol	19.18	31.04	87.79	54.23	18.69	29.43	86.73	55.22	18.76	31.44	88.10	53.76
KIS-then-CLP + Protocol	21.17	31.63	80.01	54.01	18.82	33.10	88.24	57.00	19.29	**30.15**	89.19	57.20
Chronological Recurrent Summarization (CRS)												
Translate-then-Summarize + Direct	20.27	33.36	82.06	54.58	20.14	32.37	77.99	56.28	20.08	32.80	**77.94**	55.95
Summarize-then-Translate + Direct	20.15	32.86	80.81	55.47	19.62	33.49	82.73	57.23	20.13	33.39	84.21	53.62
KIS-then-CLP + Direct	21.92	34.29	74.35	56.45	20.59	33.60	80.38	53.39	20.12	34.65	82.11	55.11
Translate-then-Summarize + Protocol	21.14	30.99	76.21	55.71	20.85	32.67	80.30	54.52	**20.48**	31.69	79.06	55.38
Summarize-then-Translate + Protocol	21.24	31.32	81.08	54.55	20.21	31.67	80.80	57.51	19.60	32.88	85.10	54.49
KIS-then-CLP + Protocol	**22.06**	32.30	**70.09**	**59.11**	**20.94**	34.92	**76.62**	**58.24**	20.19	33.28	85.15	54.86
Translate-then-Summarize.Avg	20.17	**31.40**	83.09	55.23	19.64	**31.13**	83.31	54.77	19.58	**31.83**	**82.34**	**56.32**
Summarize-then-Translate.Avg	20.16	31.45	83.57	55.11	19.40	31.32	84.42	**56.15**	19.49	32.63	86.07	54.48
KIS-then-CLP.Avg	**21.60**	32.28	**76.62**	**57.15**	**19.83**	33.09	**83.11**	55.50	**19.72**	33.29	85.38	56.07
STS.Avg	20.16	**30.90**	84.76	55.68	18.85	**30.57**	87.42	54.75	19.09	**32.05**	86.94	**56.34**
CRS.Avg	**21.13**	32.52	**77.43**	**55.98**	**20.39**	33.12	**79.80**	**56.20**	**20.10**	33.12	**82.26**	54.90
STS + Direct.Avg	**20.39**	**30.51**	**83.80**	**57.18**	18.84	**30.27**	**87.25**	54.40	**19.23**	33.46	**85.71**	**57.20**
STS + Protocol.Avg	19.92	31.29	85.71	54.18	**18.86**	30.87	87.60	**55.09**	18.94	**30.64**	88.16	55.49
CRS + Direct.Avg	20.78	33.50	79.07	55.50	20.12	33.15	80.37	55.63	**20.11**	33.61	**81.42**	54.89
CRS + Protocol.Avg	**21.48**	**31.54**	**75.79**	**56.46**	**20.67**	**33.09**	**79.24**	**56.76**	20.09	**32.62**	83.10	**54.91**

Table 5: The evaluation results for all configurations of schemas and pipelines on different LLMs. "Direct" represents direct summarization without a protocol, while "Protocol" signifies the summarization method utilizing the protocol-guided prompting approach. ↑ represents that the higher score the better and ↓ means the opposite.

Schema & Pipeline	GPT-3.5-turbo-16k			
	R-L ↑	Red ↓	NIC ↓	Con ↑
Single-turn Summaization (STS)				
None + Direct	17.60	**30.31**	85.01	49.52
KIS-only + Direct	19.60	30.82	86.26	55.58
KIS-then-CLP + Direct	**21.25**	30.91	**82.05**	**59.02**
None + Protocol	19.12	31.39	89.42	44.20
KIS-only + Protocol	19.00	**30.91**	87.87	**55.47**
KIS-then-CLP + Protocol	**21.17**	31.63	**80.01**	54.01
Chronological Recurrent Summarization (CRS)				
None + Direct	18.45	**31.58**	84.26	53.63
KIS-only + Direct	20.76	33.67	78.03	52.68
KIS-then-CLP + Direct	**21.92**	34.29	**74.35**	**56.45**
None + Protocol	19.47	**29.46**	82.72	53.94
KIS-only + Protocol	21.46	31.05	75.51	56.62
KIS-then-CLP + Protocol	**22.06**	32.30	**70.09**	**59.11**

Table 6: The evaluation results of ablation studies on *KIS-then-CLP* stage.

prompting requires the model to possess the ability to understand relatively complex prompts. We will further substantiate this conclusion in Section 5.2.

5 Analysis

5.1 Ablation Study

We conduct ablation studies to investigate the effect of the *KIS-then-CLP* stage. We run the ablations as the rows in Table 6 (see Table 7 for full ROUGE results in Appendix).

We observe in Table 6 that GPT-3.5-turbo-16k exhibits a noticeable performance decline in summarization when either the KIS or CLP steps are omitted. This also indicates that relying solely on the ability of LLMs themselves to handle long-text and multi-lingual inputs may not be an appropriate solution at present, highlighting the necessity of pre-emptively explicit text compression and cross-languages alignment for LLMs.

5.2 LLM's Scale-Effect on PGP

The sensitivity observation in Section 4.3 prompts our exploration into the impact of model scale on protocol-guided prompting. We conduct experiments on varying sizes of the llama2 (Touvron et al., 2023) series models, including 7b, 13b, and 70b (see detailed settings in Appendix 3.2).

We compare the average performance of Direct Summarization and Protocol-guided Prompting (Table 8, 9 in Appendix), the Δ results shown in Table 8 exhibit favorable changes in both omission and conflict aspects as the model size increases (NIC ↓: $13.96 \rightarrow 5.01 \rightarrow 4.36$; CRE ↑: $-4.96 \rightarrow 0.61 \rightarrow 3.32$). This indicates that with the growth of model scale, protocol-guided prompting can achieve better results compared to direct summarization. However, there is still no apparent alleviation of redundancy.

5.3 The Apathy towards Low-Resource Languages: Unjust Summarizer

Mukherjee et al. (2023) has sparked our interest in exploring LLM's prejudices across various languages. Within MCMS, we undertake relevant experiments (details can be found in Appendix 3.2).

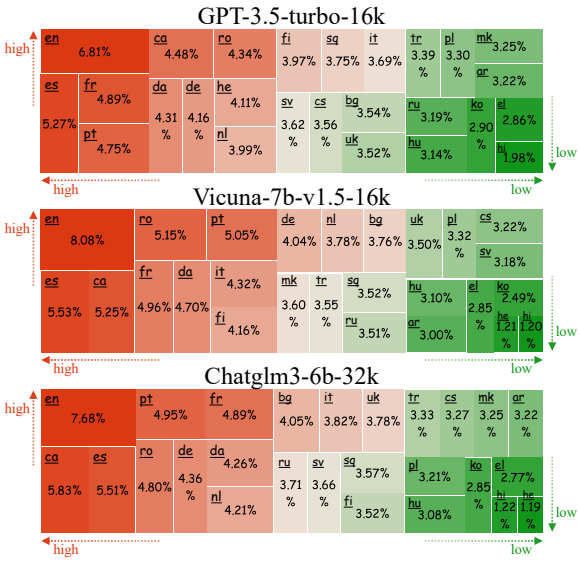

Figure 2: Normalized language attention distributions across 26 languages.

The results in Figure 2 The results on GPT-3.5-turbo-16k, Vicuna-7B-v1.5-16k, and ChatGLM3-6B-32k in Figure 2 all indicate that the content in the generated summaries tends to reflect documents in high-resource languages such as English and Spanish. Only a small portion of the content is concluded from documents in relatively low-resource languages like Hindi, Greek, Hebrew, and others. This phenomenon confirms the presence of a potential preference for high-resource languages in these models, thereby neglecting information from documents in low-resource languages. Therefore, playing the role of a fair summarizer is indeed challenging, even for current outperforming LLMs.

6 Related Work

6.1 Multi-lingual and Cross-lingual Summarization

Multi-lingual summarization (MLS) aims to process documents in multiple languages and generate their summaries in the corresponding language. The MultiLing-2015 dataset (Giannakopoulos et al., 2015) initiates interest in this task, leading to increasing subsequent studies (Vanetik and Litvak, 2015; Litvak et al., 2016; Cao et al., 2020b). Recently, with the availability of many large-scale MLS datasets (Varab and Schluter, 2021; Hasan et al., 2021; Feng et al., 2022), notable progress is achieved one after another.

Cross-lingual summarization (CLS) summarizes given documents in one language into summaries in another target language. The early work mainly focuses on pipeline methods (Yao et al., 2015; Ouyang et al., 2019; Wan et al., 2010), leading to error propagation. Recently, many large-scale CLS datasets (Zhu et al., 2019; Wang et al., 2022; Zheng et al., 2023b) are proposed, shifting the research attention to end-to-end studies (Cao et al., 2020a; Liang et al., 2022). Considering the close relation between MLS and CLS, Feng et al. (2022) evaluate the MLS models on CLS to show their zero-shot CLS ability, Wang et al. (2023) unifies MLS and CLS into a more general setting of many-to-many.

6.2 Multi-document Summarization

Multi-document summarization (MDS) refers to the task of summarizing the text in multiple documents into a concise summary. Previous studies have delved into various approaches, encompassing extractive (Angelidis and Lapata, 2018; Zheng et al., 2019; Mao et al., 2020) and abstractive techniques (Gehrmann et al., 2018; Lebanoff et al., 2018; Zhang et al., 2018). And researchers mainly focus on reducing the redundancy among documents (Peyrard et al., 2017; Xiao and Carenini, 2020; Chen et al., 2021). Due to the emergence of numerous datasets, traditional MDS tasks have become increasingly mature. Currently, there is a growing focus on MDS tasks in more diverse settings. Zhou et al. (2023) highlights the challenge of open-domain MDS, Amar et al. (2023) proposes aspect-based summarization in MDS to better fit the needs in real-world scenarios.

7 Conclusion

The capabilities required in MCMS are necessities in real-world scenarios. Our benchmark serves this demand as the first dataset for such scenario, offering high-quality summaries generated through protocol-guided prompting. Our protocol is derived from extensive human observations of news corpora, capturing human expertise to guide LLMs in summary generation. Through experiments and analysis, conducted on strong models, we unveil the shortcomings of LLMs in the MCMS and highlight the challenges of this task. Overall, we believe our benchmark holds the potential to be used in the future for evaluating the performance of LLMs in handling multi-lingual long-text tasks and how we utilize protocol-guided prompting can serve as a practical case for cost-effective annotation.

Ethics Statement

We utilize publicly available news data, which may contain viewpoints from different perspectives. The output results in the paper do not necessarily represent the views of the authors.

Limitations

While our dataset is constructed with GPT-4, budget constraints prevent us from exploring further experimental results on the GPT-4 model.

Our work primarily focuses on addressing redundancies, omissions, and conflicts among documents. However, in our attempts, we have found that while omissions and conflicts can be alleviated to some extent through our method, redundancies have not shown significant improvement.

Due to the recurrent nature of CRS, our reference summaries can cover any truncation length within the document set, as opposed to only providing a single final summary for each document set in many typical MDS datasets. However, in this work, there has not been an extensive investigation into this particular aspect, such as the impact of document quantity and language diversity on the difficulty of MCMS.

Gaining a profound understanding of a specific global news event involves more than the MCMS task discussed in our work. Exploring how to group news reports about the same event is also a worthwhile research endeavour. However, in the data construction phase of this study, the effectiveness of this step is ensured through manual post-validation without delving into its methodology.

References

Griffin Adams, Alexander Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: Gpt-4 summarization with chain of density prompting. *ArXiv preprint*, abs/2309.04269.

Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira, and Ido Dagan. 2023. Openasp: A benchmark for multi-document open aspect-based summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1991.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Matthew A Baum and Tim J Groeling. 2009. *War stories: The causes and consequences of public views of war*. Princeton University Press.

Roy F Baumeister and Stephen Hastings. 2013. Distortions of collective memory: How groups flatter and deceive themselves. In *Collective memory of political events*, pages 277–293. Psychology Press.

Maxwell T Boykoff and Jules M Boykoff. 2004. Balance as bias: Global warming and the us prestige press. *Global environmental change*, 14(2):125–136.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020a. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231, Online. Association for Computational Linguistics.

Yue Cao, Xiaojun Wan, Jin-ge Yao, and Dian Yu. 2020b. Multisumm: Towards a unified model for multilingual abstractive summarization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11–18. AAAI Press.

Wang Chen, Piji Li, and Irwin King. 2021. A training-free and reference-free summarization evaluation metric via centrality-weighted relevance and self-referenced redundancy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 404–414, Online. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2023. A survey for in-context learning. *ArXiv preprint*, abs/2301.00234.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu,

Razvan Pascanu, and Viorica Patraucean. 2021. Liro: Benchmark and leaderboard for romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. MSAMSum: Towards benchmarking multi-lingual dialogue summarization. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 1–12, Dublin, Ireland. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022a. News summarization and evaluation in the era of gpt-3. *ArXiv preprint*, abs/2209.12356.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022b. News summarization and evaluation in the era of gpt-3. *ArXiv*, abs/2209.12356.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022. A variational hierarchical model for neural cross-lingual summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2099, Dublin, Ireland. Association for Computational Linguistics.

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.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Marina Litvak, Natalia Vanetik, Mark Last, and Elena Churkin. 2016. MUSEEC: A multilingual text summarization tool. In *Proceedings of ACL-2016 System Demonstrations*, pages 73–78, Berlin, Germany. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. 2020. Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1737–1751, Online. Association for Computational Linguistics.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845, Singapore. Association for Computational Linguistics.

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2025–2031, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris D Paice. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1):171–186.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *ArXiv preprint*, abs/2309.09558.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *ArXiv preprint*, abs/2310.14799.

Murray Shanahan. 2022. Talking about large language models. *ArXiv preprint*, abs/2212.03551.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Natalia Vanetik and Marina Litvak. 2015. Multilingual summarization with polytope model. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 227–231, Prague, Czech Republic. Association for Computational Linguistics.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926, Uppsala, Sweden. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. ClidSum: A benchmark dataset for cross-lingual dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Towards unifying multi-lingual and cross-lingual summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15127–15143, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *ArXiv preprint*, abs/2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wen Xiao and Giuseppe Carenini. 2020. Systematically exploring redundancy reduction in summarizing long documents. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 516–528, Suzhou, China. Association for Computational Linguistics.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 118–127, Lisbon, Portugal. Association for Computational Linguistics.

Jianmin Zhang, Jiwei Tan, and Xiaojun Wan. 2018. Adapting neural single-document summarization model for abstractive multi-document summarization: A pilot study. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 381–390, Tilburg University, The Netherlands. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *ArXiv preprint*, abs/2301.13848.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *ArXiv preprint*, abs/2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena.

Shaohui Zheng, Zhixu Li, Jiaan Wang, Jianfeng Qu, An Liu, Lei Zhao, and Zhigang Chen. 2023b. Long-document cross-lingual summarization. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1084–1092.

Xin Zheng, Aixin Sun, Jing Li, and Karthik Muthuswamy. 2019. Subtopic-driven multi-document summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3153–3162, Hong Kong, China. Association for Computational Linguistics.

Yijie Zhou, Kejian Shi, Wencai Zhang, Yixin Liu, Yilun Zhao, and Arman Cohan. 2023. Odsum: New benchmarks for open domain multi-document summarization. *ArXiv preprint*, abs/2309.08960.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

A GLOBESUMM Construction Details

1.1 Source Data Construction

Manual Verification *"high relevance does not necessarily imply that they all present the same news event"*, here is a case for distinction:

> `[Description]` `Date: 2023-10-18`
> The U.S. Treasury Department announced the easing of certain oil, gas, and gold sanctions on Venezuela.
>
> (1) `[News1]` `Date: 2023-10-19`
> After reaching an agreement, the United States lifted sanctions on Venezuelan oil and gold ... ✓
>
> (2) `[News2]` `Date: 2023-05-09`
> Maduro calls the US takeover of oil company Citgo a violation of Venezuela's sovereignty ...

Both the `[News1]` and `[News2]` are highly relevant in overlapping terms (e.g. Venezuela, US, ...) with the given description. Obviously `[News1]` is the exact news event as described in the provided description, but it's challenging for a BM25 retriever to distinguish between them.

Therefore, we incorporate a post-retrieval manual verification. 5 annotators are invited to assess the relevance of retrieved news reports based on the specified event description. Only news meeting at least one of the following criteria is retained: (1) news that describes the same event as the given query, (2) news that involves the causes and consequences of the given query event and (3) news that reflects diverse perspectives on the given query event.

1.2 Reference Annotation Methodology

Key Information Split (KIS) In order to prevent information from becoming overly fragmented after being splitted, thereby overlooking the contextual connections, our prompt explicitly instructs the model to employ specific entity names instead of pronouns. The full request is formulated as follows:

> Please split the key information of the given news article into several simple sentences in {Source Language} and use specific entity names instead of pronouns whenever possible.
> Request: {Given news article X}

Protocol-guided Prompting (PGP) The full prompt for `[where]`, `[strategies]` and summarization process are provided in Table 11.

B Quality Assessment Details

2.1 Variations of Precision, Recall, and F_1 metrics

We have made slight modifications to the traditional precision, recall, and F_1 metrics for evaluating the performance of protocol-guided prompting with GPT-4. Here, we introduce the concept of False Positive Positive (FPP): predictions that do not align with the standard golden set but are still classified as correct answers after manual inspection. The formulas are as follows:

$$Precision^* = \frac{TP + FPP}{TP + FP}$$

$$Recall^* = \frac{TP + FPP}{TP + FN}$$

$$F_1^* = \frac{2 * Precision * Recall}{Precision + Recall}$$

Notably, the recall value may surpass 1, as GPT-4 has demonstrated the ability to detect additional redundancies, omissions, and conflicts overlooked by human annotators in practical testing. Through

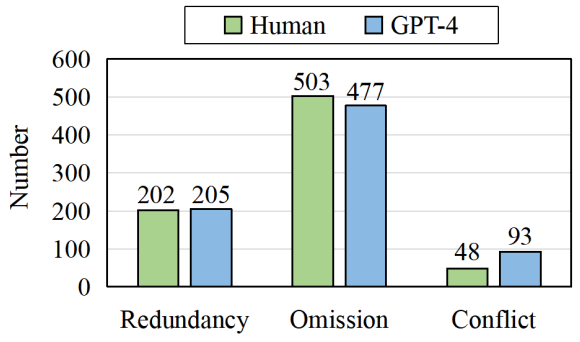

Figure 3: The comparison on the quantities of redundancies, omissions and conflicts identified by human annotators and GPT-4.

manual verification, some of these are also confirmed as accurate answers, contributing to the recall metric numerator.

2.2 Component-wise Analysis Details

The KIS step is introduced to shorten the cumulative input length of a document set, preventing excessive length. Summarizing each document before generating an overall summary is also a commonly used method for condensing. Thus, we conducted comparative experiments on XQuAD, a multi-lingual and cross-lingual QA dataset, exploring these two methods of compression. Similarly, CLP, aiming to achieve cross-lingual alignment between source and target languages, will be compared with the method of direct translation. The experiment assesses the impact of different processing methods on LLM's comprehension in Question-Answering (QA) by applying them individually to contexts from XQuAD.

C Experimental Details

3.1 Metrics Formulation

We evaluate the quality of summaries generated by different models and methods using following metrics:

- **ROUGE** (Lin, 2004) measures the overlap co-occurrence of n-grams between the candidate and reference summaries. We reported the F_1 scores for ROUGE.
- **Red** (Chen et al., 2021) is a self-referenced metric for *redundancy* evaluation. The summary itself is engaged as the reference to evaluate the degree of the semantic similarity between each summary sentences. The averaged semantic similarity result is used as the redundancy score. BERTScore (Zhang et al., 2020) is employed for similarity computation, and we use *deberta-xlarge-mnli* (He et al., 2021) as its backbone with its default setting for rescaling:

$$score_{red} = \frac{\sum_i \max_{j:i \neq j} \text{Sim}(\mathbf{x}_j, \mathbf{x}_i)}{|\mathbf{X}|},$$

where "$j : i \neq j$" means we do not consider the similarity between $\mathbf{x}_i$ and itself. We use F_1 in as the final redundancy score. Note that $score_{red} \in [-1, 1]$ and lower is better.

- **Normalized Inverse of Coverage (NIC)** captures *Omission*, as the inverse of a coverage of key information from reference summary. We employ an NLI model *t5_xxl_true_nli_mixture* (Honovich et al., 2022) to ascertain whether crucial information from the reference is entailed in the candidate summary.

$$coverage = \frac{count(entailed)}{count(sents)}$$

$$NIC = 1 - \frac{coverage}{log(|cand|)} * 10$$

where $coverage$ represents the coverage rate of a candidate summary, $count(entailed)$ means the number of sentences from reference summary entailed in candidate summary, $count(sents)$ means the total number of sentences in reference summary and $|cand|$ means the word-level length of the candidate summary[9]. Note that lower NIC is better.

- **Conflict Resolution Effectiveness (CRE)** metric evaluates how well a candidate summary addresses *conflict*. We use GPT-3.5-turbo as a referee to assess the conflict resolution strategies presented in the candidate summary. We employ conflicts identified by GPT-4 as the standard, assessing the effectiveness of the candidate summary's handling of conflicts based on the prompts provided in Table 12 and the result is present in a three-class classification of 1, 0, -1. In order to minimize errors resulting from conflicts update, we do not consider all conflicts identified in the summarization process for evaluation. Instead, we only select conflicts identified in the last 5 rounds of the CRS iteration process. The Con score is calculated as follows:

$$penalty = log(count(0) + e)$$

[9]The word-level length is calculated by nltk.

Schema & Pipeline	GPT-3.5-turbo-16k		
	R-1	R-2	R-L
Single-turn Summaization (STS)			
None + Direct	33.01	11.02	17.60
KIS-only + Direct	36.37	12.12	19.60
KIS-then-CLP + Direct	**41.55**	**13.76**	**21.25**
None + Protocol	35.01	12.16	19.12
KIS-only + Protocol	35.52	11.85	19.00
KIS-then-CLP + Protocol	**41.58**	**14.33**	**21.17**
Chronological Recurrent Summarization (CRS)			
None + Direct	34.94	11.08	18.45
KIS-only + Direct	42.12	14.16	20.76
KIS-then-CLP + Direct	**46.60**	**15.51**	**21.92**
None + Protocol	37.40	12.40	19.47
KIS-only + Protocol	41.58	15.30	21.46
KIS-then-CLP + Protocol	**47.97**	**16.41**	**22.06**

Table 7: The full ROUGE results of ablation studies on *KIS-then-CLP* stage.

$$score_{con} = \frac{count(1)}{count(1/-1) + \alpha * penalty}$$

where $count(*)$ represents the counting function, $penalty$ refers to the punitive consequence for neglecting conflict resolution. The coefficient α is set to 0.2 in our work.

3.2 Experimental implementations

Main experiments The results in Table 5 only present the ROUGE-L scores, while the full ROUGE results for the main experiment can be found in Table 10.

Ablation study The results in Table 6 only present the ROUGE-L scores, while the full ROUGE results for the ablation study can be found in Table 7.

LLM's Scale-Effect on PGP As we only investigate the impact on protocol-guided prompting, in order to reduce time and computational cost, we utilize the results of the pre-steps (translate-then-summarize, summarize-then-translate, and KIS-then-CLP) from GPT-3.5-turbo-16k model. The llama2 series models are only employed in the final summarization step. Due to the The llama2 models are running with the default settings in this project[10].

The Apathy towards Low-Resource Languages We employ an NLI model *t5_xxl_true_nli_mixture* (Honovich et al., 2022) to discern whether the sentences in the generated summary are entailed within the documents in the document set. If a sentence is entailed in a document, we consider it to be concluded from that document. Due to the NLI model's lack of capabilities in handling lengthy texts and multiple languages, we consider the output of the original document after undergoing KIS-then-CLP as the premise. Each sentence from the generated summary is then treated as a hypothesis. The entail score for each language is calculated as follows:

$$score_{lang} = \frac{\sum_S \frac{count(entailed)}{count(sents)}}{|S|}$$

$$norm_score_{lang} = \frac{score_{lang}}{\sum\limits_{lang \in langs} score_{lang}}$$

where S represents the set of generated summaries involved with the language, $count(entailed)$ means the number of sentences entailed in the document, $count(sents)$ means the total number of sentences in generated summary. $|S|$ means the number of summaries in S.

D Expenses and Compensation

The overall cost incurred for our reference annotation and experimental section utilizing GPT series models is approximately $900.

In the manual annotation phase of post-retrieval verification (Section 2.2), annotators will be compensated with $0.1 for each "yes" or "no" annotation completed. We have invited a total of 5 annotators, and they have collectively annotated 18787 news articles regarding their relevance to the provided event descriptions, resulting in a total expenditure of $1878.7. As for the construction of our protocol and the annotation process in Section 3.1, they are all undertaken by us paper authors without extra payment.

[10] https://github.com/notrichardren/llama-2-70b-hf-inference/blob/main/inference.py

Schema & Pipeline	Llama-2-7b-chat-hf				Llama-2-13b-chat-hf				Llama-2-70b-chat-hf			
	R-L ↑	Red ↓	NIC ↓	Con ↑	R-L ↑	Red ↓	NIC ↓	Con ↑	R-L ↑	Red ↓	NIC ↓	Con ↑
Chronological Recurrent Summarization (CRS)												
Translate-then-Summarize + Direct	19.18	28.82	84.66	53.01	17.71	25.66	88.51	51.21	18.15	26.26	81.94	52.47
Translate-then-Summarize + Protocol	15.52	40.35	96.04	45.07	14.88	25.24	92.95	50.28	17.49	31.30	85.93	56.35
Summarize-then-Translate + Direct	19.32	27.03	93.34	51.98	17.85	26.44	88.81	47.74	17.85	25.38	85.93	53.81
Summarize-then-Translate + Protocol	16.06	37.56	94.92	50.65	14.66	27.00	93.51	51.07	17.19	35.23	86.09	58.53
KIS-then-CLP + Direct	20.67	29.62	74.53	51.83	19.33	26.60	83.44	50.94	20.39	29.15	75.03	53.75
KIS-then-CLP + Protocol	16.39	38.94	93.45	46.29	16.37	26.48	89.33	50.37	18.01	25.77	82.19	55.10
Direct.Avg	19.72	28.49	80.84	52.27	18.30	26.23	86.92	49.96	18.80	26.93	80.97	53.34
Protocol.Avg	15.99	38.95	94.80	47.34	15.30	26.24	91.93	50.57	17.56	30.77	85.33	56.66
Δ = Protocol.Avg - Direct.Avg	- 3.73	10.46	13.96	- 4.94	- 3.00	0.01	5.01	0.61	- 1.24	3.84	4.36	3.32

Table 8: The evaluation results for Llama2 scale-effect analysis. "Direct" represents direct summarization without a protocol, while "Protocol" signifies the summarization method utilizing the protocol-guided prompting approach.

Schema & Pipeline	Llama-2-7b-chat-hf			Llama-2-13b-chat-hf			Llama-2-70b-chat-hf		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Chronological Recurrent Summarization (CRS)									
Translate-then-Summarize + Direct	41.34	13.03	19.18	34.05	10.70	17.71	35.72	11.19	18.15
Translate-then-Summarize + Protocol	29.73	7.13	15.52	28.65	8.41	14.88	35.42	10.44	17.49
Summarize-then-Translate + Direct	51.45	13.04	19.32	34.55	10.42	17.85	35.65	11.08	17.85
Summarize-then-Translate + Protocol	31.09	7.74	16.06	28.18	7.68	14.66	35.07	9.59	17.19
KIS-then-CLP + Direct	43.56	15.19	20.67	38.31	13.39	19.33	40.45	14.39	20.39
KIS-then-CLP + Protocol	31.34	8.46	16.39	32.18	9.71	16.37	36.41	11.16	18.01
Direct.Avg	45.45	13.75	19.72	35.64	11.50	18.30	37.27	12.22	18.80
Protocol.Avg	34.40	9.27	16.92	31.16	9.33	16.05	36.04	10.85	17.87
Δ = Protocol.Avg - Direct.Avg	-11.05	-4.48	-2.80	-4.48	-2.18	-2.25	-1.23	-1.37	-0.93

Table 9: The ROUGE results for Llama2 scale-effect analysis. "Direct" represents direct summarization without a protocol, while "Protocol" signifies the summarization method utilizing the protocol-guided prompting approach.

Schema & Pipeline	GPT-3.5-turbo-16k			Vicuna-7b-v1.5-16k			Chatglm3-6b-32k		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Single-turn Summaization (STS)									
Translate-then-Summarize + Direct	38.63	12.35	19.86	37.35	11.20	18.49	40.32	11.64	18.98
Summarize-then-Translate + Direct	37.94	12.36	20.07	37.72	11.25	19.08	39.90	11.94	19.45
KIS-then-CLP + Direct	41.55	13.76	21.25	38.53	11.52	18.96	40.49	11.59	19.27
Translate-then-Summarize + Protocol	37.12	11.95	19.41	38.14	11.35	19.07	38.34	11.50	18.76
Summarize-then-Translate + Protocol	36.35	11.62	19.18	38.23	11.44	18.69	38.33	11.23	18.76
KIS-then-CLP + Protocol	41.58	14.33	21.17	39.57	11.59	18.82	39.50	12.04	19.29
STS.Avg	38.86	12.73	20.16	38.26	11.39	18.85	39.48	11.66	19.09
Chronological Recurrent Summarization (CRS)									
Translate-then-Summarize + Direct	41.6	13.09	20.27	42.89	13.02	20.14	44.87	**13.87**	20.08
Summarize-then-Translate + Direct	41.59	13.18	20.15	42.21	12.90	19.62	44.31	13.12	20.13
KIS-then-CLP + Direct	46.6	15.51	21.92	44.75	13.74	20.59	**45.57**	13.62	20.12
Translate-then-Summarize + Protocol	42.78	14.45	21.14	44.46	13.77	20.85	43.90	13.42	**20.48**
Summarize-then-Translate + Protocol	42.68	14.29	21.24	43.10	13.09	20.21	42.52	12.80	19.60
KIS-then-CLP + Protocol	**47.97**	**16.41**	**22.06**	**46.64**	**14.56**	**20.94**	43.86	13.55	20.19
CRS.Avg	43.87	14.49	21.13	44.01	13.51	20.39	44.17	13.40	20.10
Translate-then-Summarize.Avg	40.03	12.96	20.17	40.71	12.34	19.64	41.86	12.61	19.58
Summarize-then-Translate.Avg	39.64	12.86	20.16	40.32	12.17	19.40	41.27	12.27	19.49
KIS-then-CLP.Avg	**44.43**	**15.00**	**21.60**	**42.37**	**12.85**	**19.83**	**42.36**	**12.70**	**19.72**
STS + Direct.Avg	**39.37**	**12.82**	**20.39**	37.87	11.32	18.84	**40.24**	**11.72**	**19.23**
STS + Protocol.Avg	38.35	12.63	19.92	**38.65**	**11.46**	**18.86**	38.72	11.59	18.94
CRS + Direct.Avg	43.26	13.93	20.78	43.28	13.22	20.12	**44.92**	**13.54**	**20.11**
CRS + Protocol.Avg	**44.48**	**15.05**	**21.48**	**44.73**	**13.81**	**20.67**	43.43	13.26	20.09

Table 10: The ROUGE F_1 scores for all configurations of schemas and pipelines on different LLMs. "Direct" represents direct summarization without a protocol, while "Protocol" signifies the summarization method utilizing the protocol-guided prompting approach. ↑ represents that the higher score the better and ↓ means the opposite.

Prompt for [where] & [strategies]

From news report 1

Request: {Given new1 X_1}

From news report 2

Request: {Given new2 X_2}

The above is the key information from two different news reports about the same event. Please clearly indicate if there are any redundancies, omissions and conflicts between each numbered sentence.
Definitions for "Redundancy", "Omission", "Conflict".
1. Redundancy: The instances where the same information or facts are repeated in both news reports, creating unnecessary duplication.
2. Omission: Omissions occur when one news report provides additional information that is not present in the other, potentially leading to an incomplete understanding of the event.
3. Conflict: Conflicts arise when there are contradictory or incompatible details between the two reports, leading to confusion or doubt about the accuracy of the information.

{Response of [where]}

Regarding the conflicts above, kindly specify the respective conflict types and provide specific solution strategies for each conflict.
- If you think the conflict arises from the updates of news events over time, please overwrite the original information with subsequent information.
- If you think the conflict arises from the contradictions of diverse perspectives, please coexist with these viewpoints and present them in an appropriate manner.
- If you think the conflict arises from linguistic misunderstandings or cultural discrepancies, kindly leverage your expertise to reconcile it, presenting them as reasonable statements from the perspectives of the languages involved.
- If you think the conflict is caused by errors in the news report itself, please correct it with accurate information deduced from the news or the common sense knowledge you already acquired.
- If you think the conflict is caused by some other unknown reasons or you can't handle the conflict with your knowledge, please use a general statement to unify them, minimizing the possibility of any misunderstanding or contradiction.

{Response of [strategies]}

Prompt for Summarization process

From news report 1

Request: {Given new1 X_1}

From news report 2

Request: {Given new2 X_2}

The above are the summarized key information from two different news reports about the same event. Please follow the given [Rules] and helpful hints [where] and [strategies] to integrate the two different summaries into a overall fluent concise summary of the event.

[Rules]
1. Redundancy: Remove the same information or facts repeated in both news reports, avoid unnecessary duplication in summary.
2. Omission: Include the additional information that is not present in the other, avoid potentially leading to an incomplete understanding of the event.
3. Conflict: Harmonize contradictory or incompatible details in news reports in a judicious manner, there are several solution strategies for dealing with different kinds of conflicts.
- If you think the conflict arises from the updates of news events over time, please overwrite the original information with subsequent information and refrain from mentioning the original information.
- If you think the conflict arises from the contradictions of diverse perspectives, please coexist with these viewpoints and present them in an appropriate manner.
- If you think the conflict arises from linguistic misunderstandings or cultural discrepancies, kindly leverage your expertise to reconcile it, presenting them as reasonable statements from the perspectives of the languages involved.
- If you think the conflict is caused by errors in the news report itself, please correct it with accurate information deduced from the news or the common sense knowledge you already acquired.
- If you think the conflict is caused by some other unknown reasons or you can't handle the conflict with your knowledge, please use a general statement to unify them, minimizing the possibility of any misunderstanding or contradiction.

[Where]
The hints guide you to identify redundancies, omissions, and conflicts.

Request: {Response of [where]}

[Strategies]
The hints offer proposed solution strategies for each conflict that you have to strictly follow.

Request: {Response of [strategies]}

Table 11: The prompts used in [where], [strategies] and summarization process.

Prompt for Conflict Evaluation

You will be given a summary of multiple news articles and the summary aims to handle the conflicts between news articles.

Your task is to determine whether the summary has effectively addressed the given potential conflicts.

If the summary effectively addresses the conflict, answer with 1.
if the summary does not involve the conflict, answer with 0.
If the summary does not address the conflict, answer with -1.

Conflicts:
{Standard conflicts identified by GPT-4}

Summary:
{Candidate summary}

For each conflict in the given conflicts, please determine whether the given summary resolves the conflict according to the rules mentioned above. Output the result in the format of a Python list, where each element in the list should be only one of the numbers 1, 0, or -1.

Table 12: The prompts used for conflict evaluation.