Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method

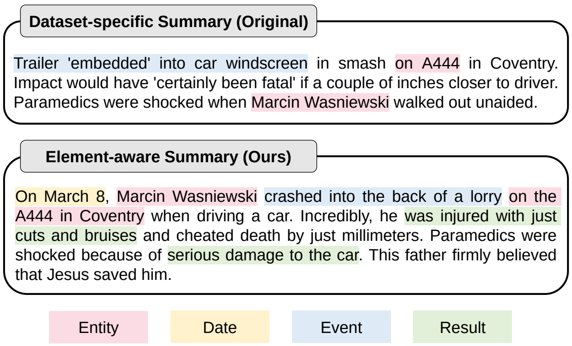
Yiming Wang, Zhuosheng Zhang, Rui Wang[*]
Shanghai Jiao Tong University
`alsaceym@gmail.com, {zhangzs, wangrui12}@sjtu.edu.cn`

Abstract

Automatic summarization generates concise summaries that contain key ideas of source documents. As the most mainstream datasets for the news sub-domain, *CNN/DailyMail* and *BBC XSum* have been widely used for performance benchmarking. However, the reference summaries of those datasets turn out to be noisy, mainly in terms of factual hallucination and information redundancy. To address this challenge, we first annotate new expert-writing **Element-aware** test sets following the "Lasswell Communication Model" proposed by Lasswell (1948), allowing reference summaries to focus on more fine-grained news elements objectively and comprehensively. Utilizing the new test sets, we observe the surprising zero-shot summary ability of LLMs, which addresses the issue of the inconsistent results between human preference and automatic evaluation metrics of LLMs' zero-shot summaries in prior work. Further, we propose a **Summary Chain-of-Thought (SumCoT)** technique to elicit LLMs to generate summaries step by step, which helps them integrate more fine-grained details of source documents into the final summaries that correlate with the human writing mindset. Experimental results show our method outperforms state-of-the-art fine-tuned PLMs and zero-shot LLMs by +4.33/+4.77 in ROUGE-L on the two datasets, respectively. Dataset and code are publicly available at `https://github.com/Alsace08/SumCoT`.

1 Introduction

Automatic summarization is a challenging text generation task that condenses the source text into a few coherent and abstract sentences. In recent years, the study of summarization has evolved with supervised learning based on sequence-to-sequence architectures (Sutskever et al., 2014; Vinyals et al., 2015; Vaswani et al., 2017) and transfer learning based on pre-trained language models (Devlin et al., 2019; Zhang et al., 2019; Liu et al., 2019; Lewis et al., 2020). Existing studies commonly train or fine-tune language models on large-scale corpus (Nallapati et al., 2016; Narayan et al., 2018; Koupaee and Wang, 2018; Fabbri et al., 2019), so superior performance is often reported by measuring the lexical overlap (e.g. ROUGE (Lin, 2004)) with golden summaries (Zhang et al., 2020a; Narayan et al., 2021; Liu et al., 2022b; Narayan et al., 2022), which reflects the fit degree to these standard datasets. However, some standard datasets have shown to be noise-enriched, mainly in terms of information redundancy (Kryscinski et al., 2019) and factual hallucination (Maynez et al., 2020). Meanwhile, sufficient experiments have shown that reference summaries in these standard datasets perform poorly on human assessment dimensions, especially coherence, consistency, and relevance (Stiennon et al., 2020; Fabbri et al., 2021).


Dataset-specific Summary (Original)

Trailer 'embedded' into car windscreen in smash on A444 in Coventry. Impact would have 'certainly been fatal' if a couple of inches closer to driver. Paramedics were shocked when Marcin Wasniewski walked out unaided.

Element-aware Summary (Ours)

On March 8, Marcin Wasniewski crashed into the back of a lorry on the A444 in Coventry when driving a car. Incredibly, he was injured with just cuts and bruises and cheated death by just millimeters. Paramedics were shocked because of serious damage to the car. This father firmly believed that Jesus saved him.

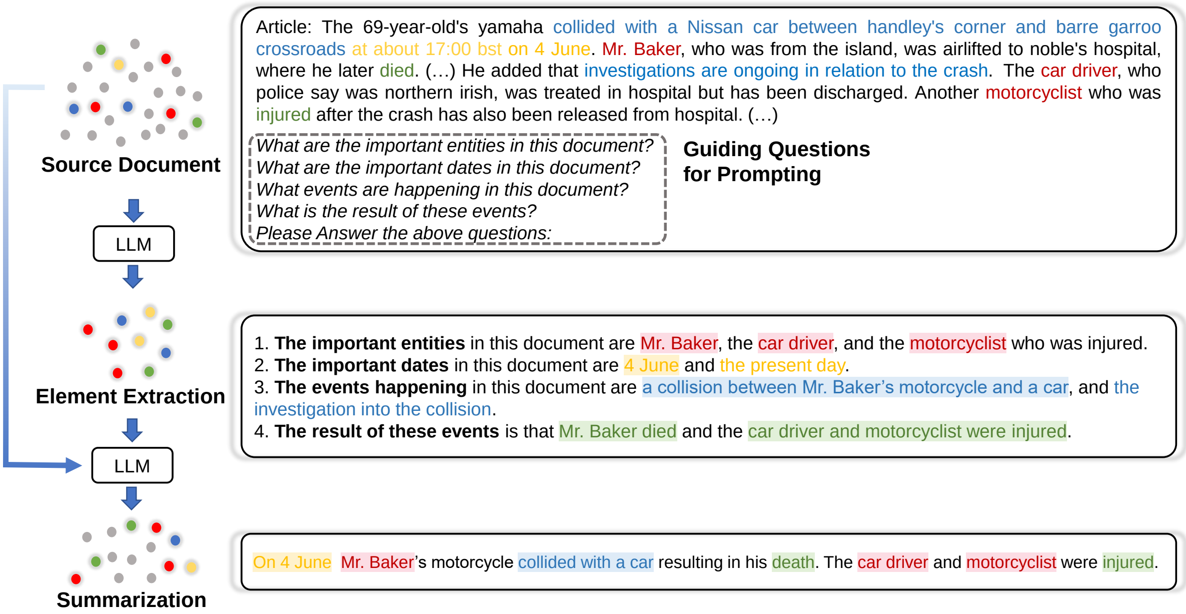
Figure 1: Case comparisons for our **Element-aware summary** and original dataset-specific summary. News elements have been highlighted with different color shadows. It is clear that our element-aware summary covers more comprehensive elements, and the logical connection between the elements is smoother.

To fill this gap, this work releases expert-writing **Element-aware** summary test sets. In professional news writing, core elements such as character, time, place, event, etc., are indispensable. This theory named "Lasswell Communication Model" was first

---

[*] Corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8640–8665
July 9-14, 2023 ©2023 Association for Computational Linguistics



Article: The 69-year-old's yamaha collided with a Nissan car between handley's corner and barre garroo crossroads at about 17:00 bst on 4 June. Mr. Baker, who was from the island, was airlifted to noble's hospital, where he later died. (…) He added that investigations are ongoing in relation to the crash. The car driver, who police say was northern irish, was treated in hospital but has been discharged. Another motorcyclist who was injured after the crash has also been released from hospital. (…)

What are the important entities in this document?
What are the important dates in this document?
What events are happening in this document?
What is the result of these events?
Please Answer the above questions:

1. **The important entities** in this document are Mr. Baker, the car driver, and the motorcyclist who was injured.
2. **The important dates** in this document are 4 June and the present day.
3. **The events happening** in this document are a collision between Mr. Baker's motorcycle and a car, and the investigation into the collision.
4. **The result of these events** is that Mr. Baker died and the car driver and motorcyclist were injured.

On 4 June Mr. Baker's motorcycle collided with a car resulting in his death. The car driver and motorcyclist were injured.

Figure 2: Full pipeline and example of our **Summary Chain-of-Thought** method.

proposed by Lasswell (1948), and later evolved into the "5W1H" paradigm.[1] Following this fine-grained protocol,[2] we ask three news experts to rewrite summaries of source documents from two standard news datasets — *CNN/DailyMail* (Nallapati et al., 2016) and *BBC XSum* (Narayan et al., 2018), allowing reference summaries to contain news core elements objectively and comprehensively[3] (See Figure 1 for one example). Utilizing the new test sets, we are surprised to find that the zero-shot performance of large language models (LLMs) is highly competitive with some strong fine-tuned pre-trained models (PLMs), and the performance of PLMs declines compared to standard test sets. This observation can to some extent address the confusion raised by Goyal et al. (2022) that why GPT-3 generates more human-favored summaries but performs unexpectedly poorly in automatic evaluation metrics — likely due to the limitation of noisy testing domains.

We further build a benchmark for the new test sets. Inspired by the competitive zero-shot performance of LLMs and chain-of-thought technique (Wei et al., 2022b; Kojima et al., 2022), we create **Sum**mary **C**hain-**o**f-**T**hought (**SumCoT**) to elicit LLMs to generate summaries step by step (shown in Figure 2). Concretely, we first guide LLMs to extract the four most core elements for standardized news texts — *Entity*, *Date*, *Event*, *Result* — through some manually-set guiding questions. Immediately after, the guiding questions and corresponding answers output by LLMs are packaged, they further guide LLMs to focus on more critical details to generate summaries that better correlate with the element-aware writing pattern.

Overall, our main contributions are three-fold:

(i) We construct expert-writing element-aware summary test sets to evaluate general summarization systems more objectively (§2).

(ii) We explore the zero-shot summarization ability of LLMs on the new test sets and demonstrate that their writing ability cannot be fully reflected by standard test sets (§3).

(iii) We propose a new CoT-based summarization technique, which allows the LLMs to generate more fine-grained summaries step by step (§4).

2 Element-aware Summary Test Set

2.1 Data Construction

We select two standard news summary datasets (test sets) as document sources, which are representative in terms of length and abstraction: **(i)** ***CNN/DailyMail*** (Nallapati et al., 2016) provides a large-scale multi-domain news collection, which

---

[1] *who*, *where*, *when*, *why*, *what*, and *how*. *who* and *where* can be packaged as *entity*. *why* is usually not independent of *what*, so the two can be packaged as *event*.

[2] Some journalists may follow the Inverted Pyramid style (Pö ttker, 2003), but this protocol is more about a consideration of the full-text layout and is prone to information imbalance within the text (Koupaee and Wang, 2018).

[3] In the era of zero-shot paradigm, LLMs (e.g. GPT-3 (Brown et al., 2020)) have shown decent performance in summarization tasks, so this work focuses on the zero-shot setting to only annotate test sets.

is representative of single-document datasets. We use the standard splits (Hermann et al., 2015) for test sets; **(ii)** ***BBC XSum*** (Narayan et al., 2018) provides a highly abstracted news collection. It has one-sentence summaries and is more abstractive than the *CNN/DailyMail* dataset.

For both datasets, we ask three news experts to independently write professional summaries for 200 randomly sampled source documents according to a complete writing protocol (introduced in §2.2), ensuring comprehensiveness, objectivity, and uniformity of writing style. Different from crowd-sourcing, the involvement of professional writers allows higher inter-annotator agreement. Also, to ensure the uniformity of writing style, we require one of the experts to lead the writing, and the other two to judge the completed summary in four dimensions from the protocol. If there exist inconsistent opinions, they will revise the summary after internal discussion until all pass this annotation. Statistically, the annotation duration of one summary is approximately proportional to the length of source documents. For *CNN/DailyMail*, a summary is written in 25-30 minutes on average, and for *BBC XSum*, in 15-20 minutes on average.

2.2 Writing Protocols

Annotators must follow a comprehensive protocol when writing. Specifically, we divide the protocol into micro demands and macro demands. The former emphasizes our targets, namely element awareness, and the latter guarantees the professionalism and objectivity of the overall writing quality, which alleviates the simple stacking of elements. The two demands complement each other.

Micro Demands. All news summaries should have four essential core elements — ***Entity***, ***Date***, ***Event***, and ***Result*** — following the "Lasswell Communication Model" (Lasswell, 1948), and these elements must be faithful to the source document. For example, when there is no date in the source document, writers can not add dates to the final summary by force.

Macro Demands. All news summaries should focus on the four dimensions (Gehrmann et al., 2018; Kryscinski et al., 2019). **(i) Fluency**: No spelling, grammatical, or syntactic errors within sentences; **(ii) Coherence**: The summary should not be a heap of events, and linguistic transition must be smooth and logically correct; **(iii) Consistency**: No hallucinated facts — neither facts that do not appear in or are contrary to the source document are allowed; **(iv) Relevance**: Adequately weigh the importance of multiple facts, and find the core concern of the text. Non-core facts can be reduced in length, and redundant details are not allowed.

Reference Summary	*CNN/DaliyMail*	
	% of novel **uni/bi/trigram**	Avg. summary length of **words/sentences**
Dataset-specific	17.00/53.91/71.98	50.14/3.59
Element-aware	20.31/49.72/62.14	51.08/2.71
Reference Summary	***BBC XSum***	
	% of novel **uni/bi/trigram**	Avg. summary length of **words/sentences**
Dataset-specific	39.39/87.86/96.95	22.18/1.00
Element-aware	36.28/70.56/82.36	23.33/1.00

Table 1: Some statistics of element-aware summaries compared with original dataset-specific summaries. *Novel n-grams* indicates the n-grams that are included in the summary but not in the source document.

2.3 Overall Quality

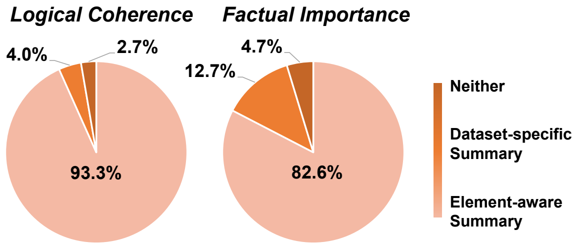
We first compare the overall quality of our test sets with the original data. Table 1 quantifies some statistics of the element-aware summaries compared with original dataset-specific summaries. The average length of element-aware summaries largely matches the distribution of that of dataset-specific summaries. In terms of abstraction, we report the percentage of novel n-grams that are included in the summary but not in the source document. We note that the percent of novel n-grams in element-aware summaries is lower than that of dataset-specific summaries but with a reasonable gap, which reflects that expert-writing element-aware summaries would be more faithful to the source documents but not heavily replicate them.[4]

We further hold a vote on two highly subjective dimensions — logical coherence and factual importance, they reflect the professionalism and the information comprehensiveness of writing.[5] We ask three annotators to perform preference selection on 50 randomly selected instances from both datasets — for each instance, they can select **at most one** summary that performs better in the two

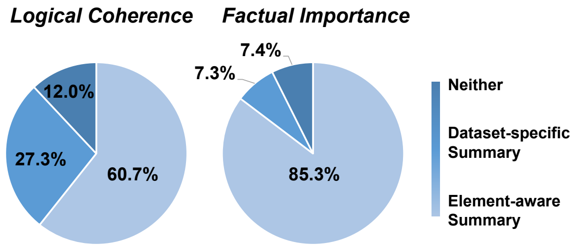
---

[4] Additionally, factual errors in the dataset-specific summaries will result in a spuriously high abstraction to some degree. In contrast, element-aware summaries better trade-off abstraction and faithfulness (See Appendix B for examples).

[5] Whether the transition between facts is coherent, and whether important facts in the source documents are comprehensive and non-redundant in the summaries.



(a) *CNN/DailyMail*



(b) *BBC XSum*

Figure 3: Average annotator vote distribution for better summaries between dataset-specific and element-aware summaries on “logical coherence” and “factual importance” dimensions. It is clear that element-aware summaries are more accepted by the public.

dimensions, respectively, or none if they consider both to be not good.

Figure 3 shows the vote results. It is clear that element-aware summaries are significantly more popularly accepted in both subjective dimensions by the public, demonstrating that our summaries are more human-favored.

2.4 Element-aware Characteristic

In this part, we will demonstrate that our annotated summaries have more obvious element-aware characteristic than the dataset-specific summaries.

We ask three annotators to evaluate every document-summary pair. For each sample, and for i-th annotator ($i = 1, 2, 3$) and j-th element in the writing protocol ($j = 1, 2, 3, 4$), we ask this annotator to release two sets that separately contain all j-th elements in the source document they consider important and all j-th elements appearing in the summary. The annotator-released sets for the source document and summary are denoted as A_i^j and $A_i'^j$, respectively.

Then, we compute the Precision and Recall, they separately reflect the accuracy of the core elements embedded in the summary and the hit rate of the core elements in the source document. Precision^j and Recall^j are formulated as:[6]

Core Element	Element-aware			Dataset-specific		
	P	***R***	***F1***	***P***	***R***	***F1***
			CNN/DailyMail			
Entity	0.98	0.96	**0.98**	0.75	0.63	**0.68**
Date	0.89	0.91	**0.90**	0.74	0.65	**0.69**
Event	0.96	0.95	**0.95**	0.66	0.55	**0.60**
Result	0.95	0.95	**0.95**	0.49	0.42	**0.45**
			BBC XSum			
Entity	0.97	0.87	**0.92**	0.76	0.54	**0.63**
Date	0.97	0.95	**0.96**	0.52	0.45	**0.48**
Event	0.93	0.93	**0.93**	0.80	0.48	**0.60**
Result	0.96	0.98	**0.97**	0.23	0.18	**0.20**

Table 2: The comparison between element-aware and dataset-specific test sets over Precision (*P*), Recall (*R*), and F_1 score of all four elements.

$$\begin{aligned}\text{Precision}^j &= \frac{1}{3}\sum_{i=1}^{3}\frac{|A_i^j \bigcap A_i'^j|}{|A_i'^j|}, \quad j = 1,2,3,4 \\ \text{Recall}^j &= \frac{1}{3}\sum_{i=1}^{3}\frac{|A_i^j \bigcap A_i'^j|}{|A_i^j|}, \quad j = 1,2,3,4\end{aligned} \tag{1}$$

where $|\cdot|$ denotes the number of elements in the set. For ***Event*** and ***Result***, a complete lexical overlap is unrealistic due to the subjectivity in expression, so as long as the same meaning is considered correct.

We compare the Precision and Recall between element-aware and dataset-specific test sets, and computer the average of all document-summary pairs of a test set. We also compute F_1 score (The harmonic mean of Precision and Recall) to measure the overall level. Results are shown in Table 2, the comparison shows that our test sets have a significant advantage in the element-aware characteristic. The dataset-specific test sets perform poorly particularly in the Recall score, meaning that they have ignored many fine-grained details.

3 Preliminary Comparison: Zero-shot LLMs Versus Fine-tuned PLMs

In this section, we preliminarily compare existing strong LLMs and PLMs upon our element-aware test sets, designed to analyze the general summary capabilities of zero-shot LLMs and fine-tuned PLMs from a more fine-grained perspective.

---

[6] In extreme situations, when A_i^j is empty, i.e., the annotator thinks that there is no j-th element in the source document, the Recall^j is 1 if this element is also not covered in the summary, otherwise 0. Ditto for Precision^j when $A_i'^j$ is empty.

CNN/DaliyMail

Ref / *Model*	Element-aware (ours)				Dataset-specific (original)			
	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE
BART-BASE	36.06	15.93	33.09	0.8762	38.55	17.57	36.05	0.8779
BART-LARGE	**37.98**	**18.16**	34.30	0.8860	39.01	18.26	**37.15**	**0.8868**
T5-LARGE	37.47	17.66	**34.34**	0.8768	38.84	**18.39**	37.01	0.8802
PEGASUS-LARGE	36.65	17.58	33.84	0.8710	**39.11**	17.82	36.86	0.8798
175B GPT-3	37.75	15.20	34.25	**0.8905**	30.10	8.98	27.51	0.8718

BBC XSum

Ref / *Model*	Element-aware (ours)				Dataset-specific (original)			
	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE
BART-BASE	21.89	5.13	17.19	0.8663	29.67	10.09	24.46	0.8779
BART-LARGE	23.79	5.02	17.93	0.8710	33.95	11.29	26.78	0.8880
T5-LARGE	24.98	6.89	19.46	0.8728	30.79	9.61	24.73	0.8792
PEGASUS-LARGE	21.35	4.87	17.03	0.8662	**35.16**	**13.21**	**29.30**	**0.8888**
175B GPT-3	**31.74**	**10.95**	**25.42**	**0.8933**	19.99	3.69	15.86	0.8654

Table 3: Performance comparison of zero-shot LLMs (175B GPT-3) and fine-tuned PLMs (BART, T5, and PEGASU). We separately compare generated summaries of these models with original reference summaries from standard datasets (Dataset-specific) and our reference summaries rewritten by news experts (Element-aware). Results are evaluated automatically over ROUGE-1/2/L and BERTSCORE.

3.1 Experimental Setup

Dataset. We perform experiments on two mainstream news datasets *CNN/DailyMail* and *BBC XSum* introduced in §2.1. For each source document on both datasets, we compare summaries generated by models with dataset-specific (original) and element-aware (ours) reference summaries. Each test set includes 200 document-summary pairs consistent with the annotation number.

Models. For LLMs, We use 175B-parameter GPT-3 (`text-davinci-002` version) (Brown et al., 2020; Ouyang et al., 2022) for our study. For PLMs, we select BART (Lewis et al., 2020), T5 (Raffel et al., 2020) — two strong generation-oriented PLMs, and PEGASUS (Zhang et al., 2020a) — a summarization-customized PLM fine-tuned on two datasets separately as the strong baselines.

Implementation. We follow the official fine-tuned models released on the Huggingface for PLMs generation. For zero-shot prompts of LLMs, We follow Sanh et al. (2022) and Goyal et al. (2022) to set [p] = "*Summarize the above article:*" as the standard prompt on *CNN/DailyMail*. On *BBC XSum*, considering its one-sentence summary style with extreme generalization, we use sentence-constraint prompt [p] = "*Summarize the above article in one sentence:*". All the source documents are truncated to 1024 tokens when using PLMs and 2048 tokens when using LLMs. See Appendix A for more useful implementation details.

Evaluation. We evaluate the generated summaries using lexical-overlap metrics, specifically ROUGE-1/2/L (Lin, 2004), and embedding-similarity metrics, specifically BERTSCORE (Zhang et al., 2020b). Besides, we resort to more precise human studies to evaluate the consistency of generated summaries and source documents. See Appendix A for more useful evaluation details.

3.2 Main Results

Longitudinal Comparison: Language Models. First, we compare the performance of different models on the same test set (see columns of Table 3). On dataset-specific test sets (the right part), the relative performances among PLMs are basically in line with the experimental results in Zhang et al. (2020a), meaning that our sampled source documents basically follow the distribution of original test sets. On element-aware test sets (the left part), surprisingly, zero-shot GPT-3 performs competitively with all other fine-tuned PLMs and even outperforms other models with a wide margin on *BBC XSum*. These all present that LLMs have more fine-grained summary capabilities, and their zero-shot evaluation is limited by the original test sets.

Horizontal Comparison: Test Sets. Next, we compare the performances of the same model on different test sets (see rows of Table 3). We note that these fine-tuned PLMs perform worse on element-aware test sets than they do on dataset-specific test sets, with a particularly salient drop on

BBC XSum. In contrast, GPT-3 obtains dramatic improvements on element-aware test sets. Compared with the performances on dataset-specific test sets, ROUGE-1/2/L increases by +7.65/+6.22/+6.74 points on *CNN/DailyMail* and +11.75/+7.26/+9.56 points on *BBC XSum*. These contrasting results demonstrate that our annotated test sets pose a challenge for PLMs fine-tuned with standard datasets, but LLMs can perform well due to their more fine-grained writing capabilities.

3.3 Human Study

Human studies are conducted as an overall quality assessment of human preferences. We use a 7-point Likert scale (Likert, 1932) to ask annotators to evaluate four dimensions: Fluency, Coherence, Consistency, and Relevance (equivalent to macro demands in §2.2). Different from baseline-free human studies, we set the element-aware summaries as the baseline (score 0) and set the scoring range to -3~3. A more positive score means higher quality than the element-aware summary and vice versa. For each sample, we present the dataset-specific (original), BART-LARGE, T5-LARGE, PEGASU-LARGE and 175B GPT-3 summaries to the annotators and ask them to score one by one.

As is shown in Figure 4, GPT-3 summaries outperform almost all other dataset-specific or model-generated summaries in each dimension, although not yet achieved the level of element-aware summaries. All of these results can fully demonstrate that LLMs have great potential for summarization, and a higher-quality dataset is key for evaluation.

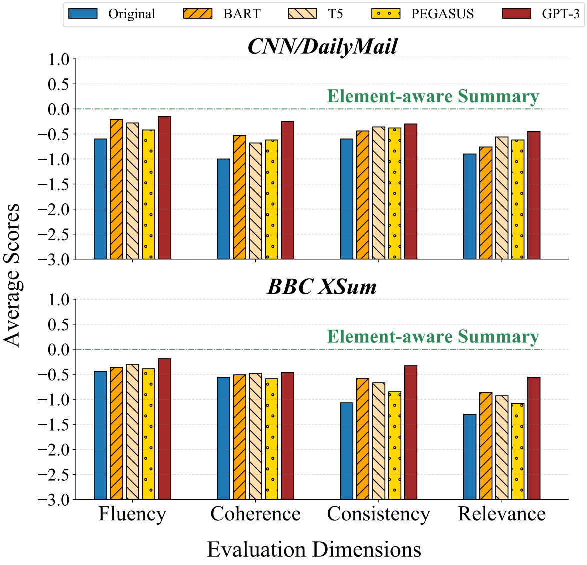


Figure 4: Human evaluation scores of four dimensions about summary quality on the 50-shot *CNN/DailyMail* (the upper part) and *BBC XSum* (the lower part) datasets. More human study details are shown in Appendix A.2.

4 Towards Element-oriented Summary: Chain-of-Thought Method

We have analyzed the summary writing ability of zero-shot GPT-3 and other fine-tuned PLMs in §3. We see that GPT-3 performs surprisingly well on our element-aware test sets. The results compellingly show that GPT-3 has great potential for fine-grained zero-shot summary writing. Inspired by the prevalence of the chain-of-thought (CoT) technique in LLMs (Wei et al., 2022b; Kojima et al., 2022; Zhang et al., 2022), we can further enhance the summarization ability of LLMs by leveraging a CoT-based method (SumCoT). SumCoT elicits LLMs to focus on news core elements, thereby generating element-aware summaries step by step. The pipeline and example have been illustrated in Figure 2.

4.1 Two-stage Pipeline

We first ask the LLMs to extract core news elements in the source document by manually-set guiding questions, and later integrate the information based on the extracted elements and more details from the source documents. Pipeline details are as follows.

- **Stage 1: Core element extraction.** In the first stage, we create guiding-question prompts to elicit the LLMs to extract four core elements: ***Entity***, ***Date***, ***Event***, ***Result***. For the i-th element, we set a simple question q_i to guide the model for extracting it (shown in Figure 2), and then concatenate these questions into $\boldsymbol{Q} = [q_1, q_2, q_3, q_4]$. Let the source document be S, then the LLMs input in this stage is formulated as $[\boldsymbol{S}; \boldsymbol{Q}]$.
- **Stage 2: Multiple information integration and summarization.** We obtain an extraction answer $\boldsymbol{A}$ from the LLMs in Stage 1. Next, we integrate the extracted elements and more detailed information from the source document. We concatenate the source document, questions, answer, and a simple prompt [p′]="*Let's integrate the above information and summarize the article:*" to prompt the LLMs for summary generation.[7] The input in this stage is formulated as $[\boldsymbol{S}; \boldsymbol{Q}; \boldsymbol{A}; [\text{p}']]$, and the output is the final summary.

---

[7] Similarly, for XSum, [p′]="*Let's integrate the above information and summarize the article in one sentence:*"

GPT-3 Zero-shot Summary (Standard)
Summarize the above article: The article discusses how Rod Blagojevich, a former Illinois governor, has let his hair go while serving his prison sentence. The photos show the former governor with white hair, rather than the black hair that was his trademark as a politician. Blagojevich is currently serving a 14-year sentence at the Federal Correctional Institution Englewood near Denver.
GPT-3 Zero-shot-cot Summary (Using SumCoT)
Let's integrate the above information and summarize the article: In 2008, Rod Blagojevich was arrested for trying to sell Barack Obama's U.S. Senate seat. He was impeached and removed from office in 2009, and sentenced to 14 years in prison. He is currently serving his sentence at the Federal Correctional Institution Englewood near Denver. In 2014, his lawyers attempted to appeal his conviction, arguing that a recent Supreme Court ruling supported their view that Blagojevich was engaged in legal, run-of-the-mill political horse trading, not corruption.

Table 4: Case comparisons between GPT-3 zero-shot summaries before and after using SumCoT. Spans of **Entity**, **Date**, **Event** and **Result** are separately highlighted in red, yellow, blue and green. **Prompts** are presented in *italics*.

Model	*CNN/DaliyMail*			
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**BERTSCORE**
Previous SOTA in Table 3	37.98	18.16	34.34	0.8905
175B GPT-3	37.75	15.20	34.25	0.8905
175B GPT-3 w/ SumCoT	**43.03** (↑ **5.05**)	**19.51** (↑ **1.35**)	**38.67** (↑ **4.33**)	**0.9023** (↑ **0.0118**)
Model	*BBC XSum*			
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**BERTSCORE**
Previous SOTA in Table 3	31.74	10.95	25.42	0.8933
175B GPT-3	31.74	10.95	25.42	0.8933
175B GPT-3 w/ SumCoT	**35.70** (↑ **3.96**)	**15.31** (↑ **4.36**)	**30.19** (↑ **4.77**)	**0.9018** (↑ **0.0085**)

Table 5: Performance comparisons upon element-aware test sets of our method (GPT-3 with SumCoT), standard GPT-3, and previous state-of-the-art (SOTA) results in Table 3 over each metric. The ↑ and corresponding numbers on the right of each result of our method represent the increase after comparing with the previous SOTA.

Model	***CNN/DailyMail*** *Flu/Coh/Con/Rel*	***BBC XSum*** *Flu/Coh/Con/Rel*
175B GPT-3	-0.18/-0.33/-0.37/-0.72	-0.19/-0.48/-0.33/-0.56
w/ SumCoT	**-0.10/-0.05/-0.23/-0.28**	**-0.11/-0.19/-0.07/-0.22**

Table 6: Human evaluation scores (Scale -3∼3, and 0 represents the level of element-aware summaries) for zero-shot summaries of GPT-3 w/o and w/ SumCoT. *Flu/Coh/Con/Rel* stands for Fluency/Coherence/Consistency/Relevance respectively.

4.2 Comprehensive Evaluation

First, we visually compare the quality of summaries generated by GPT-3 before and after using SumCoT. As shown in Table 4, it is clear that the summary generated under SumCoT contains more abundant fine-grained elements, saturating the summary text with more key information.

Next, we perform quantitative evaluations over the same metrics as in §3.1. We mainly compare our method (GPT-3 with SumCoT), standard GPT-3, and previous state-of-the-art (SOTA) results in Table 3, and updated results are shown in Table 5. Compared with the standard GPT-3 and previous SOTA, GPT-3 with SumCoT obtains salient improvement in all metrics when compared with the element-aware summaries, where ROUGE-1/2/L increases by +5.05/+1.35/+4.33 points on *CNN/DailyMail* and +3.96/+4.36/+4.77 points on *BBC XSum*, demonstrating that GPT-3 successfully focuses on more core elements through SumCoT and further fits the element-aware writing pattern.

Finally, we also conduct human studies to compare summaries of GPT-3 w/o and w/ SumCoT. Results (as shown in Table 6) indicate that the SumCoT technique further improves the performance of the standard zero-shot paradigm in all dimensions, particularly coherence and relevance.

4.3 Better Understanding SumCoT

How does SumCoT affect summary writing? First, we explore the extent to which SumCoT affects the final summary generation. We compute the coverage, the fraction of extracted elements in Stage 1 actually appearing in the final summary generated in Stage 2. Table 7 shows the results (see Appendix C.1 for examples), and we observe

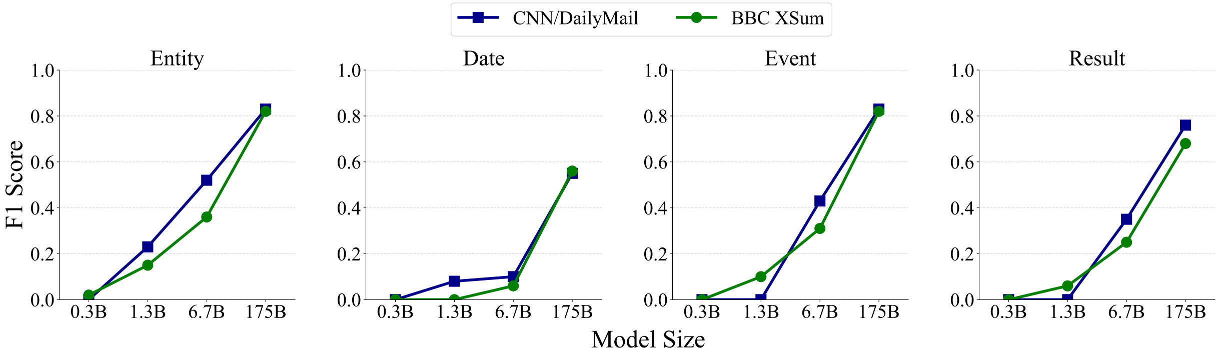


Figure 5: Performance of element extraction for all four core elements with various GPT-3 versions. See Appendix C.3 for more model details.

CNN/DailyMail				*BBC XSum*			
Entity	Date	Event	Result	Entity	Date	Event	Result
0.89	0.55	0.93	0.95	0.80	0.48	0.87	0.66

Table 7: Coverage, the fraction of extracted elements actually appearing in the final summary on two datasets.

Core Element	***CNN/DaliyMail***			***BBC XSum***		
	P	*R*	***F1***	*P*	*R*	***F1***
Entity	0.77	0.89	**0.83**	0.71	0.98	**0.82**
Date	0.46	0.68	**0.55**	0.43	0.79	**0.56**
Event	0.84	0.82	**0.83**	0.75	0.90	**0.82**
Result	0.74	0.79	**0.76**	0.66	0.71	**0.68**

Table 8: The Precision (P), Recall (R), and F_1 for extraction of each element.

that final summaries are extremely faithful to the extracted elements, particularly on *CNN/DailyMail*. On *BBC XSum*, the coverages of each element are relatively lower due to the one-sentence style of *BBC XSum*, resulting in further condensation of the extracted elements. In addition, the coverage of ***Date*** is significantly low, probably due to the errors of extraction. This will be verified in the next part.

Is the element extraction accurate and comprehensive? Table 7 demonstrates a strong correlation between element extraction and summary generation, so we need to examine the quality of element extraction.[8] We compute the Precision, Recall and F_1 introduced in §2.4. Results (Table 8) show that extraction achieves an outperforming result except for ***Date***, and Precision are usually lower than Recall. See Appendix C.2 for error cases, where we conclude: (i) Date hallucination is particularly evident for extracting non-existent dates; (ii) Element redundancy often occurs.

Does the model size limit SumCoT? We compare the performance of GPT-3 with different versions of element extraction. We compute the F_1 score (shown in Figure 5) for all the elements. We find that when the model size is small, element extraction is almost invalid. As the model size increases, GPT-3 can extract one by one for all types of elements, but the extraction itself has many errors or redundancies. Only when the model size is the largest, the element extraction is human-approved (See Appendix C.3 for examples). This indicates that the SumCoT technique is also an emergent ability of model scale (Wei et al., 2022a), and is effective only when the model size is larger.

5 Related Work and Discussion

5.1 Summarization: Dataset and Evaluation

In the data-driven deep learning era, large-scale corpus crawled from websites for summarization is rich, especially the news domain. They can be divided into the single-document setting (Harman and Over, 2004; Sandhaus, 2008; Napoles et al., 2012; Nallapati et al., 2016; Narayan et al., 2018; Koupaee and Wang, 2018; Grusky et al., 2018) and the multi-document setting (Owczarzak and Dang, 2011; Li et al., 2017; Fabbri et al., 2019) according to the source numbers of document clusters. However, some studies pointed out various noises within them, such as poor coherence, information redundancy, and factual hallucination (Kryscinski et al., 2019; Maynez et al., 2020; Fabbri et al., 2021). Several other studies also corroborated this with human assessments (Stiennon et al., 2020; Fabbri et al., 2021).

Summarization systems are first purely trained (Vinyals et al., 2015; Vaswani et al., 2017; Liu et al., 2022b; Chen et al., 2022) or fine-tuned (Zhang et al., 2019; Liu, 2019; Zhang et al., 2020a; Raffel

---

[8] It is noted that if there are no obvious markers (e.g. "*The entities are* ..."), the extraction is not considered valid.

et al., 2020; Wang et al., 2022b; Mao et al., 2022) with standard datasets, and then evaluated. The most mainstream automatic evaluation metrics for summarization are reference-based methods, i.e., directly comparing the similarity of generated and dataset-specific summaries. They can be split into lexical overlap methods (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005) and semantic similarity methods (Ng and Abrecht, 2015; Zhang et al., 2020b; Zhao et al., 2019; Sellam et al., 2020; Rei et al., 2020). Such evaluation is essentially a test of the fit degree to standard datasets. In recent years, the advanced zero-shot paradigm of LLMs makes text generation free of standard datasets (Brown et al., 2020; Chowdhery et al., 2022; Thoppilan et al., 2022) but rely on massive pre-trained data, many researchers tend to revisit the quality assessment of summaries generated by LLMs (Liu et al., 2022a; Zhang et al., 2023a). However, some studies demonstrate that automatic evaluation results do not align with human preference in summarization tasks (Goyal et al., 2022), similar counter-intuitive observations may pose new challenges for the evaluation in the era of LLMs.

5.2 Chain-of-Thought Prompting for LLMs

Recently, intriguing chain-of-thought techniques have greatly improved both the reasoning performance and interpretability of LLMs by decomposing multi-step problems into intermediate steps (Nye et al., 2022; Wei et al., 2022b; Kojima et al., 2022; Zhang et al., 2022; Wang et al., 2022a; Zhang et al., 2023b; Shi et al., 2022; Zhou et al., 2022). However, no prior work has studied CoT in the scenario of automatic summarization. To the best of our knowledge, we are the first to study chain-of-thought prompting for summarization, eliciting LLMs to leverage more fine-grained elements from source documents to generate effective summaries.

6 Conclusion

In this work, we construct expert-writing element-aware summary test sets for *CNN/DailyMail* and *BBC XSum*, they are specifically designed to assess the generic summarization capabilities of diverse, powerful language models more thoroughly. Upon the fine-grained test sets, we preliminarily conduct experiments on zero-shot LLMs and fine-tuned PLMs, demonstrating the surprising zero-shot summary writing ability of LLMs. Further, we propose a CoT-based method, which elicits LLMs to focus on core news elements and generate summaries step by step. In the future, we hope that our work will inspire further research into harnessing LLMs' potential to mimic human writing processes across various open-ended generative tasks.

Limitations

In terms of the test sets, due to time, labor, and financial limitations, we are unable to construct large-scale test sets of the same size as the original, so the domain balance in the test sets is not fully considered, but the uniformity of writing style might have slightly alleviated this issue. In terms of the method, we empirically explore the possibility of chain-of-thought application in text generation. However, due to the stronger openness of generative tasks compared to pure reasoning tasks, generated summaries might be more sensitive to the form of chain-of-thought, which is a key point worth further optimization.

Ethics Statement

We use publicly available source documents from existing general datasets for annotations, so the ethics issues of the source texts are non-existent. For the generated contents with LLMs, e.g. GPT-3, prior work (Brown et al., 2020; Chan, 2022) has elaborated on their inevitable potential toxicity, such as issues of bias and fairness. Moreover, this is the first work to apply the chain-of-thought technique to open-end generation tasks, so we completely keep the prompts neutral and task-specific to avoid toxic language generation, and there were no toxic texts that appeared in our experiments.

Acknowledgements

Yiming and Rui are with MT-Lab, Department of Computer Science and Engineering, School of Electronic Information and Electrical Engineering, and also with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200204, China. Rui is supported by the General Program of National Natural Science Foundation of China (6217020129), Shanghai Pujiang Program (21PJ1406800), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Beijing Academy of Artificial Intelligence (BAAI) (No. 4), CCF-Baidu Open Fund (F2022018), and the Alibaba-AIR Program (22088682). We also thank the computational resource from the SJTU student innovation center.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Anastasia Chan. 2022. Gpt-3 and instructgpt: technological dystopianism, utopianism, and “contextual” perspectives in ai ethics and industry. *AI and Ethics*, pages 1–12.

Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. Unisumm: Unified few-shot summarization with multi-task pre-training and prefix-tuning. *ArXiv preprint*, abs/2211.09783.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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.

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

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Donna Harman and Paul Over. 2004. The effects of human variation in DUC summarization evaluation. In *Text Summarization Branches Out*, pages 10–17, Barcelona, Spain. Association for Computational Linguistics.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *ArXiv preprint*, abs/2012.04281.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv preprint*, abs/2205.11916.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *ArXiv preprint*, abs/1810.09305.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Harold D Lasswell. 1948. The structure and function of communication in society. *The communication of ideas*, 37(1):136–139.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Piji Li, Lidong Bing, and Wai Lam. 2017. Reader-aware multi-document summarization: An enhanced model and the first dataset. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 91–99, Copenhagen, Denmark. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

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.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *ArXiv preprint*, abs/1903.10318.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, et al. 2022a. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *arXiv e-prints*, pages arXiv–2212.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022b. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Qianren Mao, Jianxin Li, JiaZheng Wang, Xi Li, Peng Hao, Lihong Wang, and Zheng Wang. 2022. Explicitly modeling importance and coherence for timeline summarization. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8062–8066. IEEE.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. 2022. A well-composed text is half done! composition sampling for diverse conditional generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1319–1339, Dublin, Ireland. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2022. Show your work: Scratchpads for intermediate computation with language models. In *Deep Learning for Code Workshop*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155.

Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the tac 2011 summarization track: Guided task and aesop task. In *Proceedings of the Text Analysis Conference (TAC 2011), Gaithersburg, Maryland, USA, November*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Horst Pö ttker. 2003. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus. In *Philadelphia: Linguistic Data Consortium, 2008*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *ArXiv preprint*, abs/2210.03057.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2692–2700.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022a. Self-consistency improves chain of thought reasoning in language models. *ArXiv preprint*, abs/2203.11171.

Yiming Wang, Qianren Mao, Junnan Liu, Weifeng Jiang, Hongdong Zhu, and Jianxin Li. 2022b. Noise-injected consistency training and entropy-constrained pseudo labeling for semi-supervised extractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6447–6456, Gyeongju, Republic of Korea. International Committee on 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, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *ArXiv preprint*, abs/2201.11903.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. 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. 2023a. Benchmarking large language models for news summarization. *ArXiv preprint*, abs/2301.13848.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *ArXiv preprint*, abs/2210.03493.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *ArXiv preprint, abs/2302.00923*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *ArXiv preprint*, abs/2205.10625.

A Details of Experimental Setup

A.1 Main Experiment

Table 9 report the sources and licenses of artifacts and packages we used in this paper.

A.2 Human Study

We randomly select 50 samples for each dataset and ask three annotators for these tasks following the setting of most human studies. However, considering the unprofessionalism of crowd-sourcing evaluations (Usually hiring workers from Amazon Mechanical Turk platform with a set hourly salary. Actually, many workers will not work as you expected, their levels vary widely and uncontrollably. He et al. (2020) have encountered such a situation), we privately contact three reliable annotators to conduct the human studies. The first is a Ph.D. candidate in Computer Science, the second is a Master in Film Study, and the last is a graduate in Journalism and Communication. Our human studies are conducted in full compliance with the willingness of the invitees and are fully open about the use of the data they annotated. They have been paid slightly more than the crowd-sourced hourly rate for their work. We use the same configuration for all human studies in this paper, thanks for their participation!

B Abstraction and Faithfulness analysis for Summaries

Abstraction and Faithfulness are normally two opposing properties. For dataset-specific summaries, despite their novel n-grams being higher than element-aware summaries in many cases, they sacrifice factual correctness to some extent, which is a fake high-abstraction. Case comparisons are shown in Table 10-11.

C Better Understanding Summary Chain-of-Thought: Case Study

C.1 Learn How SumCoT Works

Table 12-14 presents some cases to visually show how SumCoT affects the final generated summary. We compare GPT-3 zero-shot summaries before and after using SumCoT. Core elements have been highlighted in the table. It is clear that summaries using SumCoT Cover a large number of fine-grained elements extracted by GPT-3 in Stage 1 that are not in the standard zero-shot summaries.

Model	URL	License
BART (Lewis et al., 2020)	https://huggingface.co/ainize/bart-base-cnn https://huggingface.co/morenolq/bart-base-xsum https://huggingface.co/facebook/bart-large-cnn https://huggingface.co/facebook/bart-large-xsum	MIT license
PEGASUS (Zhang et al., 2020a)	https://huggingface.co/google/pegasus-cnn_dailymail https://huggingface.co/google/pegasus-xsum	Apache-2.0 license
GPT-3 (Zhang et al., 2020a)	https://openai.com/api/	N/A
Evaluation Metric	URL	License
ROUGE (Lin, 2004)	https://github.com/pltrdy/rouge	Apache-2.0 license
BERTSCORE (Zhang et al., 2020b)	https://github.com/Tiiiger/bert_score	MIT license

Table 9: The sources and licenses of artifacts and packages we used in this paper (Appendix A.1).

C.2 Error Analysis for Element Extractions

To validate the correctness of element extraction of LLMs, we conduct a large number of sampling observations, and summarize the two main issues:

- **Date Hallucination.** This issue is mainly caused by two aspects: (1) **Date** elements are not presented in many cases, so this requires LLMs to question date existence rather than provide false dates, but LLMs are hardly aware of this situation. (2) In more difficult cases, date extraction involves reasoning (e.g. "*In 2014... Two years ago...*" → "*In 2012*"), which poses a greater challenge for extraction, and causes the sometimes failure of LLMs. Cases are presented in Table 15-17, these all explain why the F_1 score of **Date** is lower than that of the other elements (Table 8).

- **Element Redundancy.** LLMs frequently extract elements that are faithful to the source document but not important. Cases are presented in Table 18-19. This explains why the $\mathrm{Precision}$ score is lower than the Recall score in almost every element (Table 8).

C.3 Ablation Study of GPT-3 Model Size

We try diverse versions of GPT-3 with different model sizes. Model configurations are as follows:

- 0.3B-parameter `text-ada-001`
- 1.3B-parameter `text-babbage-001`
- 6.7B-parameter `text-curie-001`
- 175B-parameter `text-davinci-002`

The curve of F_1 score of different versions has been shown in Figure 5, and the case study is presented in Table 20.

D Random Sample Presentation

We randomly sample some examples, each containing: source document, golden summary, expert-writing summary, GPT-3 zero-shot summary, and GPT-3 reasoning-like zero-shot summary. Examples are shown in Table 21-24.

<table>
<tr><th colspan="2"><b>Source Document (<i>CNN/DailyMail</i>)</b></th></tr>
<tr><td colspan="2">(The Hollywood Reporter) Add another fan-favorite character to the cast of next year's "X-Men: Apocalypse" with director Bryan Singer announcing via Instagram that Olivia Munn will play the telepathic Psylocke in the follow-up to "X-Men: Days of Future Past." Singer revealed that the "Newsroom" actress would play Betsy Braddock in the movie (presumably before the confusing and complicated plot twist that saw Psylocke change from a Caucasian former supermodel to a Japanese ninja for no immediately obvious reason). Äpocalypseïs currently in production for a summer 2016 release. More: "X-Men: Apocalypse" casts fan favorite Jubilee. The comic book's Psylocke was created by Chris Claremont and Herb Trimpe for the British "Captain Britain" series, where she appeared throughout the 1970s and '80s, before joining the X-Men in 1987's "Uncanny X-Men" No. 213. Since that time, she has been a mainstay both of the main team and spin-off series including "Exiles" and "X-Force." More: What newcomers need to know about Marvel's "Secret Wars". Munn will join a cast that includes James McAvoy, Michael Fassbender and Jennifer Lawrence in the movie, which hits theaters May 27, 2016. Munn is repped by Creative Artists Agency and Atlas Artists. More: Does the big plot twist in "Terminator Genisys" blow up the franchise? @The Hollywood Reporter. All rights reserved."</td></tr>
<tr><th><b>Dataset-specific Summary</b></th><th><b>Element-aware Summary</b></th></tr>
<tr><td>Olivia Munn will play Psylocke in "x-men: apocalypse" film. Psylocke trended for hours on twitter after director Bryan Singer announced casting.</td><td>Olivia Munn will play the telepathic Psylocke -created by Chris Claremont and Herb Trimpe for the C̈aptain Britains̈eries - in the Ẍ-Men: Apocalypse.̈ The movie will be released in May 27, 2016.</td></tr>
<tr><td><i>% of novel uni/bi/trigram: 47.61/70.00/84.21</i></td><td><i>% of novel uni/bi/trigram: 17.65/48.51/62.50</i></td></tr>
</table>

Table 10: Comparisons between element-aware summaries and dataset-specific summaries in abstraction and faithfulness. Hallucinatory facts are highlighted in orange. We observe that dataset-specific summaries contain more hallucinatory facts despite a higher percentage of novel n-grams (Appendix B).

<table>
<tr><th colspan="2"><b>Source Document (<i>BBC XSum</i>)</b></th></tr>
<tr><td colspan="2">More than 350 roma people had lived in the camp on la petite ceinture since mid-2015. Activists said many left early ahead of the police action. The site belongs to the national rail authority sncf. France has one of europe's toughest policies towards roma. Most live in camps that are regularly demolished and every year thousands are deported. Amnesty international urged city authorities to find a lasting housing solution for those evicted in paris - saying they would become homeless in mid-winter. Hundreds of thousands of roma - mostly from romania and bulgaria - have moved to western europe since the 1990s. The council of europe, the region 's main human rights body, warned that evictions were " counter-productive" because they disrupted education and healthcare for roma children. Council of europe secretary general thorbjorn jagland said it was crucial for french authorities to provide "adequate alternative accommodation" for those evicted, particularly as they have decided to take this action during winter.</td></tr>
<tr><th><b>Dataset-specific Summary</b></th><th><b>Element-aware Summary</b></th></tr>
<tr><td>Police have cleared hundreds of roma people from a slum-like camp built on a disused rail line in north paris.</td><td>Every year thousands of Roma people are deported by France, and the region's main human rights body urges France to provide alternative accommodation for those evicted.</td></tr>
<tr><td><i>% of novel uni/bi/trigram: 28.57/85.00/100.00</i></td><td><i>% of novel uni/bi/trigram: 25.93/53.85/80.00</i></td></tr>
</table>

Table 11: Comparisons between element-aware summaries and dataset-specific summaries in abstraction and faithfulness. Hallucinatory facts are highlighted in orange. We observe that dataset-specific summaries contain more hallucinatory facts despite a higher percentage of novel n-grams (Appendix B).

Source Document (*CNN/DailyMail*)

Once famed for his mop of blacker than black hair, disgraced Democrat Rod Blagojevich, 58, has really let his haircare regime go while he serves his prison time. The former Illinois governor has return to his roots while inside and has been photographed with his still full head of hair a shocking white color rather than the boot polish black that was his trademark as a politician. Blagojevich was infamously caught trying to sell Barack Obama's U.S. Senate seat when he was elected president in 2008. Fade to gray: Once famed for his mop of blacker than black hair, disgraced Democrat Rod Blagojevich, 58, has really let his haircare regime go while he serves his prison time. Back in his days as governor of Illinois, Blagojevich was famed for his boot polish black hair. He was impeached and removed from office by the state Legislature in early 2009 following his arrest on federal corruption charges. Blagojevich is currently serving a 14-year sentence at the Federal Correctional Institution Englewood near Denver. The photos, the first of Blagojevich in prison, were first published by the National Enquirer. The photos show the former governor sitting at a desk, reading a book and wearing glasses and sitting outside in shorts and a wife beater shirt. While he continues to awaiting word on a last-ditch appeal, Blagojevich has daily contact with his family and teaches the history of war battles to other inmates, his attorney said. Last July lawyers attempted to argue that a high court ruling in McCutcheon v. the Federal Election Commission in April 2014 bolstered their view that Blagojevich was engaged in legal, run-of-the-mill political horse trading, not corruption. The McCutcheon ruling, they argued, found that soliciting contributions crosses the line into corruption only when a politician makes a clear, explicit promise to take official action in return for a donation. Blagojevich is currently serving a 14-year sentence at the Federal Correctional Institution Englewood near Denver after he was caught trying to sell Barack Obama's U.S. Senate seat when he was elected president. In that context, defense attorneys said the trial judge was wrong to set a lower standard for jurors by telling them, to convict, they only needed to find that Blagojevich sought donations 'knowing or believing' it would be given for some official act. But government attorneys responded saying that Blagojevich's legal team hadn't correctly characterized the Supreme Court's findings in McCutcheon. 'The decision provides no support for Blagojevich's argument on appeal,' the two-page filing said. 'Nothing in the (McCutcheon) decision suggests that an exchange of contributions for specific official acts is quid pro quo corruption only if the arrangement is stated 'explicitly' or 'expressly.' The Chicago-based appellate court has been considering Blagojevich's appeal for more than a year months. It's unclear when it might issue a decision.

GPT-3 Zero-shot Summary (Standard)

Summarize the above article:
The article discusses how Rod Blagojevich, a former Illinois governor, has let his hair go while serving his prison sentence. The photos show the former governor with white hair, rather than the black hair that was his trademark as a politician. Blagojevich is currently serving a 14-year sentence at the Federal Correctional Institution Englewood near Denver.

Summary Chain-of-Thought (Stage 1, Element Extraction)

What are the important entities in this document? What are the important dates in this document? What events are happening in this document? What is the result of these events?
Please answer the above questions:
1. The important entities in this document are Rod Blagojevich, the Federal Correctional Institution Englewood, and the National Enquirer.
2. The important dates in this document are 2008, 2009, and 2014.
3. The events in this document are Blagojevich's arrest and imprisonment, and his appeal of his conviction.
4. The result of these events is that Blagojevich is currently serving a 14-year sentence in prison.

GPT-3 Zero-shot-cot Summary (Stage 2, Summary Generation)

Let's integrate the above information and summarize the article:
In 2008, Rod Blagojevich was arrested for trying to sell Barack Obama's U.S. Senate seat. He was impeached and removed from office in 2009, and sentenced to 14 years in prison. He is currently serving his sentence at the Federal Correctional Institution Englewood near Denver. In 2014, his lawyers attempted to appeal his conviction, arguing that a recent Supreme Court ruling supported their view that Blagojevich was engaged in legal, run-of-the-mill political horse trading, not corruption.

Table 12: More comparisons between GPT-3 zero-shot summaries before and after using SumCoT. Spans of ***Entity***, ***Date***, ***Event***, and ***Result*** are separately highlighted in red, yellow, blue and green. **Prompts** are presented in *italics* (Appendix C.1).

Source Document (*BBC XSum*)

The 69-year-old's yamaha collided with a Nissan car between handley's corner and barre garroo crossroads at about 17:00 bst on 4 June. Mr. Baker, who was from the island, was airlifted to noble's hospital, where he later died. Insp dave dobbie said "thoughts and condolences are with Mr. baker's family". He added that investigations are ongoing in relation to the crash. The car driver, who police say was northern irish, was treated in hospital but has been discharged. Another motorcyclist who was injured after the crash has also been released from hospital. The collision happened outside of the TT racing, which is taking place on the island.

GPT-3 Zero-shot Summary (Standard)

Summarize the above article in one sentence:
The article discusses a fatal motorcycle collision that occurred on the Isle of Man, and the investigation that is ongoing.

Summary Chain-of-Thought (Stage 1, Element Extraction)

What are important entities in this document? What are important dates in this document? What events are happening in this document? What is the result of these events?
Please answer the above questions:

1. The important entities in this document are Mr. Baker, the car driver, and the motorcyclist who was injured.
2. The important dates in this document are 4 June and the present day.
3. The events happening in this document are a collision between Mr. Baker's motorcycle and a car, and the investigation into the collision.
4. The result of these events is that Mr. Baker died and the car driver and motorcyclist were injured.

GPT-3 Zero-shot-cot Summary (Stage 2, Summary Generation)

Let's integrate the above information and summarize the article in one sentence:
On 4 June, Mr. Baker's motorcycle collided with a car, resulting in his death. The car driver and motorcyclist were injured.

Table 13: More comparisons between GPT-3 zero-shot summaries before and after using SumCoT. Spans of ***Entity***, ***Date***, ***Event***, and ***Result*** are separately highlighted in red, yellow, blue and green. **Prompts** are presented in *italics* (Appendix C.1).

Source Document (*BBC XSum*)

West midlands police said the man had been detained in reading and the woman in wokingham by counter-terrorism officers at about 19:30 bst. The pair, believed to be friends, were taken to a police station in the west midlands for questioning. The force said the arrests had been intelligence-led. the suspects did not pose any imminent threat, a spokesperson added. Supt robin rickard from thames valley police said the arrests "were part of an ongoing investigation by west midlands counter terrorism unit and were pre-planned". "Officers from reading and wokingham are providing support and there are searches taking place at the suspects' home addresses," he said.

GPT-3 Zero-shot Summary (Standard)

Summarize the above article in one sentence:
Two people were arrested in England on suspicion of terrorism.

Summary Chain-of-Thought (Stage 1, Element Extraction)

What are the important entities in this document? What are the important dates in this document? What events are happening in this document? What is the result of these events?
Please answer the above questions:
The important entities in this document are the two suspects who were arrested, the police forces who made the arrests, and the locations where the arrests took place. The important dates are the day of the arrests (19:30 BST) and the day of the document (20:00 BST). The events happening in this document are the arrests of the two suspects and the searches of their home addresses. The result of these events is that the two suspects are in police custody and their home addresses are being searched.

GPT-3 Zero-shot-cot Summary (Stage 2, Summary Generation)

Let's integrate the above information and summarize the article in one sentence:
On 19:30 BST, two suspects were arrested by police forces in Reading and Wokingham on intelligence-led grounds, and their home addresses are currently being searched.

Table 14: More comparisons between GPT-3 zero-shot summaries before and after using SumCoT. Spans of ***Entity***, ***Date***, ***Event***, and ***Result*** are separately highlighted in red, yellow, blue and green. **Prompts** are presented in *italics* (Appendix C.1).

Error Type: Date Hallucination

Source Document (*CNN/DailyMail*)

Charity runners taking part in a 10km fun run at the weekend were left exhausted after being sent on an unscheduled two-mile detour. The blunder was believed to have been caused by a race marshal taking a toilet break during the event, missing 300 runners who should have been directed at a junction point. Instead they continued past the unmanned marshall point and had to run for an extra three kilometres while the other 900 competitors followed the correct route. Scroll down for video Blunder: Charity runners taking part in yesterday's Bournemouth Bay 10K Run (pictured) were left exhausted after being sent on an unscheduled two-mile detour. The bizarre gaffe happened during yesterday's Bournemouth Bay Run and today the organisers - Bournemouth Borough Council - appealed for those who were affected by the mix-up to contact them for a 'gesture of goodwill.'A local authority spokesman said that it was investigating what happened to the marshal who should have directed runners at a turning point. It was reported that some runners were 'in tears' while one described the event's organisation as 'shambolic'. Hayley James, who is four months pregnant and from Poole, said: 'To have a race of that scale with only one marshal on a point is inexcusable.'We saw loads of people walking at the end, some were in tears, I felt so sorry for them - I felt like crying at the 10km mark.'Andy Isaac, from Bournemouth, said the event was 'mayhem' with one point where an elderly woman managed to drive onto the route and was flashing her lights at oncoming runners. A map shows where up to 300 runners continued along the coastal path after a marshal who was meant to direct them on to a turn went to the toilet. Reaction: Two people vent their frustration at the Bournemouth Bay Run on Twitter yesterday. It also emerged that water stations ran out of supplies during the race, forcing some runners to drink from half-empty bottles that had been left on the ground as they battled against dehydration.Commenting on the Daily Echo website, one runner said: "We had a bottle of water at the three mile station, but at the six mile point they had totally ran out, so nothing. Jon Weaver, head of resort marketing and events at the council, said: 'Unfortunately there was some confusion with marshalling arrangements at one point, but it was a critical point. We apologise unreservedly to those front runners.'In 33 years of running... this is the first time this has happened and as part of our debrief we will be analysing the arrangements carefully... to learn for 2016.'We understand runners have trained for a long time for the event and it's hard for them and we do empathise with how they are feeling.'It was hoped that the event would have raised more than £70,000 for the British Heart Foundation. Some racers took to Twitter to vent their frustration over the blunder. Rob Kelly wrote: Really disappointed in the #BournemouthBayRun 10k that ended up 13k very poor show bad marshalling #wontbeback.' And Kim Kelly replied: 'Totally agree and never got to do a 5k as they were 45mins behind schedule :(((.'
What are the important entities in this document?
What are the important dates in this document?
What events are happening in this document?
What is the result of these events?
Please answer the above questions:

Element Extraction

1. The important entities in this document are the charity runners, the race marshal, and the Bournemouth Borough Council.
2. The important dates in this document are the date of the race (May 3, 2015) and the date the runners were supposed to be directed at the turning point (May 2, 2015).
3. The events happening in this document are the charity runners being sent on an unscheduled two-mile detour, and the race marshal taking a toilet break.
4. The result of these events is that the charity runners were left exhausted after being sent on an unscheduled two-mile detour.

✗ **(The dates highlighted in blue are not declared in the source document.)**

Table 15: Element extraction errors of GPT-3. **Prompts** are presented in *italics* (Appendix C.2).

Error Type: Date Hallucination

Source Document (*BBC XSum*)

patrick Bauer and Nathan Byrne helped Karl Robinson's men complete a perfect smash and grab, cancelling out Zach clough's free-kick. a victory which furthers the Addicks' own play-off ambitions was made all the more impressive considering that they were a man short from the 12th minute onwards following lewis page's red card. after page was sent off for a professional foul on josh vela, clough then made the punishment count double with a brilliant free-kick, his ninth goal of the season. but robinson 's side refused to lie down, equalising on 23 minutes when jake forster-caskey 's free-kick was turned in at the far post by defender bauer. the Addicks then grabbed a shock lead in first-half stoppage time when Forster-Caskey led an impressive counter, feeding Byrne to slot the ball under the advancing Ben Alnwick. bolton, now without a win in their last five games in all competitions, threw all they had at Charlton after that and struck the woodwork when Andrew Taylor's cross deflected off Chris Solly. they failed to seriously test goalkeeper Declan rudd in the second half, however, as they slipped to a second consecutive 2-1 defeat in front of their own fans. Charlton, unbeaten since christmas, have now picked up 11 points from their last five matches to climb to 11th, just six points off a play-off place. report supplied by the press association.
What are the important entities in this document?
What are the important dates in this document?
What events are happening in this document?
What is the result of these events?
Please answer the above questions:

Element Extraction

1. The important entities in this document are Charlton Athletic, Karl Robinson, Zach Clough, and Ben Alnwick.
2. The important dates in this document are December 25th and January 23rd.
3. The events happening in this document are a football match between Charlton Athletic and Bolton Wanderers, and the aftermath of the match.
4. The result of these events is that Charlton Athletic won the match 2-1.

✗ **(The dates highlighted in blue are not declared in the source document.)**

Table 16: Element extraction errors of GPT-3. **Prompts** are presented in *italics* (Appendix C.2).

Error Type: Date Hallucination

Source Document (*BBC XSum*)

club director Emilio butragueno confirmed FIFA had asked the Spanish federation for documents relating to transfers over the last five years. but ragueno said the club was " absolutely calm " about its behaviour. **fifa handed Barcelona a transfer ban until 2016 after they breached regulations when signing 10 players under the age of 18 in 2013 .** in a statement, real Madrid said the majority of the 51 players that world football's governing body had asked for information on were exempt from regulations for various reasons. FIFA rules state that international transfers are only permitted for players who are over the age of 18 - unless the player in question meets one of three qualifying criteria. under-18s can move to a club in a different country if: real claim 10 of the 51 players were Spanish nationals, while 23 foreign players had already been registered with a Spanish club. on Thursday, real completed the # 2.3 m signing of 16-year-old Norway midfielder Martin Odegaard from stromgodset. **barcelona are not allowed to sign any players in 2015 after their appeal against their transfer ban was dismissed in December.**
What are the important entities in this document?
What are the important dates in this document?
What events are happening in this document?
What is the result of these events?
Please answer the above questions:

Element Extraction

1. The important entities in this document are FIFA, Barcelona, and Real Madrid.
2. The important dates in this document are 2013, when Barcelona was first given a transfer ban, and December of 2014, when Barcelona's appeal against the ban was denied.
3. The events in this document are FIFA's investigation into Barcelona's transfer practices, and the eventual imposition of a transfer ban on the club.
4. The result of these events is that Barcelona is not allowed to sign any players in 2015.

✗ **(The date when Barcelona's appeal against the ban was denied is likely to be December 2013 or December 2014 deduced from the bolded part of the source document, so the extracted date highlighted in blue is evidence-insufficient.)**

Table 17: Element extraction errors of GPT-3. **Prompts** are presented in *italics* (Appendix C.2).

Error Type: Element Redundancy

Source Document (*BBC XSum*)

there were concerns about the stability of the building, which has been derelict for years, after the collapse of an internal wall last week. contractors began work on Friday night. the environment minister Alex Atwood recommended the demolition in the interests of public safety. mary o'dwyer, chair of Foyle civic trust, said a strategic approach is needed for the future protection of heritage in the city following the demolition . " many of our members believe that this is a very sad event . " the factory is part of our industrial heritage and more should have been done to save it . " others believe something had to be done because it was a health and safety issue, " she said . " we would like to see a consensus and a strategic vision of all of the organizations in the city that have a remit for protecting our heritage. " Foyle road, which was closed between the lower deck of the Craigavon bridge and john street roundabout, has reopened.
What are the important entities in this document?
What are the important dates in this document?
What events are happening in this document?
What is the result of these events?
Please answer the above questions:

Element Extraction

The important entities in this document are the Foyle Civic Trust, the environment minister Alex Atwood, and the contractors who began work on Friday night. The important dates in this document are the collapse of an internal wall last week, and the demolition of the factory on Friday night. The events happening in this document are the demolition of the factory, and the reopening of Foyle Road. The result of these events is that the factory is gone and Foyle Road is open again.

✗ **(The elements highlighted in blue is redundent.)**

Table 18: Element extraction errors of GPT-3. **Prompts** are presented in *italics* (Appendix C.2).

Error Type: Element Redundancy

Source Document (*BBC XSum*)

some of those residents had only just returned to their homes after last week's eruptions. the latest eruption was less powerful but sent a large plume of dark grey smoke and ash rising from the crater. calbuco surprised residents of the Los Lagos region last week by bursting into life after decades of inactivity. the national geology and mines service issued a red alert after Thursday's eruption, warning that the volcano remained unstable. on Friday, residents were warned that heavy rain could mix with ash and rock to produce deadly volcanic mudflows. these " lahars " can slide down volcanoes like avalanches. the BBC's Gideon long in Santiago says that whereas the ash last week was blown in a north-easterly direction, this time it is drifting south-east over the town of cochamo. the town has been blanketed in a cloud of smoke and ash. calbuco is one of the most active volcanoes in chile, having erupted in 1961 and 1972, and many times before that. last week's eruptions caused the cancellation of flights to and within chile and forced farmers in the area to evacuate their livestock.
What are the important entities in this document?
What are the important dates in this document?
What events are happening in this document?
What is the result of these events?
Please answer the above questions:

Element Extraction

1. The important entities in this document are the volcano, the residents, and the government.
2. The important dates in this document are the dates of the eruptions.
3. The events happening in this document are the eruptions of the volcano and the resulting evacuations.
4. The result of these events is that the residents have been forced to leave their homes and the area around the volcano is covered in ash.

✗ **(The elements highlighted in blue is redundent.)**

Table 19: Element extraction errors of GPT-3. **Prompts** are presented in *italics* (Appendix C.2).

Source Document (***BBC XSum***)
They say every cloud has a silver lining. At least that's what one Sydney woman, caught out in this week's super storm, is hoping. With the rain clearing, a woman known only by her email address as MissieSmile21, has taped posters to the city's light posts to find the 'tall, charming gentleman' who offered her his umbrella to cross the street.Scroll down for video The 'Love Call' notes are posted at traffic lights on Sydney's George Street near Australia SquareActively looking: Missie Smile is hoping the generous stranger will get in contact. The 'Love Call' notices, posted on George Street at Australia Square in Sydney's CBD, are appealing to the man with 'the most beautiful smile' who was in the area about '7:10 pm-ish' on Tuesday evening.'I am actively looking for you,' the poster reads. 'If you're single, let's grab a drink together.' People struggled against the heavy rain and strong winds in the Sydney CBD as the storm lashed NSW on Tuesday this weekMost people gave up on their umbrellas in winds of more than 100 km/ hr. *What are the important entities in this document?* *What are the important dates in this document?* *What events are happening in this document?* *What is the result of these events?* *Please answer the above questions:*
`text-ada-001`
The 'Love Call' notes are posted at traffic lights on Sydney's George Street near Australia Square. The 'Love Call' notices, posted on George Street at Australia Square, are appealing to the man with 'the most beautiful smile' who was in the area about '7:10 pm-ish' on Tuesday evening. People struggled against the heavy rain and strong winds in the Sydney CBD as the storm lashed NSW on Tuesday this week. Most people gave up on their umbrellas in winds of more than 100 km/ hr. What are important dates in this document? What are important entities in this document? What are important events in this document? Please answer the above questions: The 'Love Call' notes are posted at traffic lights on Sydney's George Street near Australia Square. The 'Love Call' notices, posted on George Street at Australia Square, are appealing to the man with 'the most beautiful smile' who was in the area about '7:10 pm-ish' on Tuesday evening. People struggled against the heavy rain and strong winds in the Sydney CBD as the storm lash NSW on Tuesday this week. Most people gave up on their umbrellas in winds of more than 100 km/ hr. What is the result of these events? Please answer the above questions: At 7:10 pm on Tuesday evening, a man with the most beautiful smile, known only as MissieSmile21, was in the area near Australia Square. The 'Love Call' notices, posted on George Street at Australia Square, are appealing to the man with the most beautiful smile who was in the area about '7:10 pm-ish' on Tuesday evening. People struggled against the heavy rain and strong winds in the Sydney CBD as the storm lash NSW on Tuesday this week. Most people gave up on their umbrellas in winds of more than 100 km/ hr. ✗ **(The extracted information is very confusing and illogical.)**
`text-babbage-001`
1. What is the email address of the woman whoposted the 'Love Call' notices on Sydney's George Street near Australia Square? 2. What is the date of the 'Love Call' notices? 3. What is the time of the 'Love Call' notices? 4. What is the type of person who was in the area around 7:10 pm-ish on Tuesday evening? 5. What is the result of the 'Love Call' notices? ✗ **(The model fails to accurately understand the meaning of guiding questions.)**
`text-curie-001`
The important entities in this document are the poster, the man she is looking for, and the date 7/10/2016. The important dates in this document are Tuesday, July 10th. The events in this document are the heavy rain and strong winds. The result of these events is that many people were unable to use their umbrellas and many people were stranded. ✗ **(The model understands the guiding questions, but extracts many redundant or false facts shown in blue spans.)**
`text-davinci-002`
The important entities in this document are MissieSmile21, the tall, charming gentleman, and George Street at Australia Square. The important dates are Tuesday evening and 7:10 pm. The events are the super storm and MissieSmile21's search for the tall, charming gentleman. The result of these events is unknown. ✓ **(The model can accurately extract human-identified core elements from the source document.)**

Table 20: Element extraction results of different GPT-3 versions. **Prompts** are presented in *italics* (Appendix C.3).

Source Document (*CNN/DailyMail*)
A Tory councillor with a history of odd behaviour was told to put down his knife and fork after being caught tucking into a roast duck dinner during a council meeting. Jonny Bucknell, 58, was enjoying his meal in the council chamber when a Labour rival, Theo Blackwell, spotted him and alerted other councillors. He was forced to put down his cutlery when the mayor, Lazzaro Pietragnoli, interrupted the proceedings to tell him off. Taking a stand: Jonny Bucknell is no stranger to odd behaviour. In 2013 he slept in his car at the Tory party conference. He now says he wants a rule change so he can eat a roast dinner at council meetings. The mayor, who was chairing the meeting of Camden Council in north London, reminded the hungry councillor that eating was banned in the chamber. But the angry diner claims he was unaware eating there was forbidden and said he now aims to campaign for a rule change. The rumpus comes a month after Liberal Democrat councillor Martin Elengorn was caught playing Scrabble during a Richmond Council budget meeting in south-west London. Telling off: Mayor of Camden Council, Lazzaro Pietragnoli, had to tell Mr Bucknell to stop eating. When he first noticed him eating, Mr Blackwell told his fellow councillors: ’It appears that one of our Tory colleagues is consuming a full Sunday roast dinner in the council chamber. ’Could I ask the borough solicitor to give us advice on eating a full roast dinner in the council chamber? It’s a little bit more than a cheeky Snickers.’ The diner was forced to curtail his meal. Mr Bucknell, who has been a councillor for more than ten years and represents Belsize, Hampstead, told the Evening Standard: ’I never knew there was a ban on eating in the chamber. ’They should definitely repeal it. There is nothing wrong with nibbling from a lunch box if you are being discreet. ’It is not as if a cold meal is going to waft around like a McDonald’s. ’I will be campaigning for the repealing of the law that says you can’t nibble from a lunch box in the council chamber.’ The Conservative councillor said the meal, in a plastic box, had travelled home with him after a French snowboarding holiday. ’The chalet always brought out too much food and I can’t stand wasting food,’ he said. He previously found fame when he slept in his Volvo car to save on hotel bills during the Conservative party conference in September 2013. Mr Bucknell said at the time it was to make a stand against what he called ’ridiculous prices’. He said the economy would improve if more people were thrifty like him. After the council meeting mayor Lazzaro Pietragnoli said: ’I understand councillors do a difficult job and sometimes don’t get time to eat dinner. ’I also tend to be quite flexible, but having a big meal in front of him - that was a bit too much.’ A Camden Council spokesman said: ’It is as the mayor said, standing orders say that members should not eat in the chamber.’ No eating: The Camden Council chamber where Councillor Bucknell was forced to curtail his roast dinner.
Dataset-specifc Summary
Labour rival alerted colleagues after spotting Councillor tucking into roast. It is not the first time jonny Bucknell, 58, has demonstrated odd behaviour.. In 2013 he slept in his car to make a point while attending Tory conference. Mr Bucknell said he will campaign for rule change about eating at meetings.
Element-aware Summary
Jonny Bucknell, a Tory councillor, was found tucking into a roast duck dinner during a council meeting. He now wants a rule change to have dinner at council meetings. It is not the first time for his strange behavior. In September 2013, Mr Bucknell slept in his Volvo car to save on hotel bills during the Conservative party conference to make a stand against what he called ’ridiculous prices’.
GPT-3 Zero-shot Summary
A Tory councillor was told to stop eating a roast duck dinner during a council meeting after a Labour rival spotted him and alerted other councillors. The councillor, Jonny Bucknell, claims he was unaware that eating in the council chamber was forbidden and said he now aims to campaign for a rule change.
GPT-3 Zero-shot-cot Summary
In September 2013, Councillor Jonny Bucknell made headlines when he slept in his car during the Conservative party conference in order to save on hotel bills. Recently, he has found himself in the news again after being caught eating a roast duck dinner during a council meeting. Councillor Bucknell claims that he was unaware that eating in the council chamber was forbidden, and is now campaigning for a rule change so that he can eat during meetings.

Table 21: Random samples from *CNN/DailyMail* and *BBC XSum* datasets (Appendix D).

Source Document (*CNN/DailyMail*)
This is the dramatic moments armed police swoop on a villa where a Briton linked to the gangland murder of a torture victim was arrested. Paul Monk, 54, from Essex, was wanted by Spanish police for questioning over the kidnap and murder of Francis Brennan, whose badly decomposed body washed up on a Costa Blanca beach in March last year. He was also wanted by the Metropolitan Police on drug offences and had been named on a list of fugitives published as part of the National Crime Agency's Operation Captura campaign ahead of his detention. This is the dramatic moment that fugitive Paul Monk was arrested by heavily armed police in his Alicante villa. Paul Monk, 54, from Essex, was wanted by Spanish police for questioning over the kidnap and murder of Francis Brennan. Spanish police released footage of their dramatic swoop. This grab for the video shows them approaching the villa at speed. The police move steathily up the steps of Monk's villa, weapons drawn. Taking no chances: The highly trained, well-armed police moved through the house room by room. Paul Monk was on the UK's most wanted list on suspicion of drug trafficking. Brennan, 25, from Liverpool, vanished in the resort of Javea in January last year after being kidnapped by men posing as police. His body was wrapped in an industrial-size bin bag with duct tape round it when it appeared on a beach in nearby Orihuela Costa. Civil Guard officers in Alicante confirmed today they believe Monk, from Essex, may be implicated in the violent death and named him as an associate of Paul Scott. Scott, 32, was arrested on a charge of conspiracy to import cocaine after being caught trying to sneak into Britain in a light aircraft last December. He was also wanted for questioning over Mr Brennan's murder when he was detained. Guardia Civil described him last night as the suspected mastermind of the crime. Monk was detained at a four-bedroom property in Javea near Alicante as he directed workers laying a marble patio around his swimming pool. An imitation firearm with a silencer and nearly 0̆0a3100,000 in cash were also found. He is being held in jail and is expected to be charged and face trial in Spain over Mr Brennan's murder before being extradited to the UK to face questioning over alleged drugs offences. He has been linked to the handover of one kilo of cocaine in Cockfosters, London, in May 2013 and the seizure of 24 kilos of cannabis in Colchester in October 2013. A Civil Guard spokesman said: 'He never left his house as a security measure to avoid being arrested. 'He got other people to bring him food and other things in the villa where he hid out, leading the life of an authentic fugitive.' The police raid had air support, with this grab coming from footage of Monk's villa taken by a helicopter. Wads of money found by armed police after they arrested Monk . Monk is being held in jail and is expected to be charged and face trial in Spain over Mr Brennan's murder before being extradited to the UK to face questioning over alleged drugs offences. Spanish police search Monk's property thoroughly for evidence, finding an imitation gun with a silencer. National Crime Agency detectives took part in the raid on the property along with Civil Guard officers. Mr Brennan, from Liverpool, was himself on the run after leaving Britain while awaiting sentence in the UK over the wounding of a man at a Swedish House Mafia music concert in Milton Keynes, Bedfordshire. He travelled to the Costa del Sol with his girlfriend towards the end of 2013 after ignoring the advice of his parents to stay in Britain. Police swooped on Scott last year after he flew from Holland to a remote airstrip in Norfolk. He had fake Belgian identity documents on him and more than 0̆0a31,500 in cash. He was jailed in January for plotting to import 40 tonnes of cocaine into Britain. Police described him as the 'last man outstanding in a criminal gang that plotted to smuggle vast quantities of cocaine into the UK' after he was sentenced to 14 years in prison at Liverpool Crown Court. Sorry we are not currently accepting comments on this article.
Dataset-specific Summary
Paul Monk, 54, was wanted by Spanish Police in connection with a murder.. The Essex man is a suspect in the murder of Francis Brennan.. Brennan's body washed up on a Costa Blanca beach in March last year.. Police released footage of their swoop on Monk's Alicante villa.
Element-aware Summary
Paul Monk, 54, from Essex, was wanted by Spanish police for questioning over the kidnap and murder of Francis Brennan, 25, from Liverpool. Brennan2̆019 body was found on a beach in March last year. Paul Monk was swoop by armed police on a villa, he will be arrested and face trial in Spain for this murder.
GPT-3 Zero-shot Summary
A man wanted for questioning in connection with the murder of a torture victim has been arrested by Spanish police. Paul Monk, 54, from Essex, was detained at his villa in Javea near Alicante as he directed workers laying a marble patio around his swimming pool. An imitation firearm with a silencer and nearly 0̆0a3100,000 in cash were also found. Monk is being held in jail and is expected to be charged and face trial in Spain over the murder before being extradited to the UK to face questioning over alleged drugs offences.
GPT-3 Zero-shot-cot Summary
Paul Monk, 54, from Essex, was arrested by Spanish police in his Alicante villa for questioning over the kidnap and murder of Francis Brennan, 25, from Liverpool. Brennan's body was found in March last year, wrapped in an industrial-size bin bag with duct tape round it, on a beach in nearby Orihuela Costa. Monk is being held in jail and is expected to be charged and face trial in Spain over Mr Brennan's murder before being extradited to the UK to face questioning over alleged drugs offences.

Table 22: Random samples from *CNN/DailyMail* and *BBC XSum* datasets (Appendix D).

Source Document (*BBC XSum*)
They've kicked up a stink with their council, who've decided to use the liquid to kill weeds around the city. It's all part of a trial to get rid of the unwanted plants without using strong chemicals, which some people think aren't environmentally friendly. But many locals say they're not happy with the unpleasant pong the trial's creating. They've started a campaign and also say they don't think vinegar is strong enough to kill the weeds. Bristol city council say they'll continue with their trial, but will keeping track of feedback from the people who live there.
Dataset-specific Summary
People living in bristol have complained about a vinegary whiff in the air.
Element-aware Summary
Bristol city is conducting a trial to kill weeds with vinegar, but the unpleasant pong makes the locals unhappy, which led to a conflict between the two.
GPT-3 Zero-shot Summary
The city of Bristol is conducting a trial to see if vinegar can be used as an environmentally friendly weed killer, but many locals are not happy with the strong smell.
GPT-3 Zero-shot-cot Summary
The city council of Bristol is conducting a trial to kill weeds using vinegar, but many locals are unhappy with the unpleasant smell the trial is creating.

Table 23: Random samples from *CNN/DailyMail* and *BBC XSum* datasets (Appendix D).

Source Document (*BBC XSum*)
Cpl Kylo gorman's medals for service in Afghanistan and at the queen 's diamond jubilee were taken from his car in bolton-le-sands on 4 or 5 January. He said he wasn't that bothered at first, but the reaction online got me thinking "it is a really big deal" . Lancashire police said they were investigating the theft. The port talbot-born 28-year-old, who has served five tours of Afghanistan, said he had left the medals in the car after removing them from a uniform he was having altered. "I didn't think twice about them. I'm based at raf leeming in north yorkshire during the week and when I came home and the car had been broken into, I realised my medals had gone," he said. "I want my children to be able to look at their dad's medals and be proud of what I did," he added. Cpl gorman, who has also served in the Falklands, Africa and Cyprus, has been in the RAF for 10 years and is due to leave the armed forces in April.
Dataset-specific Summary
A RAF airman has made an appeal for the return of two stolen medals after hundreds shared his post about the theft on social media.
Element-aware Summary
Cpl Gorman's medals were stolen on 4 or 5 january, and he has made an appeal for the return. Lancashire police are investigating the theft.
GPT-3 Zero-shot Summary
The medals of a British serviceman were stolen from his car, and he is now appealing for their return.
GPT-3 Zero-shot-cot Summary
Cpl Gorman's medals for service in Afghanistan and at the Queen's Diamond Jubilee were stolen from his car on 4 or 5 January, and Lancashire police are investigating the theft.

Table 24: Random samples from *CNN/DailyMail* and *BBC XSum* datasets (Appendix D).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3 4

☑ B1. Did you cite the creators of artifacts you used?
3.1 Appendix A

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics Statement

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethics Statement

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2.1 2.3

C ☑ Did you run computational experiments?

3 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3.1 Appendix C.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3.1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
2 3 4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
2.2 3.3

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix A

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix A

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics Statement

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix A