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

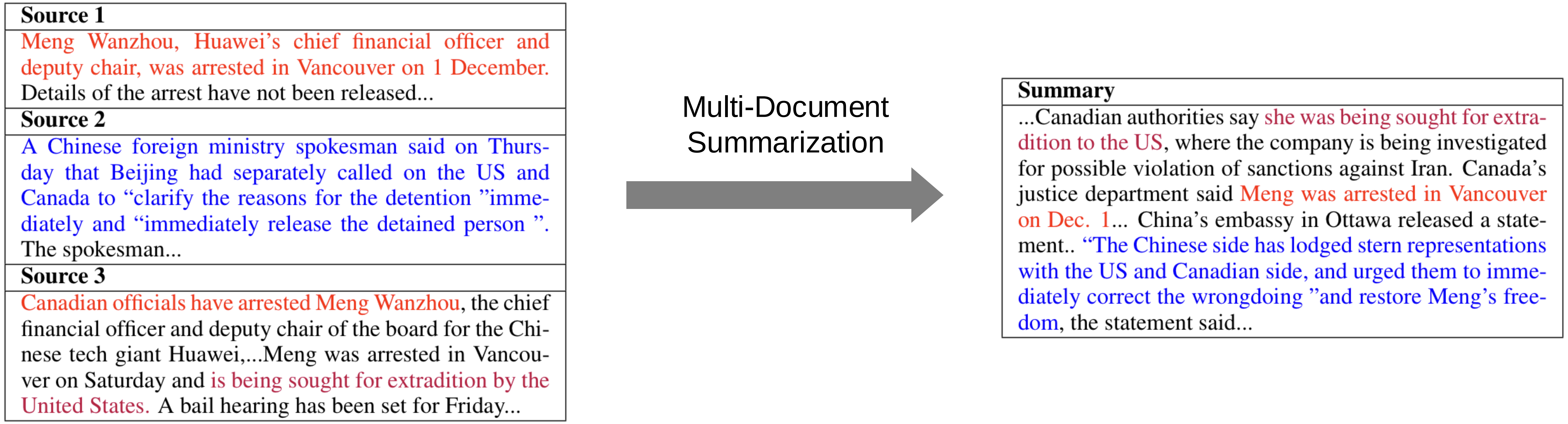
Juhwan Choi, Jungmin Yun, Kyohoon Jin and Youngbin Kim

Chung-Ang University

Motivation: Multi-News for Multi-Document Summarization

Multi-News[1] is widely used dataset for multi-document summarization tasks:

- It contains 56,216 sets of 156,289 news documents and their summarizations
- The dataset was constructed by automatic web crawling



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...

Multi-Document Summarization

Summary

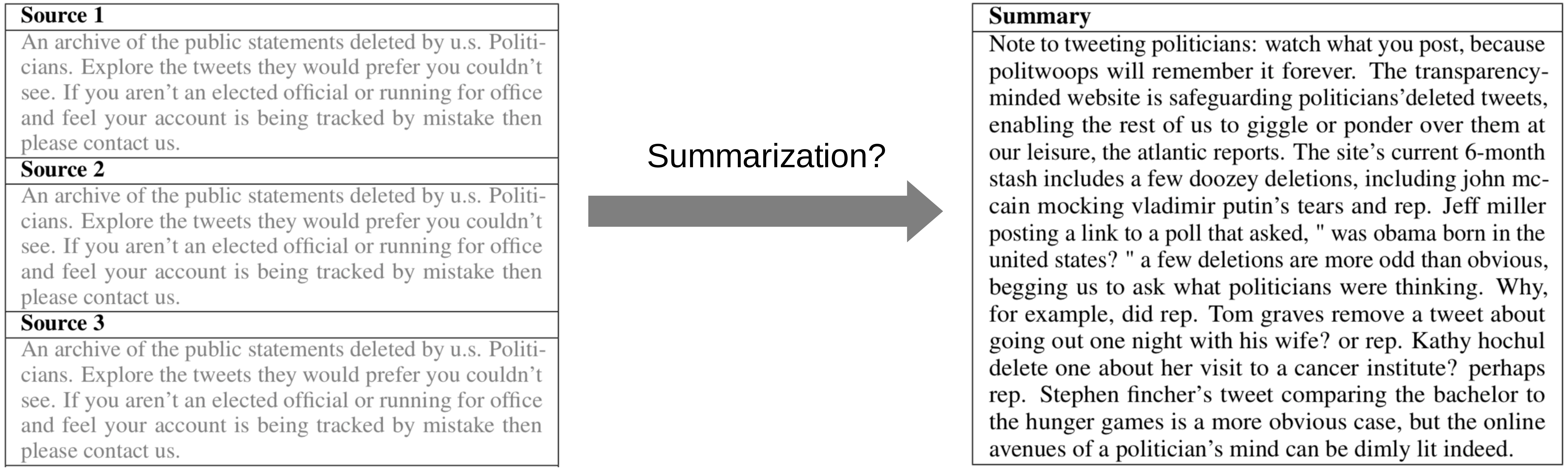
...Canadian authorities say she was being sought for extradition to the US, where the company is being investigated for possible violation of sanctions against Iran. Canada’s justice department said Meng was arrested in Vancouver on Dec. 1... China’s embassy in Ottawa released a statement.. “The Chinese 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...

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

Motivation: Quality Issues of Multi-News

Despite the value of Multi-News, we found that **Multi-News has quality issues**:

- During the crawling process, the crawler gathered unrelated, noisy documents
  - Such as system messages, promotional texts
- These noisy documents may disturb the training procedure
- Therefore, a method to classify and remove such noisy documents is required



Source 1

An archive of the public statements deleted by u.s. Politicians. Explore the tweets they would prefer you couldn’t see. If you aren’t an elected official or running for office and feel your account is being tracked by mistake then please contact us.

Source 2

An archive of the public statements deleted by u.s. Politicians. Explore the tweets they would prefer you couldn’t see. If you aren’t an elected official or running for office and feel your account is being tracked by mistake then please contact us.

Source 3

An archive of the public statements deleted by u.s. Politicians. Explore the tweets they would prefer you couldn’t see. If you aren’t an elected official or running for office and feel your account is being tracked by mistake then please contact us.

Summarization?

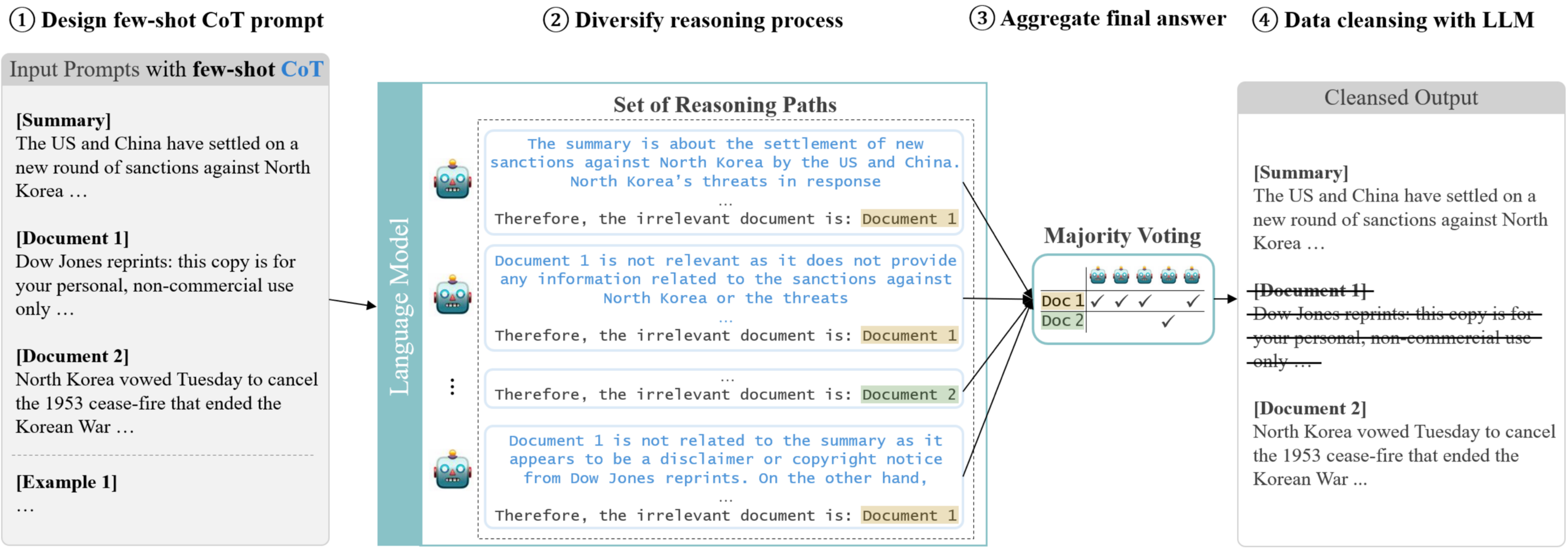
Summary

Note to tweeting politicians: watch what you post, because politwoops will remember it forever. The transparency-minded website is safeguarding politicians’deleted tweets, enabling the rest of us to giggle or ponder over them at our leisure, the atlantic reports. The site’s current 6-month stash includes a few doozey deletions, including john mccain mocking vladimir putin’s tears and rep. Jeff miller posting a link to a poll that asked, " was obama born in the united states? " a few deletions are more odd than obvious, begging us to ask what politicians were thinking. Why, for example, did rep. Tom graves remove a tweet about going out one night with his wife? or rep. Kathy hochul delete one about her visit to a cancer institute? perhaps rep. Stephen fincher’s tweet comparing the bachelor to the hunger games is a more obvious case, but the online avenues of a politician’s mind can be dimly lit indeed.

Dataset Cleansing via LLM-Based Annotation

However, it is expensive to hire human annotators to recognize noisy documents:

- Inspired by recent studies to establish large language models (LLMs), we proposee leveraging LLMs instead of human annotators
- Using LLM-based data annotation, we introduce a dataset cleansing framework to identify and remove noisy documents
- We construct **Multi-News+**, an enhanced version of Multi-News with this framework



① **Design few-shot CoT prompt**

Input Prompts with **few-shot** CoT

[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]
…

② **Diversify reasoning process**

Language Model

Set of Reasoning Paths

```
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
```

③ **Aggregate final answer**

Majority Voting

	🤖	🤖	🤖	🤖	🤖
Doc 1	✓	✓	✓		✓
Doc 2				✓	

④ **Data cleansing with LLM**

Cleansed Output

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

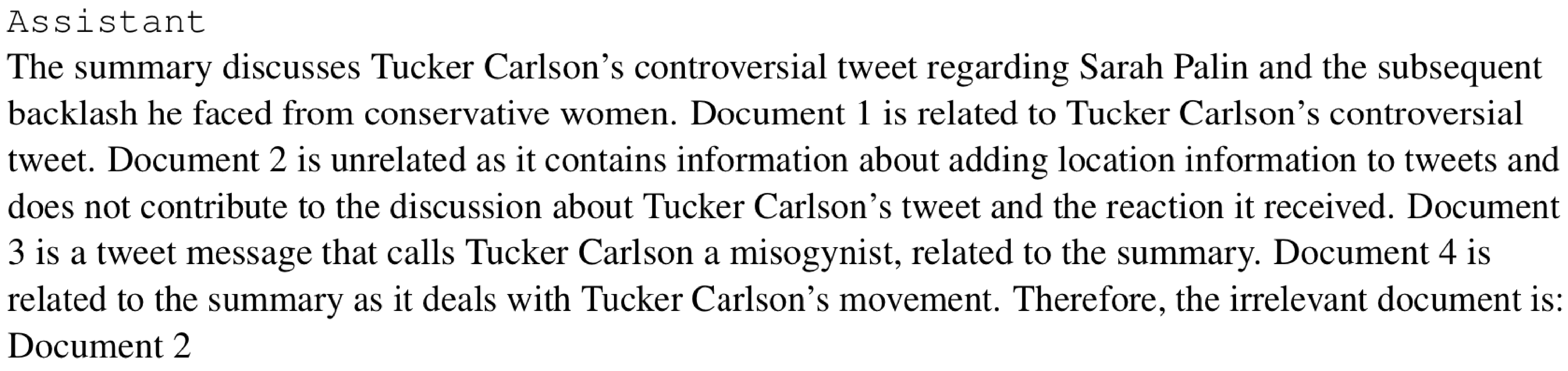
~~**[Document 1]**~~
~~Dow Jones reprints: this copy is for your personal, non-commercial use only …~~

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

Proposed Method: Usage of Chain-of-Thought

We use **chain-of-thought (CoT)**[1] prompting for identifying noisy documents:

- This integration of CoT enables the LLM to consider and evaluate the relevance of each document regarding the summary
- Accordingly, we can expect the improved accuracy of identification
- Furthermore, as this CoT can be used as a rationale for the decision of LLM, it is able to investigate these rationales in case of need for analysis



`Assistant`

The summary discusses Tucker Carlson's controversial tweet regarding Sarah Palin and the subsequent backlash he faced from conservative women. Document 1 is related to Tucker Carlson's controversial tweet. Document 2 is unrelated as it contains information about adding location information to tweets and does not contribute to the discussion about Tucker Carlson's tweet and the reaction it received. Document 3 is a tweet message that calls Tucker Carlson a misogynist, related to the summary. Document 4 is related to the summary as it deals with Tucker Carlson's movement. Therefore, the irrelevant document is: Document 2

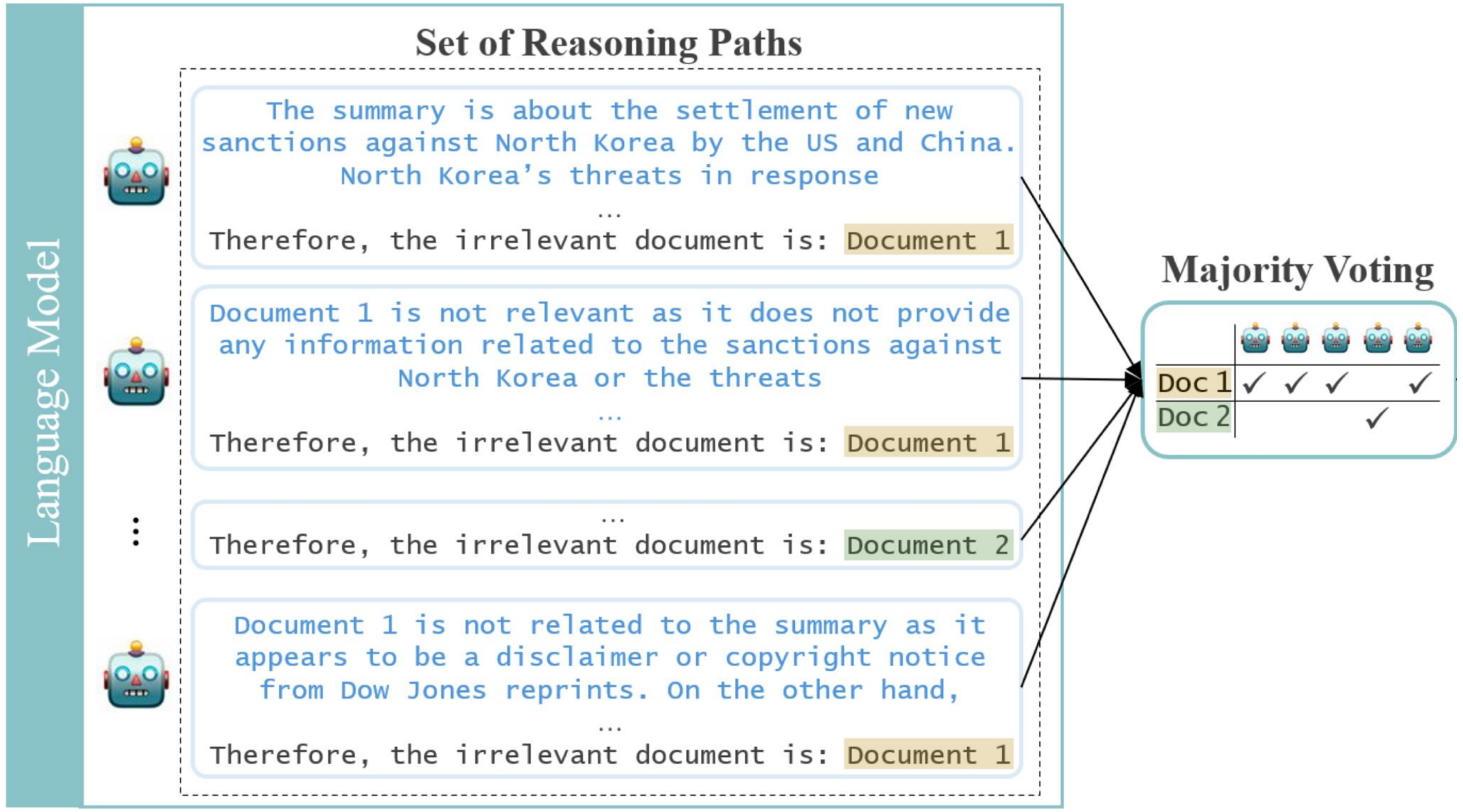
---

[1] Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, NeurIPS 2022.

Proposed Method: Inclusion of Majority Voting

Instead of relying on the decision from a single LLM, we incorporate **majority voting**:

- This is similar to majority voting when using human annotators
  - We used five LLMs and removed documents where more than three agents agreed on the noise
- This majority voting acts as a safeguard against errors from individual LLMs, enhancing the accuracy and quality of the annotation process



Set of Reasoning Paths

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

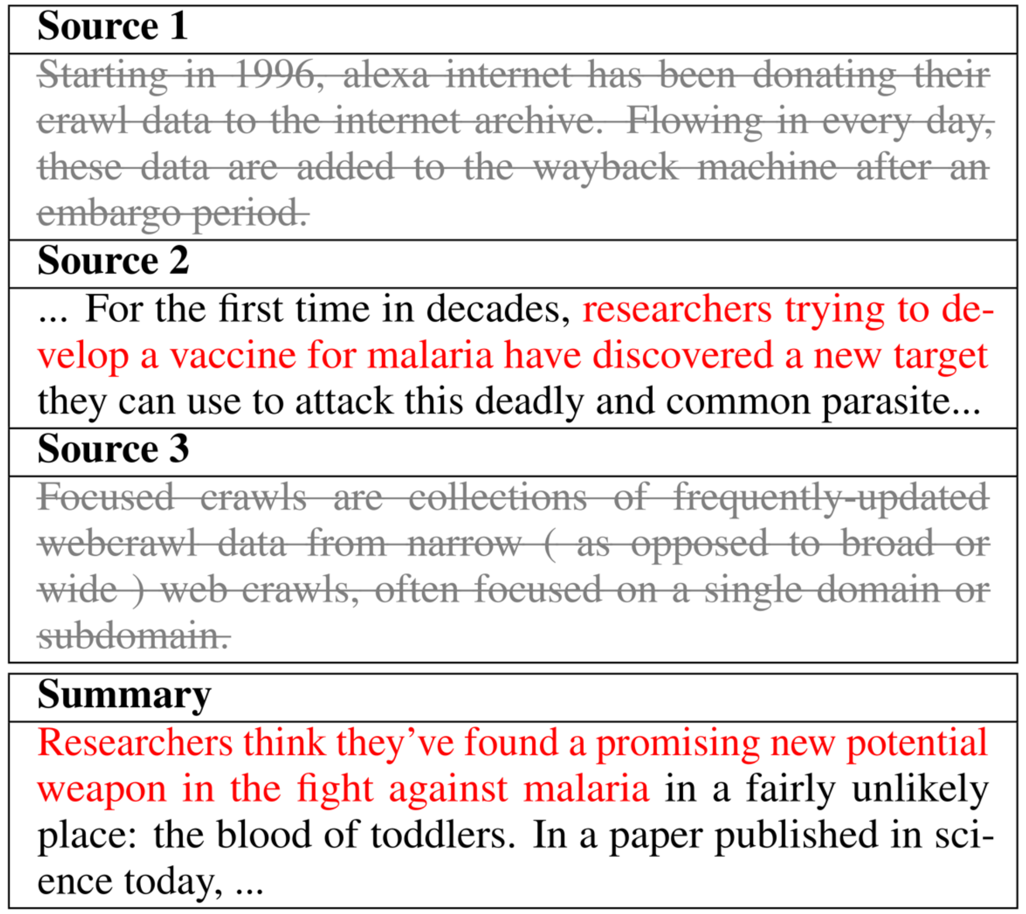
Majority Voting

	🤖	🤖	🤖	🤖	🤖
Doc 1	✓	✓	✓		✓
Doc 2				✓	

Construction of Multi-News+

Based on the proposed framework, we constructed **Multi-News+** from previous Multi-News:

- We removed noisy documents from Multi-News using the proposed framework
- We found that 18% of documents from Multi-News can be regarded as noisy documents
- The overall cost was approximately 550$
    - As we used GPT-3.5-turbo from OpenAI,
the cheaper models or open-source models could further reduce the cost



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

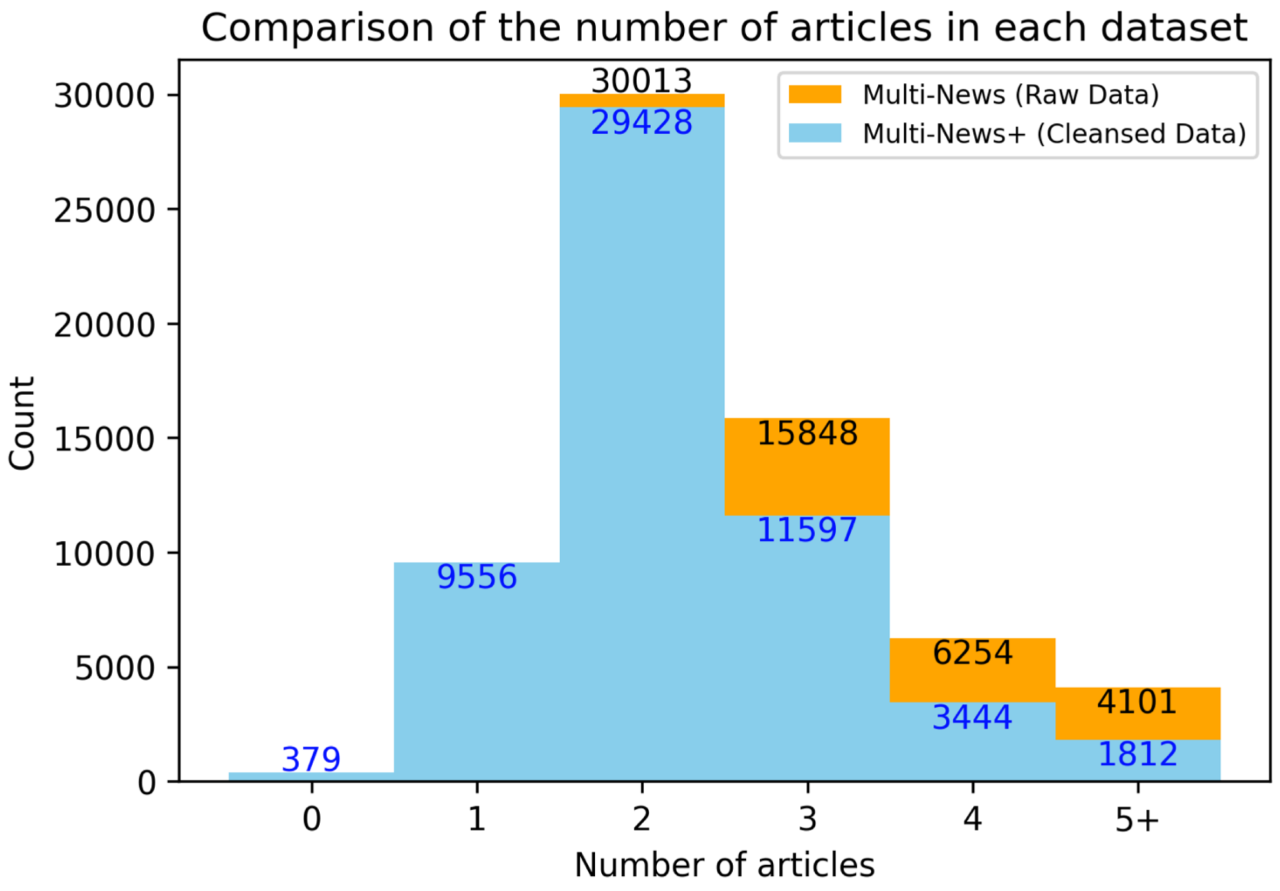
... For the first time in decades, researchers trying to develop a vaccine for malaria have discovered a new target they can use to attack this deadly and common parasite...

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

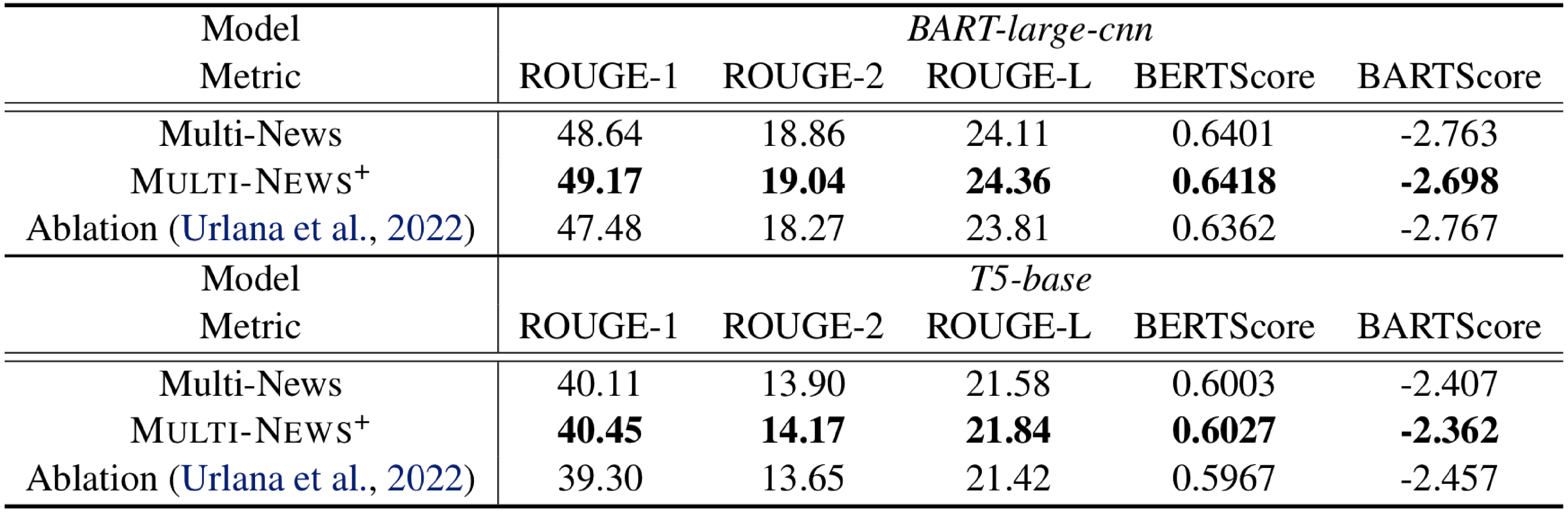
Researchers think they've found a promising new potential weapon in the fight against malaria in a fairly unlikely place: the blood of toddlers. In a paper published in science today, ...



Experiment: Comparison between Two Datasets

We performed experiment to evaluate the effectiveness of using Multi-News+:

- We trained BART and T5 models on Multi-News+ and Multi-News
- We found that the usage of Multi-News+ brings superior performance compared to the model trained with Multi-News
  - The exclusion of noisy documents improved task understanding of the model, leading to better performance
- Additionally, we found that the previous automatic data cleansing method for single document summarization task was not beneficial for Multi-News[1]



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

Conclusion

We proposed:

- A dataset cleansing framework using LLMs with CoT and majority voting, establishing cost-efficient option for identifying irrelevant and noisy data
- Multi-News+, the improved version of Multi-News, using the proposed framework

We found that:

- The usage of Multi-News+ brings performance gain on downstream task
- The proposed framework is useful tool for improving the quality of the dataset

We plan to:

- Extending our approach to other tasks and datasets
- Expanding our method to incorporate the supervision from superior LLMs and weighted majority voting by such expert LLMs

Thank You!