EMNLP 2024

Multi-News+: Cost-efficient Dataset Cleansing via LLM-based Data Annotation

Paper

Code

Juhwan Choi
gold5230@cau.ac.kr

Jungmin Yun
cocoro357@cau.ac.kr

Kyohoon Jin
fhzh123@cau.ac.kr

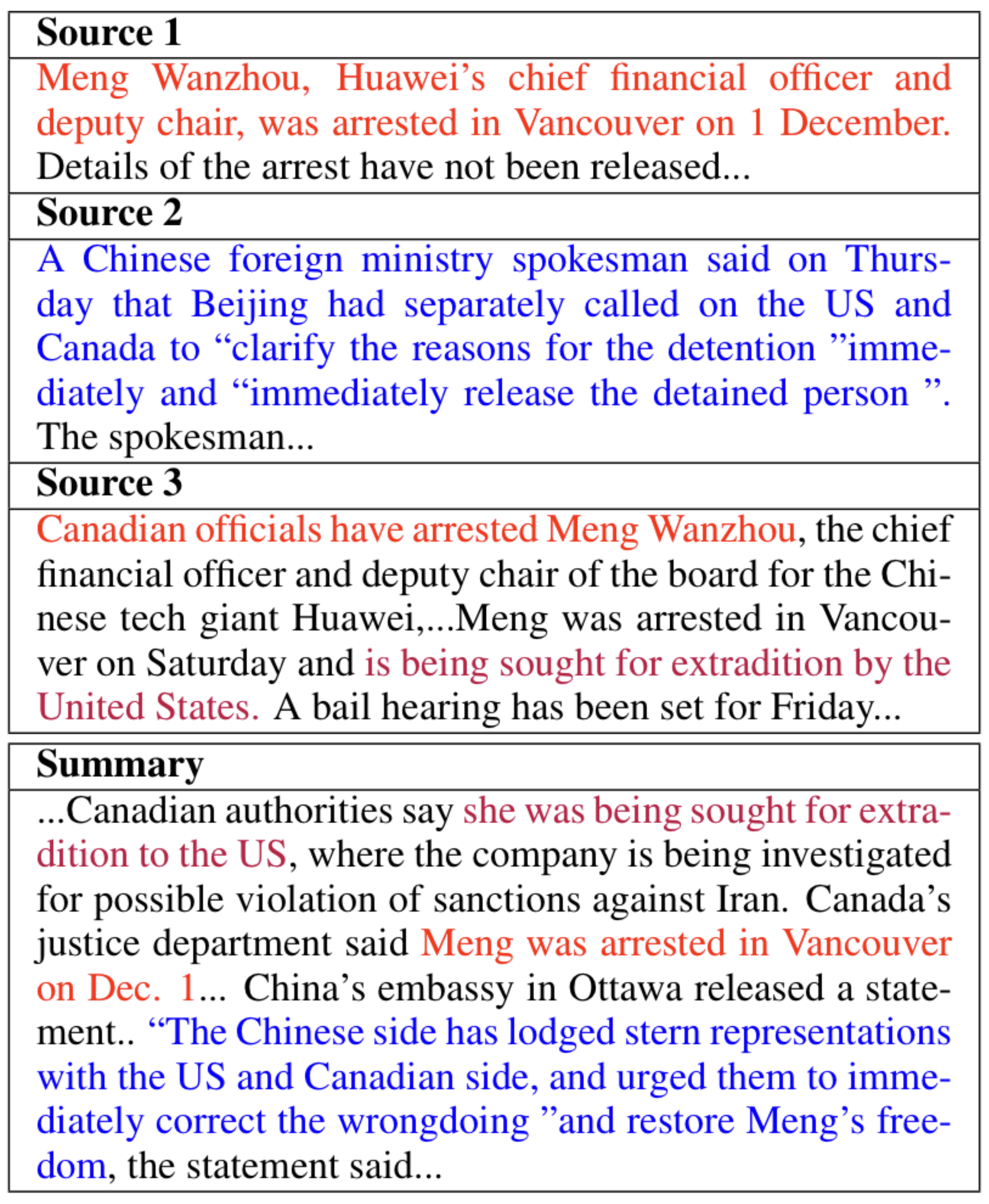
YoungBin Kim
ybkim85@cau.ac.kr

Chung-Ang University

Preliminary and Motivation: Multi-News

- **Multi-News**[1] is widely used dataset for multi-document summarization research
- It contains more than 50,000 sets and 150,000 documents
- However, Multi-News suffers from noisy documents because of error of crawling agents during the construction process
- We propose to leverage LLMs to classify these noisy documents instead of human annotators, establishing cost-efficient dataset cleansing method
- We construct **Multi-News+**, an enhanced version of Multi-News with our framework

Good Case



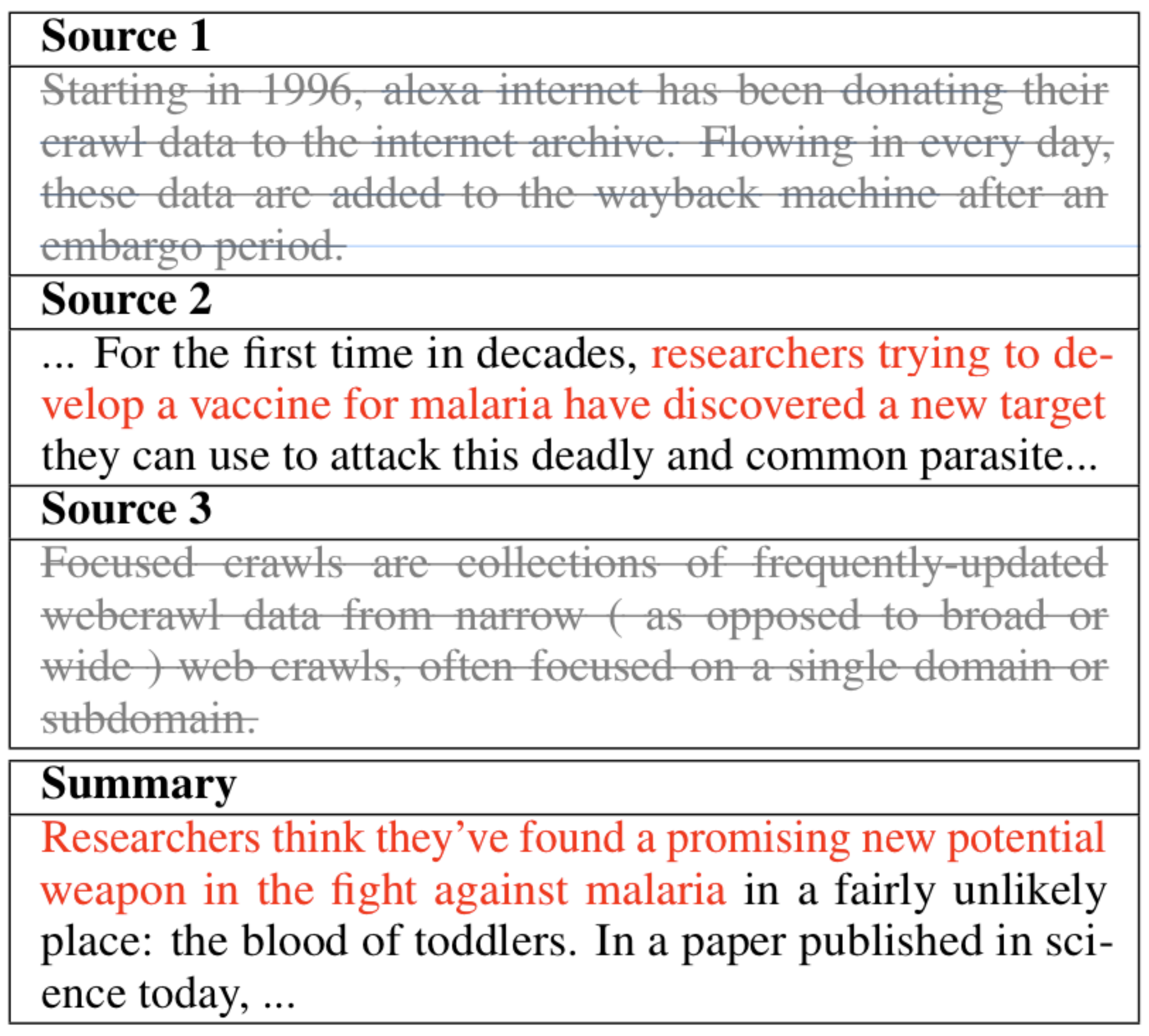
Source 1
Meng Wanzhou, Huawei's chief financial officer and deputy chair, was arrested in Vancouver on 1 December. Details of the arrest have not been released...

Source 2
A Chinese foreign ministry spokesman said on Thursday that Beijing had separately called on the US and Canada to "clarify the reasons for the detention "immediately and "immediately release the detained person ". The spokesman...

Source 3
Canadian officials have arrested Meng Wanzhou, the chief financial officer and deputy chair of the board for the Chinese tech giant Huawei,...Meng was arrested in Vancouver on Saturday and is being sought for extradition by the United States. A bail hearing has been set for Friday...

Summary
...Canadian authorities say she was being sought for extradition to the US, where the company is being investigated for possible violation of sanctions against Iran. Canada's justice department said Meng was arrested in Vancouver on Dec. 1... China's embassy in Ottawa released a statement.. "The Chinese side has lodged stern representations with the US and Canadian side, and urged them to immediately correct the wrongdoing "and restore Meng's freedom, the statement said...

Bad Case



Source 1
~~Starting in 1996, alexa internet has been donating their crawl data to the internet archive. Flowing in every day, these data are added to the wayback machine after an embargo period.~~

Source 2
... For the first time in decades, researchers trying to develop a vaccine for malaria have discovered a new target they can use to attack this deadly and common parasite...

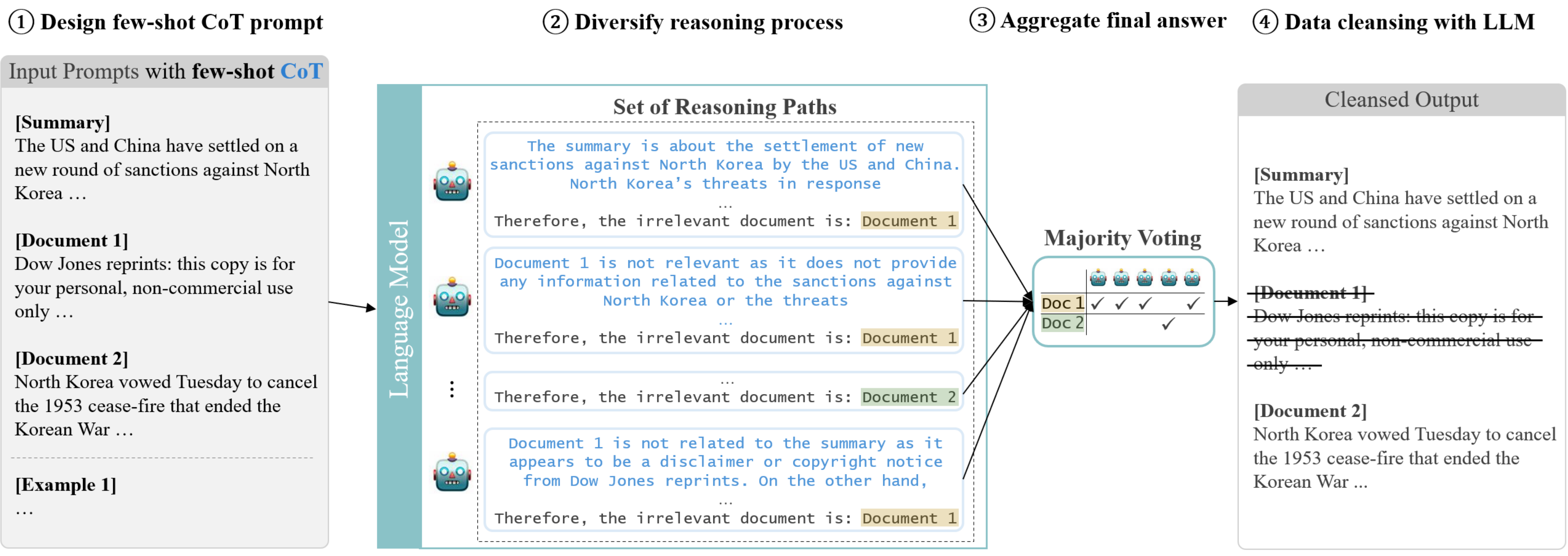
Source 3
~~Focused crawls are collections of frequently-updated webcrawl data from narrow (as opposed to broad or wide) web crawls, often focused on a single domain or subdomain.~~

Summary
Researchers think they've found a promising new potential weapon in the fight against malaria in a fairly unlikely place: the blood of toddlers. In a paper published in science today, ...

[1] Fabbri et al., *Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model*, ACL 2019

Method

- **Chain-of-Thought**: We apply CoT prompting to improve the accuracy of identification and providing rationale for future investigation
- **Majority Voting**: Similar to human annotation with majority voting, we use five LLMs to perform annotation and only identify documents where more than three agents agreed



① Design few-shot CoT prompt
② Diversify reasoning process
③ Aggregate final answer
④ Data cleansing with LLM
Input Prompts with few-shot CoT
Language Model
Set of Reasoning Paths
Majority Voting
Cleansed Output

[Summary]
The US and China have settled on a new round of sanctions against North Korea …

[Document 1]
Dow Jones reprints: this copy is for your personal, non-commercial use only …

[Document 2]
North Korea vowed Tuesday to cancel the 1953 cease-fire that ended the Korean War …

[Example 1]
…

```
The summary is about the settlement of new sanctions against North Korea by the US and China. North Korea's threats in response
…
Therefore, the irrelevant document is: Document 1

Document 1 is not relevant as it does not provide any information related to the sanctions against North Korea or the threats
…
Therefore, the irrelevant document is: Document 1

…
Therefore, the irrelevant document is: Document 2

Document 1 is not related to the summary as it appears to be a disclaimer or copyright notice from Dow Jones reprints. On the other hand,
…
Therefore, the irrelevant document is: Document 1
```

	🤖	🤖	🤖	🤖	🤖
Doc 1	✓	✓	✓		✓
Doc 2				✓	

[Summary]
The US and China have settled on a new round of sanctions against North Korea …

~~[Document 1]~~
~~Dow Jones reprints: this copy is for your personal, non-commercial use only …~~

[Document 2]
North Korea vowed Tuesday to cancel the 1953 cease-fire that ended the Korean War ...

Experiments and Insights

- We constructed Multi-News+ with our proposed framework, removing 18% of documents from Multi-News
- We trained BART and T5 model by Multi-News and Multi-News+, and found that Multi-News+ brings better performance than Multi-News
- Additionally, our ablation study with the dataset cleansing method for single document summarization shows its ineffectiveness for multi-document summarization

Model	*BART-large-cnn*				
Metric	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	BARTScore
Multi-News	48.64	18.86	24.11	0.6401	-2.763
MULTI-NEWS+	**49.17**	**19.04**	**24.36**	**0.6418**	**-2.698**
Ablation (Urlana et al., 2022)	47.48	18.27	23.81	0.6362	-2.767
Model	*T5-base*				
Metric	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	BARTScore
Multi-News	40.11	13.90	21.58	0.6003	-2.407
MULTI-NEWS+	**40.45**	**14.17**	**21.84**	**0.6027**	**-2.362**
Ablation (Urlana et al., 2022)	39.30	13.65	21.42	0.5967	-2.457