Exploring Iterative Controllable Summarization with Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated remarkable performance in abstractive summarization tasks. However, their ability to precisely control summary attributes (e.g., *length* or *topic*) remains underexplored, limiting their adaptability to specific user preferences. In this paper, we systematically explore the controllability of LLMs. To this end, we revisit summary attribute measurements and introduce iterative evaluation metrics, *failure rate* and *average iteration count*, to more precisely evaluate controllability beyond assessment of errors. Our findings show that LLMs struggle more with numerical attributes than with linguistic attributes. To address this challenge, we propose a guide-to-explain framework (GTE) for controllable summarization. GTE enables the model to identify misaligned attributes in the initial draft and guides it to self-explain errors in the previous output. By encouraging reflection on attribute misalignment, GTE generates well-adjusted summaries that satisfy the desired attributes with robust effectiveness while requiring surprisingly fewer iterations than other iterative approaches.

1 Introduction

Large language models (LLMs) have demonstrated superior performance in text summarization, outperforming encoder-decoder models by generating more contextually appropriate and natural summaries (Goyal et al., 2023; Zhang et al., 2024; Pu et al., 2023; Ryu et al., 2024b). However, given the diversity of individual preferences for summary styles, it is essential to generate summaries tailored to specific user needs (Zhang et al., 2023b). For example, some users may prefer topic-focused summaries or wish to retain exact phrases. Although LLMs excel at generating fluent summaries, their ability to precisely control attributes remains underexplored (Liu et al., 2024), limiting their adaptability to diverse user preferences. Typical requests

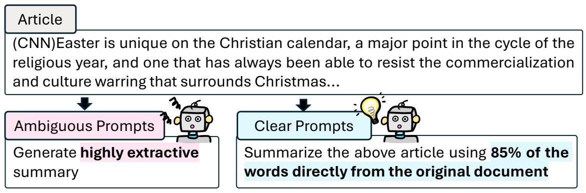

Figure 1: Ambiguous instructions hinder LLMs' ability to follow control signals and complicate the evaluation process (e.g., how should "highly" be judged in a generated summary?).

can be ambiguous, such as "summarize in 3 sentences" or "generate a highly extractive summary". Sentence lengths can vary significantly, and vague terms such as "highly" hinder reliable instruction-following and complicate evaluating whether the instructions are properly satisfied (Figure 1).

Therefore, we systematically explore the controllability of LLMs. We begin by revisiting the measurements for four key attributes: *extractiveness*, *length*, *topic*, and *speaker*, and refine them for more precise measurement. Specifically, rather than relying solely on word presence as in previous strategies for measuring *topic*- or *speaker*-focused summaries, we adopt embedding-based similarity to incorporate semantic information into the measurements. With more precise attribute measurements in place, we next investigate how reliably LLMs can control these attributes. To fully explore LLM controllability, we evaluate whether LLMs can accurately control specified attributes through iterative refinement. Even if initial attempts fail, we test whether they can eventually succeed without external guidance. To this end, we introduce two evaluation metrics: the *failure rate*—the proportion of control failures within the maximum iterations—and the *average iteration count* until successful control. In Section 4, we show that while LLMs excel at controlling linguistic attributes such as *topic* and *speaker*, they struggle significantly with numerical attributes such as *extractiveness* and *length*. We assume that, unlike linguistic attributes, which rely on semantic coherence, numerical ones require adherence to strict structural constraints, making fine-grained control challenging.

To address this challenge, we propose a guide-to-explain (GTE) framework, which enables precise attribute control solely through LLMs without relying on additional attribute-specific training. We first design a step-by-step attribute identification phase that instructs the model to identify misaligned attributes in its previously generated summary and then guides it to explain the rationale behind its errors. Through self-reflection, the model corrects its prior mistakes and generates a well-aligned summary in the regeneration phase. By integrating a self-refinement strategy—proven effective in complex reasoning tasks (Weng et al., 2023; Madaan et al., 2023; Dhuliawala et al., 2024; Gou et al., 2024)—into controllable summarization, we improve the controllability of LLMs while enhancing summary quality.

We evaluate GTE on mixed-attribute summarization datasets, MACSum$_{Doc}$ and MACSum$_{Dial}$ (Zhang et al., 2023b). GTE successfully controls each attribute with minimal iterations, significantly outperforming other iterative methods and demonstrating robustness by consistently adjusting attributes across data samples. Furthermore, we demonstrate the high quality of the controlled summaries across multiple generic summarization evaluation metrics, including UniEval (Zhong et al., 2022) and QuestEval (Scialom et al., 2021). Finally, we analyze whether LLMs can control multiple attributes simultaneously, revealing their difficulty in jointly managing correlated numerical attributes. Our contributions are as follows:

- We systematically explore the controllability of LLMs in text summarization.
- We refine the measurement of summarization attributes and introduce *iterative evaluation* metrics to evaluate LLM controllability.
- We propose a guide-to-explain (GTE) framework, which guides the model to explain its misalignments and effectively adjust attributes within just a few iterations.

2 Related work

Controllable summarization Controllable summarization has recently gained attention due to its practical applications (Zhong et al., 2021; Ahuja et al., 2022; Maddela et al., 2022; Mehra et al., 2023; Xu et al., 2023; Zhang et al., 2023b; Ribeiro et al., 2023). Previous research has employed encoder-decoder models to control attributes (Fan et al., 2018; Liu and Chen, 2021; Dou et al., 2021; He et al., 2022; Mao et al., 2022; Zhang et al., 2022; Goyal et al., 2022; Vig et al., 2022; Bahrainian et al., 2022; Liu et al., 2022; Pagnoni et al., 2023; Wang et al., 2023; Urlana et al., 2024). For example, CTRLSum (He et al., 2022) trains models by prepending a keyword sequence to the source document. Similarly, MACSum (Zhang et al., 2023b) adopts prompt learning by prepending each attribute's value to the input using a combination of hard prompts and soft prefixes. HYDRASUM (Goyal et al., 2022) leverages a single encoder, multiple decoder framework with a mixture-of-experts approach, where the decoders share probabilities to effectively control the attributes.

Most controllable summarization research has relied on encoder-decoder frameworks. In addition, these methods often require attribute-specific training or custom datasets to control each attribute, limiting the flexibility of attribute manipulation. Therefore, we propose a generalizable approach that enables flexible attribute control without the need for tailored training, leveraging LLMs for controllable summarization (Tang et al., 2023; Yuan et al., 2024; Liu et al., 2024).

Self-correction Recently, self-correction approaches have been used to refine initial attempts at solving complex problems (Weng et al., 2023; Shinn et al., 2023; Madaan et al., 2023; Dhuliawala et al., 2024; Gou et al., 2024), mirroring human behavior. In summarization tasks, self-correction has been employed to enhance the overall quality of summaries (Zhang et al., 2023a; Sun et al., 2024). Zhang et al. (2023a) utilizes iterative feedback from an evaluator to instruct ChatGPT to produce higher-quality summaries. Unlike prior work, we focus on generating summaries tailored to user preferences, which involve multiple factors to consider.

3 Attribute Measurement and Evaluation Framework for LLM Controllability

We first analyze how each summarization attribute has traditionally been measured and redefine those that were not clearly defined. In particular, we refine linguistic attributes—often measured by word count, using embedding-based similarity. These

Attribute	Metrics	Paper
Extractiveness	ROUGE, word overlap	Goyal et al. (2022); Zhang et al. (2023b)
Length	Absolute length, length ratio	Goyal et al. (2022); He et al. (2022); Maddela et al. (2022); Zhang et al. (2023b)
Topic, Query	ROUGE, LDA, topic word count, classifier	Zhong et al. (2021); He et al. (2022); Zhang et al. (2023b); Xu et al. (2023)
Speaker, Entity	ROUGE, speaker utterance word overlap	Maddela et al. (2022); Zhang et al. (2023b)

Table 1: Previous methods for measuring attributes. they typically relied on word count–based metrics to assess linguistic aspects such as *topic* or *speaker*.

refined measurements allow us to more accurately capture the attributes of generated summaries. Building on this, we propose iterative evaluation metrics to assess the controllability of LLMs—that is, their ability to precisely adjust attributes through multiple rounds of control.

3.1 Revisiting attribute measurements for controllable summarization

We revisit attribute measurement to quantify key attributes for controllable summarization: *extractiveness*, *length*, *topic*, and *speaker*. Table 1 summarizes how previous controllable summarization studies have measured each attribute. However, the measurements for certain attributes have not yet been clearly defined. Thus, we outline our newly defined approach for attribute measurements below.

Extractiveness quantifies the degree of lexical overlap between a summary and its source document. a highly extractive summary is preferred when users need to retain the original context, such as in legal documents, whereas paraphrasing is often favored in general cases. Following the definition of *extractiveness*, we measure the attribute as the proportion of words in the summary directly reused from the source text.

Length refers to the number of words or sentences in the summary or the ratio of the summary's length to that of the original text. By controlling the length, the amount of information in the summary can be adjusted according to user preferences. Prompts used in earlier work often specify a fixed number of sentences (e.g., "3 sentences"), but this approach fails to account for variations in sentence length and does not accurately reflect the summary's actual length (Goyal et al., 2023; Liu et al., 2024; Yuan et al., 2024). Since summary length may vary depending on the complexity of the document (Ryu et al., 2024a), we use the length ratio rather than absolute length in our experiments.

Topic refers to generating a summary centered around one or more themes. Query-focused summarization (QFS), which generates summaries based on a specific query, and entity-based summarization, which focuses on a particular individual, are variations of topic-focused summarization. Zhang et al. (2023b) measured topic word frequency in summaries. Similarly, most QFS methods have relied solely on ROUGE scores, evaluating generated summaries by comparing them to human-annotated references (Zhong et al., 2021). However, even when topic words do not explicitly appear, a summary can still reflect the core context of the topic—especially in LLM-generated summaries, which tend to paraphrase content. Therefore, rather than simply counting word occurrences, we evaluate the semantic similarity between the summary and each topic-related word. We compute the embedding similarity $\mathcal{B}$ between the topic word and each word in the summary s as follows: $\frac{1}{n}\sum_{i\in s}\mathcal{B}(topic, word_i)$, where n is the number of words in the summary. If multiple topics k are present, we use the average embedding similarity across all topics: $\frac{1}{k}\sum_{j\in k}\frac{1}{n}\sum_{i\in s}\mathcal{B}(topic_j, word_i)$.

Speaker refers to generating a summary that focuses on the utterances of a specific speaker within a long document or dialogue. Zhang et al. (2023b) calculate the frequency of the speaker's spoken words appearing in the summary. Similar to *topic* measurement, simply counting the proportion of words from a specific speaker's dialogue included in the summary does not fully capture semantic alignment. Therefore, we extract the speaker's utterances to construct a speaker set $\mathcal{U}$ and leverage BERTScore F1 (Zhang et al., 2020) to compute the embedding similarity between the summary s and $\mathcal{U}$: BERTScore$(s, \mathcal{U})$.

3.2 Iterative controllability evaluation

Building on these refined measurements of summary attributes, we introduce iterative evaluation metrics to assess whether LLMs can iteratively refine and adjust attributes over multiple revisions. Specifically, we introduce two metrics: (1) the *fail-*

	Extractiveness	Length	Topic
Phi-3-medium	100.00% / ↻	100.00% / ↻	38.08% / 0.22
Llama3-8B	100.00% / ↻	100.00% / ↻	57.14% / 0.12
Llama3-70B	49.91% / 8.05	49.36% / 8.24	0.00% / 0.24
GPT-3.5	49.73% / 9.80	76.42% / 0.00	0.00% / 0.00
GPT-4o	39.31% / 6.63	69.84% / 0.00	0.38% / 0.02

Table 2: We evaluate the controllability of LLMs by iteratively testing their ability to accurately adjust specified attributes. The left number represents the averaged control *failure rate*, and the right side denotes the *average iteration count* for successful control.

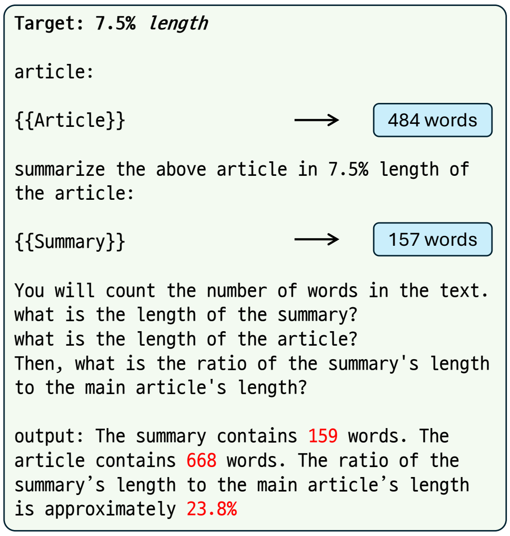

Figure 2: LLMs show notable errors in word count estimation: for an article with 484 words and a summary with 157 words, the model predicts 668 and 159 words, respectively—revealing limitations in self-critique within controllable summarization.

ure rate, proportion of cases in which the model reaches the predefined maximum number of iterations without achieving the desired modifications, and (2) the *average iteration count* required for successful attribute control. We set the maximum number of iterations to 20 due to cost constraints.

Iteration threshold We set attribute-specific thresholds and iteratively regenerate summaries until those thresholds are met. Each attribute is measured using the criteria outlined in Section 3.1 to determine its respective threshold. For *extractiveness* and *length*, we consider control successful if the attribute values fall within ± 5 of the target value. For *topic* and *speaker*, we use the minimum embedding similarity values of the reference summaries in the training dataset as thresholds to determine whether a summary is *topic*-focused or *speaker*-focused. Theses thresholds can be adjusted based on the strictness of the evaluation criteria. The distribution of the datasets used in our experiments is provided in Appendix A.

Label reinterpretation We use the two publicly available MACSum datasets (Zhang et al., 2023b) for controllable summarization. However, existing labels are ambiguous, as the criteria are not numerically defined (e.g., how short must a summary be to qualify as short?). We believe that such ambiguity may confuse LLMs, so we assign clear numerical values to each label. To provide detailed criteria, we reinterpret the labels based on the attribute distributions in each training set (see Appendix A). For *extractiveness*, we define the labels as normal: 85%, high: 90%, and fully: 100%, based on the average values. For the *length* attribute, we follow the annotation criteria of the MACSum dataset—short: 5–10%, normal: 15–25%, and long: 30–35%—and set our target values to short: 7.5%, normal: 15%, and long: 32.5%. Importantly, our method generates summaries based on the specified numerical values, regardless of predefined labels.

4 Analysis on Controllability of LLMs

4.1 Iterative Evaluation on LLMs

As research on leveraging LLMs for controllable summarization remains limited, we evaluate the controllability of various LLMs using the iterative evaluation method described in Section 3.2. We first provide an initial control prompt and generate a summary. If the generated summary fails to meet the specified attribute threshold, the result is fed back into the LLM’s input, prompting it to regenerate until the attribute is correctly controlled. As shown in Table 2, smaller-scale LLMs such as Phi-3-medium (Abdin et al., 2024) and Llama3-8B (Dubey et al., 2024), partially control *topic*, but fail to control *extractiveness* and *length*. Similarly, large-scale LLMs such as Llama3-70B, GPT-3.5 (Brown, 2020), and GPT-4o (Achiam et al., 2023) effectively control *topic*, demonstrating low failure rates. However, they struggle with *extractiveness* and *length*, with failure rates of around 50%. Notably, when initial attempts fail, even GPT-4o is unable to adjust after multiple iterations, ultimately reaching the maximum iteration limit, resulting in an iteration count of zero. These findings suggest that generating summaries while controlling attributes remains challenging for LLMs, even with iterative attempts, especially for numerical attributes such as *extractiveness* and *length*.

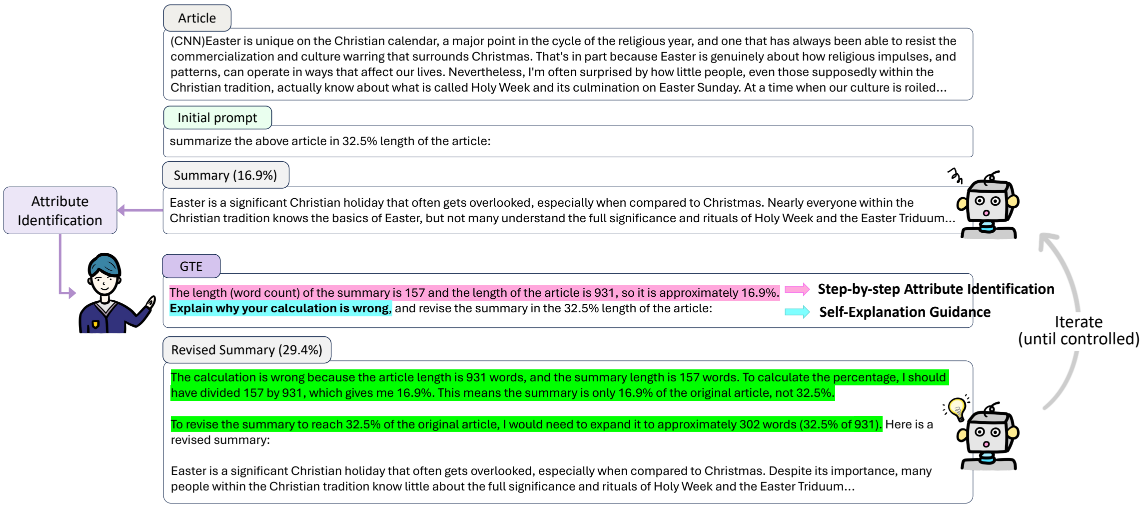

Figure 3: Overview of guide-to-explain system (GTE). The pink parts (■) represent the step-by-step attribute-identification, and the blue parts (■) correspond to the self-explanation guidance.

4.2 Self-critique for controllable summarization

We evaluate whether LLMs can self-adjust summary attributes through self-critique, which has previously improved performance in generic summarization (Zhang et al., 2023a; Sun et al., 2024). As shown in Figure 2, however, unlike in generic summarization tasks, LLMs struggle with measuring attributes. Specifically, they fail to accurately count words in either the source or the summary, making it infeasible for them to revise summaries to match target attribute values on their own.

5 Guide-to-Explain (GTE)

Therefore, we introduce a guide-to-explain (GTE) framework to control attributes via LLMs. As shown in Figure 3, the GTE framework consists of two key phases: step-by-step attribute identification and self-explanation guidance. Since LLMs struggle to reliably measure summary attributes on their own (Figure 2), we explicitly provide the attribute values and teach the model step by step how each attribute should be identified. We then guide the LLM to reflect by explaining the rationale behind its mistakes, enabling it to make appropriate adjustments in subsequent iterations.

5.1 Step-by-step attribute identification

We first instruct the LLM to generate an initial summary s' that reflects the specified attribute. If the LLM fails to control the attributes accurately, we provide step-by-step attribute identification (`SAI`) to guide the model on how to adjust them. Since LLMs struggle with measuring numerical attributes such as *extractiveness* or the *length* ratio, we explicitly instruct the model on how to measure each attribute step by step, enabling it to revise its previously generated summary more precisely.

5.2 Self-explanation guidance

After the identification phase, we provide self-explanation guidance (`SEG`) to the model, guiding the model to explain why it initially failed to control the attributes. This mirrors how humans solve complex problems by reviewing their mistakes to improve future responses. Building on this, in the next iteration, the document (d), initial instruction (i), and previously generated summary (s') are provided as inputs, along with `SAI` and `SEG`. Although LLMs are known to struggle with number-related tasks (Akhtar et al., 2023; Imani et al., 2023), our guidance helps the model effectively control numerical attributes by self-explaining its miscalculations before generating summaries, especially when combined with the step-by-step attribute identification phase. We introduce GTE as a framework that integrates step-by-step attribute identification and self-explanation guidance.

5.3 Overall process

Figure 3 illustrates in detail how the GTE framework operates. By receiving [d; i, s'; `SAI`; `SEG`] as input, the model first reflects on the reasons for its initial error before generating a revised summary. If the revised summary still fails to satisfy the attributes, GTE repeats the process until the model generates an attribute-aligned summary. See

Model	Extractiveness (↓ / ↓)				Length (↓ / ↓)				Topic(↓)
	normal	high	fully	avg	short	normal	long	avg	
Phi-3-medium-Iter	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	38.08% / 0.22
Phi-3-medium-GTE	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	37.97% / 0.04
Llama3-8B-Iter	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	100.00% / ↻	57.14% / 0.12
Llama3-8B-GTE	12.63% / 3.52	11.63% / 2.53	0.00% / 1.46	11.70% / 3.26	26.40% / 3.08	10.92% / 2.26	13.18% / 3.85	14.99% / 2.80	25.56% / 0.91
Llama3-70B-Iter	54.82% / 8.44	37.21% / 7.47	2.70% / 3.78	49.91% / 8.05	18.40% / 6.58	54.61% / 10.42	67.44% / 12.00	49.36% / 8.24	0.00% / 0.24
Llama3-70B-SAI	26.55% / 6.57	18.60% / 7.81	0.00% / 1.86	24.14% / 6.52	4.80% / 5.42	2.73% / 3.81	10.85% / 4.84	5.12% / 4.39	0.00% / 0.10
Llama3-70B-GTE	**0.21%** / 3.28	**0.00%** / 2.83	**0.00%** / 1.50	**0.18%** / 3.22	**0.00%** / 1.10	**0.00%** / 1.61	**2.32%** / 3.14	**0.55%** / 1.90	0.00% / 0.01
GPT-3.5-Iter	45.18% / 9.80	60.47% / 0.00	94.59% / 0.00	49.73% / 9.80	53.60% / 0.00	80.89% / 0.00	88.37% / 0.00	76.42% / 0.00	0.00% / 0.00
GPT-3.5-GTE	17.56% / 3.86	51.16% / 5.00	67.57% / 4.00	23.58% / 3.90	5.60% / 4.63	44.03% / 6.62	78.29% / 7.00	43.33% / 5.95	0.00% / 0.00
GPT-4o-Iter	34.69% / 6.77	55.81% / 0.00	78.38% / 3.00	39.31% / 6.63	72.00% / 0.00	64.85% / 0.00	79.07% / 0.00	69.84% / 0.00	0.38% / 0.02
GPT-4o-SAI	35.12% / 5.50	48.84% / 15.50	62.16% / 6.00	38.03% / 6.13	60.00% / 8.79	61.09% / 9.40	78.29% / 2.00	64.90% / 8.60	0.00% / 0.04
GPT-4o-GTE	**0.00%** / 2.76	**0.00%** / 4.70	**0.00%** / 2.03	**0.00%** / 2.87	**0.00%** / 1.20	**0.00%** / 1.21	**0.00%** / 1.96	**0.00%** / 1.42	0.00% / 0.02

Table 3: The results of controllability measured on the MACSum$_{Doc}$ dataset. Surprisingly, GTE achieves near-zero failure rates across all attributes with only a few iterations. The bold denotes the best performance. Failure or reaching the maximum number of iterations is denoted as ↻.

Model	Extractiveness (↓ / ↓)				Length (↓ / ↓)				Topic (↓)	Speaker (↓)
	normal	high	fully	avg	short	normal	long	avg		
Llama3-70B-Iter	31.78% / 8.13	43.59% / 8.40	8.16% / 5.39	29.63% / 7.59	12.00% / ↻	93.75% / 6.00	98.00% / ↻	81.79% / 6.00	0.00% / 0.01	0.00% / 0.00
Llama3-70B-SAI	14.41% / 5.91	23.08% / 5.31	**0.00%** / 3.72	13.27% / 5.50	**0.00%** / 1.25	62.05% / 5.70	92.00% / 9.33	57.10% / 5.62	0.00% / 0.02	0.00% / 0.00
Llama3-70B-GTE	**0.00%** / 2.31	**0.00%** / 2.56	4.08% / 3.64	**0.61%** / 2.49	**0.00%** / 1.00	**36.61%** / 4.73	**80.00%** / 5.70	**37.65%** / 4.53	0.00% / 0.01	0.00% / 0.00
GPT-4o-Iter	79.24% / 4.36	82.05% 3.67	59.18% / 1.00	76.54% / 4.00	6.00% / ↻	98.21% / ↻	100.00% / ↻	84.26% / ↻	0.31% / 0.01	0.00% / 0.00
GPT-4o-SAI	84.75% / 4.00	87.18% 1.50	53.06% 5.10	80.25% / 4.32	2.00% / 4.50	96.43% / ↻	100.00% / ↻	82.41% / 4.50	0.00% / 0.01	0.00% / 0.00
GPT-4o-GTE	**17.80%** / 7.94	**25.64%** / 7.92	**8.16%** / 4.58	**17.28%** / 7.53	**0.00%** / 1.40	**9.82%** / 2.75	**44.00%** / 4.21	**13.58%** / 2.90	0.00% / 0.02	0.00% / 0.00

Table 4: The results of controllability measured on the MACSum$_{Dial}$ dataset.

Appendix B for the detailed prompts.

6 Experimental setup

We evaluate the controllability of various LLMs, including Phi-3-medium (Abdin et al., 2024), the Llama3 series (Dubey et al., 2024), and the GPT series (Brown, 2020; Achiam et al., 2023). To analyze model performance by size, we utilize both the 8B [1] and quantized 70B versions[2] of Llama3, as well as GPT-3.5 and GPT-4o. We use BERTScore (Zhang et al., 2020) to measure embedding similarity. We used two datasets for our experiments: MACSum$_{Doc}$ and MACSum$_{Dial}$ (Zhang et al., 2023b), which comprise committee meeting transcripts and news content, respectively. Both datasets are designed for mixed-attribute summarization that controls multiple attributes simultaneously. Notably, only MACSum$_{Dial}$ include *speaker* attribute. Since we evaluate LLM performance on individual attributes, we use attributes separately.

7 Results and Discussions

Main results We denote the naive iteration approach, which repeatedly adjusts attributes, as `Iter`. The strategy that provides only step-by-step attribute identification is defined as `SAI`—a stronger version of self-critique that provides the correct attribute values, since LLMs struggle to measure summary attributes on their own. As shown in Table 3, our GTE framework demonstrates remarkably lower failure rates and requires fewer iterations when adjusting summaries across all attributes, including challenging numerical attributes in MACSum$_{Doc}$. Surprisingly, our method reduced the failure rate to nearly 0% when applied to Llama3-70B and GPT-4o, successfully controlling both *extractiveness* and *length* within just 1–3 iterations. For smaller models such as Phi-3-medium and Llama3-8B, which initially exhibited high failure rates, our approach significantly reduced those rates, demonstrating its effectiveness across different model scales. In particular, for long *length*—the most challenging attribute—our method achieved a remarkably low failure rate of just 2.32% within an average of 3.14 iterations.

LLMs encounter greater difficulty with the MACSum$_{Dial}$ dataset (Table 4). The dataset, derived from QMSum (Zhong et al., 2021), consists of lengthy and diverse content from parliamentary and committee meetings, making it more challenging than the CNN-news-based MACSum$_{Doc}$. Notably, length control proved to be the most challenging attribute in MACSum$_{Dial}$. This challenge is likely due to the dataset's origin in long parliamentary transcripts, which makes it inherently difficult to generate summaries of a specific tar-

[1] meta-llama/Meta-Llama-3-8B-Instruct

[2] casperhansen/llama-3-70b-instruct-awq

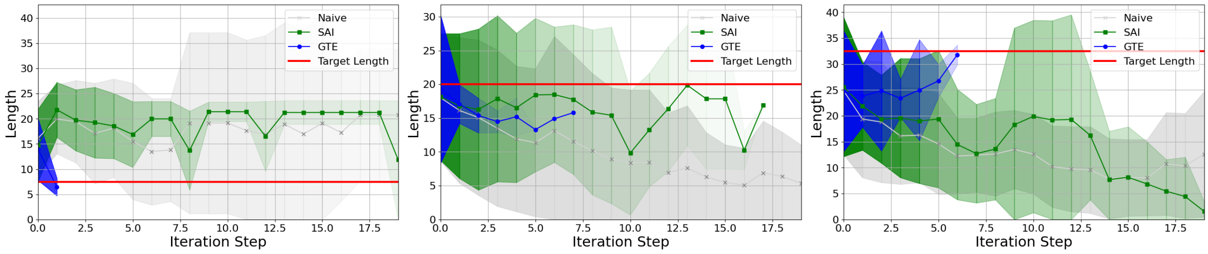

Figure 4: The graphs show how the length ratio changes for each iteration. The intensity of the distribution color is proportional to the number of data points, and the markers represent the average values. The red line indicates the target length, with values of 7.5%, 20%, and 32.5% from left to right.

get length. While the model handled short-length summaries relatively well, difficulty increased significantly as the requested summary length grew. In fact, for long-length summaries, both GPT-4o-`Iter` and GPT-4o-`SAI` showed a 100% failure rate. However, our framework showed meaningfully improved length controllability. With GPT-4o, the average failure rate dropped below 50%. Notably, for normal-length summaries, the failure rate further reduced from over 90% to 9.82%. Regarding *extractiveness*, the `Iter` and `SAI` of GPT-4o exhibit relatively low iteration counts, as the models often exceed the maximum iteration. While their failure rates were close to 80%, our GTE framework achieved a markedly lower failure rate at 17.28% with low iterations, demonstrating the effectiveness of our framework.

Gradual change across iteration steps To analyze how the attribute changes at each step, we track *length* adjustments per iteration (Figure 4). While all methods start with a similar distribution at the initial point, GTE consistently converges within approximately three iterations, maintaining a stable length adjustment pattern across samples. In contrast, `Iter` and `SAI` show inconsistent changes across samples, resulting in higher variance in length adjustments. This demonstrates that our method enables robust attribute control with fewer iterations, regardless of the data sample. For this experiment, we use Llama3-70B and randomly select 110 samples from the MACSum$_{Doc}$ test set.

Attribute types We observe that LLMs control linguistic attributes (*topic* and *speaker*) better than numerical attributes (*extractiveness* and *length*). This aligns with previous research in mathematical reasoning, where LLMs struggle with numerical features (Akhtar et al., 2023), highlighting a broader challenge in precisely handling numerical constraints. From the perspective of the summarization task, *extractiveness* and *length* control the structure of the summary, whereas *topic* and *speaker* influence its content. Our findings suggest that LLMs are proficient at adjusting content to align with user preferences but struggle to generate summaries with specific structural constraints.

Quality of controlled summary We evaluate the quality of summaries generated by GTE. We mainly use UniEval (Zhong et al., 2022) and QuestEval (Scialom et al., 2021), as they correlate highly with human judgments and assess the overall quality of the summary itself. UniEval is a multi-dimensional evaluator that assesses *coherence*, *consistency*, *fluency*, and *relevance* of summaries. QuestEval measures precision and recall by leveraging a question-answering framework to compare the content between the source document and the generated summary without relying on the reference summary. Table 5 shows that our method's summaries outperform across all UniEval dimensions and QuestEval, demonstrating effective attribute control while maintaining overall summary quality. *Relevance* assesses how well a summary retains key information compared to the reference. While `Iter` and `SAI` generate misaligned summaries with lower *relevance* scores, GTE effectively aligns them, resulting in a substantial gain.

Previous studies have shown that ROUGE scores (Lin, 2004) are insufficient for assessing summary quality (Scialom et al., 2021; Zhong et al., 2022; Ryu et al., 2024a). However, since our goal is to control the summary rather than match the reference, we still include ROUGE and BERTScore (Zhang et al., 2020) in our evaluation to provide a more comprehensive assessment. Our framework achieves higher scores than other approaches, demonstrating across various evaluation metrics that GTE not only enhances controllability but also improves overall summary quality.

Model	UniEval					QuestEval	BERTScore	ROUGE-1
	Coherence	Consistency	Fluency	Relevance	Overall			
`Iter` (Ext)	0.820	0.800	0.859	0.696	0.794	0.523	0.826	0.194
`SAI` (Ext)	0.884	0.843	0.905	0.785	0.864	0.554	0.848	0.229
`Iter` (Len)	0.836	0.803	0.836	0.759	0.808	0.484	0.829	0.235
`SAI` (Len)	0.934	0.834	0.942	0.887	0.899	0.548	0.867	0.270
GTE (Ext)	**0.941**	**0.873**	0.937	0.880	**0.908**	**0.590**	0.861	0.236
GTE (Len)	0.937	0.840	**0.944**	**0.901**	0.905	0.553	**0.868**	**0.272**

Table 5: Among the iterative methods, GTE demonstrates both effective attribute control and noticeable improvements in summary quality.

8 Mixed attributes

We extend our evaluation to assess whether LLMs can precisely handle mixed-attribute control. While models manage to control linguistic attributes, they struggle with numerical attributes. Simultaneous control over all attributes remains challenging for all iterative methods, including GTE. Our GTE framework guides LLMs to identify the causes of their errors and regenerate summaries by incorporating this feedback. However, in a mixed attribute setting, the model must process multiple instances of `SAI` and `SEG` for each attribute simultaneously, increasing the cognitive load and making precise control of all attributes more difficult. Therefore, unlike single-attribute evaluation—which assesses whether individual attributes are accurately controlled—we evaluate mixed-attribute control by measuring errors using mean absolute deviation (MAD). This approach compares the differences between the attributes of the generated summary and the requested values, providing a more flexible evaluation of attribute control.

Sequential-planning Recognizing the challenges in precisely controlling all attributes simultaneously, we introduce a sequential planning strategy, *min-planning*, which gradually adjusts attributes—starting with those that are most poorly controlled in the initial draft—using GTE. Figure 5 shows the results comparing single-attribute control with iterations to mixed-attribute control using *min-planning* on the MACSum$_{Doc}$ dataset. We refer to the initial summary in the mixed-attribute control setting as the *mixed-draft*. The *min-planning* method shows a modest improvement in controlling both attributes compared to the *mixed-draft*. However, attributes are still not fully controlled as in single-attribute models, highlighting the difficulty of balancing multiple attributes. We anticipate that modifying one attribute often disrupts previously adjusted attributes due to underlying correlations. For example, even if *length* is adjusted first, it may still change when *extractiveness* is subsequently controlled. Additionally, *min-planning* adjusts each attribute only once without iteration, which may explain its inability to fully control the attributes. A single refinement is often insufficient, whereas GTE iteratively regenerates the summary until the target attribute is successfully adjusted in single-attribute control. Exploring ways for LLMs to control multiple attributes simultaneously would be promising future work.

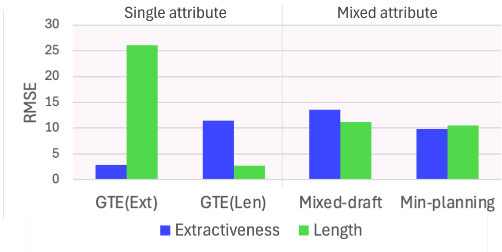

Figure 5: Correlations among attributes hinder LLMs' ability to control them jointly in mixed-attribute setting.

9 Conclusion

In this work, we systematically explore the controllability of LLMs. To this end, we revisit the measurement of summary attributes. We evaluate the controllability of LLMs via iterative assessment and find that they struggle more with numerical attributes than linguistic ones. To address this, we propose a GTE framework, in which the model is guided to explain its misalignments through attribute identification and then uses this explanation to generate better-controlled summaries in subsequent iterations. GTE enables LLMs to control challenging numerical attributes with lower failure rates and fewer iterations. Furthermore, we demonstrate the high quality of controlled summaries using various evaluation metrics.

Limitation

We explore the controllability of various attributes in LLMs and introduced a novel guide-to-explain (GTE) framework to address challenges in numerical attributes. While GTE enhanced successfully control over challenging numerical attributes, it still struggled with highly correlated mixed numerical attributes. Additionally, *min-planning*, which adjusts attributes in order of least alignment, also faced difficulties achieving precise control. Even after properly adjusting one attribute, modifying the correlated numerical attribute caused the previously adjusted attribute to change. We believe further research could explore more effective methods for addressing these challenges.

Ethics

We used publicly available MACSum datasets for our research, conducting experiments with Phi-3, Llama3 [3], GPT-3.5, and GPT-4o from April to October 2024.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506, Dublin, Ireland. Association for Computational Linguistics.

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15391–15405, Singapore. Association for Computational Linguistics.

[3]Meta Llama3 Community License, Copyright © Meta Platforms, Inc. All Rights Reserved. More details can be found at: Llama3 License

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland. Association for Computational Linguistics.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Tanya Goyal, Nazneen Rajani, Wenhao Liu, and Wojciech Kryscinski. 2022. HydraSum: Disentangling style features in text summarization with multi-decoder models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 464–479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tanya Goyal et al. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

Yizhu Liu, Qi Jia, and Kenny Zhu. 2022. Length control in abstractive summarization by pretraining information selection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6885–6895, Dublin, Ireland. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. EntSUM: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Awadallah, and Dragomir Radev. 2022. DYLE: Dynamic latent extraction for abstractive long-input summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1687–1698, Dublin, Ireland. Association for Computational Linguistics.

Dhruv Mehra, Lingjue Xie, Ella Hofmann-Coyle, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2023. EntSUMv2: Dataset, models and evaluation for more abstractive entity-centric summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5538–5547, Singapore. Association for Computational Linguistics.

Artidoro Pagnoni, Alex Fabbri, Wojciech Kryscinski, and Chien-Sheng Wu. 2023. Socratic pretraining: Question-driven pretraining for controllable summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12737–12755, Toronto, Canada. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Lee, and Jungseul Ok. 2024a. Multi-dimensional optimization for text summarization via reinforcement learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5858–5871, Bangkok, Thailand. Association for Computational Linguistics.

Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Geunbae Lee, and Jungseul Ok. 2024b. Key-element-informed sllm tuning for document summarization. In *Interspeech 2024*, pages 1940–1944.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Shichao Sun, Ruifeng Yuan, Ziqiang Cao, Wenjie Li, and Pengfei Liu. 2024. Prompt chaining or stepwise prompt? refinement in text summarization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7551–7558, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023. In-context learning of large language models for controlled dialogue summarization:

A holistic benchmark and empirical analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67, Singapore. Association for Computational Linguistics.

Ashok Urlana, Pruthwik Mishra, Tathagato Roy, and Rahul Mishra. 2024. Controllable text summarization: Unraveling challenges, approaches, and prospects - a survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1603–1623, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468, Seattle, United States. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023. Instructive dialogue summarization with query aggregations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7630–7653, Singapore. Association for Computational Linguistics.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for Computational Linguistics.

Ruochen Xu, Song Wang, Yang Liu, Shuohang Wang, Yichong Xu, Dan Iter, Pengcheng He, Chenguang Zhu, and Michael Zeng. 2023. LMGQS: A large-scale dataset for query-focused summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14764–14776, Singapore. Association for Computational Linguistics.

Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. 2024. Following length constraints in instructions. *Preprint*, arXiv:2406.17744.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. SummIt: Iterative text summarization via ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023b. MACSum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A Attribute details

Table 6 presents the distributions of the MACSum$_{Doc}$ and MACSum$_{Dial}$ training datasets used in our study. For each attribute, we report the distribution of attribute values corresponding to each assigned label, with the average shown in parentheses. For *extractiveness*, both datasets show a wide range of values within each label but exhibit similar average values: around 85% for normal, 90% for high, and 100% for fully. These averages are used as the relabeled target values. For *length*, the observed averages deviate from the annotation guide. In MACSum$_{Doc}$, the means are 4.6% (short), 6.9% (normal), and 13.9% (long), while in MACSum$_{Dial}$, they are 2.0%, 3.7%, and 6.0%, respectively. Due to the small gaps between label means, relabeling based on these values would not sufficiently capture LLM controllability for length. Therefore, we follow the annotation guide and relabel with target values of 7.5% (short), 15% (normal),

Attribute	Label	MACSum$_{Doc}$			MACSum$_{Dial}$		
		Distributions	Relabel	# of summaries	Distributions	Relabel	# of summaries
Extractiveness	normal	35.7 - 100.0% (85.2%)	85.0%	3731	53.2 - 100.0% (86.4%)	85.0%	1661
	high	55.0 - 100.0% (90.0%)	90.0%	287	63.0 - 100.0% (88.9%)	90.0%	340
	fully	84.6 - 100.0% (99.7%)	100.0%	260	75.9 - 100.0% (98.4%)	100.0%	337
Length	short	0.7 - 15.0% (4.8%)	7.5%	1059	0.2 - 20.8% (2.0%)	7.5%	300
	normal	0.5 - 48.6% (6.9%)	20.0%	2194	0.3 - 41.9% (3.7%)	20.0%	1693
	long	1.5 - 39.8% (13.9%)	32.5%	1025	0.7 - 32.4% (6.0%)	32.5%	345
Topic	-	74.8 - 88.8	74.0	2013	73.6 - 87.0	74.0	2317
Speaker	-	-	-	-	75.6 - 92.0	75.0	1796

Table 6: Data distributions of MACSum$_{Doc}$ and MACSum$_{Dial}$.

and 32.5% (long). For *topic*, both datasets show similar scores. We consider summaries with scores above the minimum threshold of 74 to be topic-focused. Similarly, for *speaker*, we use a minimum threshold of 75, derived from the distribution of reference summaries, to define speaker-focused outputs.

B Guide-to-explain (GTE) prompts

Below, we present the prompts and example outputs used for each attribute within the GTE framework.

Article

(CNN)Easter is unique on the Christian calendar, a major point in the cycle of the religious year, and one that has always been able to resist the commercialization and culture warring that surrounds Christmas. That's in part because Easter is genuinely about how religious impulses, and patterns, can operate in ways that affect our lives. Nevertheless, I'm often surprised by how little people, even those supposedly within the Christian tradition, actually know about what is called Holy Week and its culmination on Easter Sunday. At a time when our culture is roiled by questions of identity and ethics (and tolerance) that have profound religious implications, it's worth pausing to explore this crucial holiday -- and the awareness of the human condition, in all its sadness and glory, that it engenders. After all, Holy Week calls mostly to those who incline their minds and hearts in its direction with seriousness of intent..

Initial prompt

summarize the above article in 32.5% length of the article:

Summary (16.9%)

Easter is a significant Christian holiday that often gets overlooked, especially when compared to Christmas. Nearly everyone within the Christian tradition knows the basics of Easter, but not many understand the full significance and rituals of Holy Week and the Easter Triduum, which are three days of prayer that begin on Good Friday and end on Easter Sunday. Holy Communion originated from the Last Supper, a Passover meal that Jesus shared with his disciples before being crucified. On Maundy Thursday, Jesus washed the disciples' feet to demonstrate humility...

GTE

The length (word count) of the summary is 157 and the length of the article is 931, so it is approximately 16.9%.
Explain why your calculation is wrong, and revise the summary in the 32.5% length of the article:

Revised Summary (29.4%)

The calculation is wrong because the article length is 931 words, and the summary length is 157 words. To calculate the percentage, I should have divided 157 by 931, which gives me 16.9%. This means the summary is only 16.9% of the original article, not 32.5%.

To revise the summary to reach 32.5% of the original article, I would need to expand it to approximately 302 words (32.5% of 931). Here is a revised summary:

Easter is a significant Christian holiday that often gets overlooked, especially when compared to Christmas. Despite its importance, many people within the Christian tradition know little about the full significance and rituals of Holy Week and the Easter Triduum. This three-day period of prayer begins on Good Friday, commemorating Jesus' crucifixion, and concludes on Easter Sunday, celebrating his resurrection...

Figure 6: Length guide-to-explain (GTE).

Article

(CNN)Easter is unique on the Christian calendar, a major point in the cycle of the religious year, and one that has always been able to resist the commercialization and culture warring that surrounds Christmas. That's in part because Easter is genuinely about how religious impulses, and patterns, can operate in ways that affect our lives. Nevertheless, I'm often surprised by how little people, even those supposedly within the Christian tradition, actually know about what is called Holy Week and its culmination on Easter Sunday. At a time when our culture is roiled by questions of identity and ethics (and tolerance) that have profound religious implications, it's worth pausing to explore this crucial holiday -- and the awareness of the human condition, in all its sadness and glory, that it engenders. After all, Holy Week calls mostly to those who incline their minds and hearts in its direction with seriousness of intent. Still, the fuss must puzzle those looking on, wondering what it all means. Why do Christians make so much of this springtime week, and make so much of Easter weekend? There is a phrase that many never come across, even among Christians: Easter Triduum .This refers to the three days of Easter that begin with Good Friday, proceed through Holy Saturday, and conclude with Easter Sunday. It's definitely a progression, although the word itself -- triduum -- can refer to any three days of prayer...

Initial prompt

summarize the above article in 32.5% length of the article:

Summary (16.9%)

Easter is a significant Christian holiday that often gets overlooked, especially when compared to Christmas. Nearly everyone within the Christian tradition knows the basics of Easter, but not many understand the full significance and rituals of Holy Week and the Easter Triduum, which are three days of prayer that begin on Good Friday and end on Easter Sunday. Holy Communion originated from the Last Supper, a Passover meal that Jesus shared with his disciples before being crucified. On Maundy Thursday, Jesus washed the disciples' feet to demonstrate humility...

SAI

The length (word count) of the summary is 157 and the length of the article is 931, so it is approximately 16.9%.
Summarize the above article in the 32.5% length of the article:

Revised Summary (23.4%)

Easter, a significant Christian holiday, remains untouched by commercialization and culture wars. Many people, including Christians, are unfamiliar with Holy Week and its culmination on Easter Sunday. The Easter Triduum, a three-day period, begins with Maundy Thursday, where Jesus shared the Last Supper with his disciples, demonstrating his humility by washing their feet. Good Friday, the day of the crucifixion, is a somber reminder of death, sacrifice, and fear. The Gospels provide varying accounts of Jesus' final words, including his quote from the 22nd Psalm. Holy Saturday, often overlooked, serves as a passage between the darkness of Good Friday and the hope of Easter Sunday. It's a day of anxiety and uncertainty, when the pain of the crucifixion still lingers, but hope begins to emerge. Easter Sunday, with its promise of transformation and resurrection, is a celebration of the human experience. The resurrection is not just a physical revival but a total transformation, symbolizing the possibility of change and renewal in our own lives...

Figure 7: Length attribute-identification (SAI).

Article

(CNN)Easter is unique on the Christian calendar, a major point in the cycle of the religious year, and one that has always been able to resist the commercialization and culture warring that surrounds Christmas. That's in part because Easter is genuinely about how religious impulses, and patterns, can operate in ways that affect our lives. Nevertheless, I'm often surprised by how little people, even those supposedly within the Christian tradition, actually know about what is called Holy Week and its culmination on Easter Sunday. At a time when our culture is roiled by questions of identity and ethics (and tolerance) that have profound religious implications, it's worth pausing to explore this crucial holiday -- and the awareness of the human condition, in all its sadness and glory, that it engenders. After all, Holy Week calls mostly to those who incline their minds and hearts in its direction with seriousness of intent. Still, the fuss must puzzle those looking on, wondering what it all means. Why do Christians make so much of this springtime week, and make so much of Easter weekend? There is a phrase that many never come across, even among Christians: Easter Triduum .This refers to the three days of Easter that begin with Good Friday, proceed through Holy Saturday, and conclude with Easter Sunday. It's definitely a progression, although the word itself -- triduum -- can refer to any three days of prayer...

Initial prompt

summarize the above article in 32.5% length of the article:

Summary (16.9%)

Easter is a significant Christian holiday that often gets overlooked, especially when compared to Christmas. Nearly everyone within the Christian tradition knows the basics of Easter, but not many understand the full significance and rituals of Holy Week and the Easter Triduum, which are three days of prayer that begin on Good Friday and end on Easter Sunday. Holy Communion originated from the Last Supper, a Passover meal that Jesus shared with his disciples before being crucified. On Maundy Thursday, Jesus washed the disciples' feet to demonstrate humility...

Iter

summarize the above article in 32.5% length of the article:

Revised Summary (9.1%)

Easter is a unique Christian event that resisted commercialization. Many Christians lack understanding of Holy Week and Easter. The Easter Triduum - Good Friday, Holy Saturday, and Easter Sunday - marks death, transition, and resurrection. This pattern reflects our daily struggles, suffering, and rebirth, making Easter a celebration of moving from darkness to light.

Figure 8: Length iteration (`Iter`).

Article

(CNN)We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. We created a culture of such deep distrust and disrespect that violence seemed the inevitable response. We let the violence flow. We let the violence stand for everything that's wrong with the things we already didn't like. By now you may be asking, "Who's we? "You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People. Yes we can. We are family. I use "we" a lot when I talk about our country's achievements. I like to say we won the Second World War, we put a man on the moon, we invented the Internet, we gave the world jazz. Well, if I -- a son of immigrants whose family had nothing to do with any of those accomplishments -- if I get to claim those aspects of American history, then surely I have to claim the unsavory aspects too. "We" cuts both ways. We enslaved Africans. We cut Reconstruction short and made a mockery of equal citizenship. We supported Jim Crow, then redlined, subordinated, and ghettoized African-Americans...

Initial prompt

summarize the above article using 85% of words directly from the original document:

Summary (100.0%)

We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. By now you may be asking, "Who's we?" You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People...

GTE

This summary contains 100.0% of words directly from the original document.
Explain why your calculation is wrong, and revise the summary so that 85% of the words are used directly as in the original document:

Revised Summary (86.8%)

I apologize for the mistake. Upon re-calculating, I realized that the original summary contained 79.4% of words directly from the original document. To revise the summary to meet the 85% threshold, I'll incorporate more words directly from the original document. Here's the revised summary:

"We did it again, in another American city", perpetuating a cycle of violence and chaos. "We set Baltimore on fire this time", and in doing so, "we brutalized black bodies". The author emphasizes the significance of "we", underscoring collective responsibility for both accomplishments and injustices. "We the People" implies a shared history, encompassing achievements like "we won the Second World War", "we put a man on the moon", "we invented the Internet", and "we gave the world jazz", as well as darker aspects like "we enslaved Africans", "we cut Reconstruction short and made a mockery of equal citizenship", "we supported Jim Crow, then redlined, subordinated, and ghettoized African-Americans", and "we failed"...

Figure 9: Extractiveness guide-to-explain (GTE).

Article

(CNN)We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. We created a culture of such deep distrust and disrespect that violence seemed the inevitable response. We let the violence flow. We let the violence stand for everything that's wrong with the things we already didn't like. By now you may be asking, "Who's we? "You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People. Yes we can. We are family. I use "we" a lot when I talk about our country's achievements. I like to say we won the Second World War, we put a man on the moon, we invented the Internet, we gave the world jazz. Well, if I -- a son of immigrants whose family had nothing to do with any of those accomplishments -- if I get to claim those aspects of American history, then surely I have to claim the unsavory aspects too. "We" cuts both ways. We enslaved Africans. We cut Reconstruction short and made a mockery of equal citizenship. We supported Jim Crow, then redlined, subordinated, and ghettoized African-Americans...

Initial prompt

summarize the above article using 85% of words directly from the original document:

Summary (100.0%)

We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. By now you may be asking, "Who's we?" You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People...

SAI

This summary contains 100.0% of words directly from the original document.
Summarize the above article using 85% of the words directly from the original document:

Revised Summary (78.0%)

Here's a rewritten summary that meets the 85% requirement:

We allowed Baltimore to burn, brutalizing black bodies and turning a funeral into a riot. We created chaos, replaying the images and fostering a culture of distrust and disrespect. When asked "Who's we?", some may object, saying "Don't lump me in with them." But the truth is, we as a nation are responsible. The word "we" is synonymous with American identity. If we claim America's achievements, we must also claim its dark past, including slavery and racial inequity. We are the authors of Baltimore's story, and it's our responsibility to address the problem. We can't just point to individual faults or structural flaws; we must acknowledge both. It's time to put aside our differences, acknowledge the truth, and work together to create opportunity and change.

Figure 10: Extractiveness attribute-identification (`SAI`).

Article

(CNN)We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. We created a culture of such deep distrust and disrespect that violence seemed the inevitable response. We let the violence flow. We let the violence stand for everything that's wrong with the things we already didn't like. By now you may be asking, "Who's we? "You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People. Yes we can. We are family. I use "we" a lot when I talk about our country's achievements. I like to say we won the Second World War, we put a man on the moon, we invented the Internet, we gave the world jazz. Well, if I -- a son of immigrants whose family had nothing to do with any of those accomplishments -- if I get to claim those aspects of American history, then surely I have to claim the unsavory aspects too. "We" cuts both ways. We enslaved Africans. We cut Reconstruction short and made a mockery of equal citizenship. We supported Jim Crow, then redlined, subordinated, and ghettoized African-Americans...

Initial prompt

summarize the above article using 85% of words directly from the original document:

Summary (100.0%)

We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. By now you may be asking, "Who's we?" You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People...

Iter

Summarize the above article using 85% of the words directly from the original document:

Revised Summary (100.0%)

We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. The word "we" is one of the great American words. We the People. I use "we" a lot when I talk about our country's achievements. Well, if I get to claim those aspects of American history, then surely I have to claim the unsavory aspects too. "We" cuts both ways. We enslaved Africans. We cut Reconstruction short and made a mockery of equal citizenship. We supported Jim Crow, then redlined, subordinated, and ghettoized African-Americans. We cut blacks out of the New Deal. We created a polity in which racial inequity and economic inequality magnify each other unrelentingly. We tried to put a lid on it with heavy policing and a War on Drugs. We failed...

Figure 11: Extractiveness iteration (`Iter`).

Article

(CNN)Prison life won't be pretty for Aaron Hernandez, the former NFL player and convicted murderer sentenced to life without parole.After correction officers evaluate him, he will be shipped to Massachusetts' flagship maximum-security prison, one of the most high-tech jails in the United States with no history of breakouts: the Souza-Baranowski Correctional Center in Shirley, about 40 miles outside downtown Boston.It's called Souza, for short, and it's the state's newest prison, opened in 1998, with a matrix of 366 cameras recording live 24 hours a day and a microwave detection perimeter withtaut wire."I don't know the date, but he'll be going there.That's the maximum-security facility," Department of Corrections spokesman Darren Duarte said.Legal advocates for inmates describe Souza as sterile and violent at once.Its diverse demographic includes the young and the old, many of whom are also doing life.One stubborn problem is that opiates are smuggled to inmates, the legal advocates said."It's very shiny and clean looking and very sterile," said Leslie Walker, executive director of Prisoners' Legal Services of Massachusetts, who has been visiting the Souza prison about every six weeks for the past 15 years and serves indigent prisoners there.But, she added: "It is a very dangerous prison that is right now experiencing a veritable flood of opiates."...

Initial prompt

summarize the above article focusing on hernandez:

Summary (72.6)

We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. By now you may be asking, "Who's we?" You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People...

GTE

The similarity between this summary and the topic hernandez is 72.6, so this summary is not centered around the topic hernandez.
Explain why you were unable to summarize around the topic, and revise the summary to center around the topic hernandez:

Revised Summary (76.7%)

+I was unable to summarize around the topic hernandez because the original article provided a lot of context about the prison system and the conditions that hernandez will face, but did not provide much information that is specific to hernandez himself. The article does not delve into hernandez's thoughts, feelings, or personal experiences, and instead focuses on the general details of his situation.

Here is a revised summary that centers around the topic hernandez:

Aaron Hernandez, the former NFL player and convicted murderer, is facing a harsh reality as he begins his life sentence without parole. After being convicted, Hernandez was immediately transferred to the Massachusetts Correctional Institution-Cedar Junction in Walpole, where he will be processed before being sent to the Souza-Baranowski Correctional Center in Shirley. As a high-profile inmate, Hernandez's safety will be a concern for prison officials, who may need to take measures to protect him from other inmates or gang members. Despite his celebrity status, Hernandez will face harsh conditions, including being locked in his cell for 19 hours a day.

Figure 12: Topic guide-to-explain (GTE).