CCSBench: Evaluating Compositional Controllability in LLMs for Scientific Document Summarization

Yixi Ding
National University of Singapore
yixi.d@comp.nus.edu.sg

Jiaying Wu*
National University of Singapore
jiayingw@nus.edu.sg

Tongyao Zhu
National University of Singapore
tongyao.zhu@u.nus.edu

Yanxia Qin
Singapore University of Technology and Design
yanxia_qin@sutd.edu.sg

Qian Liu
Sea AI Lab
liuqian@sea.com

Min-Yen Kan*
National University of Singapore
kanmy@comp.nus.edu.sg

Abstract

To broaden the dissemination of scientific knowledge to diverse audiences, it is desirable for scientific document summarization systems to simultaneously control multiple attributes such as length and empirical focus. However, existing research typically focuses on controlling single attributes, leaving the *compositional control* of multiple attributes underexplored. To address this gap, we introduce CCSBench, the first evaluation benchmark for compositional controllable summarization in the scientific domain. Our benchmark enables fine-grained control over both *explicit* attributes (e.g., length), which are objective and straightforward, and *implicit* attributes (e.g., conceptual or empirical focus), which are more subjective and abstract. We conduct extensive experiments using various large language models (LLMs) under various settings, including in-context learning, parameter-efficient fine-tuning, and two-stage modular methods for balancing control over different attributes. Our findings reveal significant limitations in LLMs capabilities in balancing trade-offs between control attributes, especially implicit ones that require deeper understanding and abstract reasoning. [1]

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Information systems** → **Summarization**.

Keywords

Abstractive Summarization, Scientific Document Summarization, Large Language Models

ACM Reference Format:
Yixi Ding, Jiaying Wu, Tongyao Zhu, Yanxia Qin, Qian Liu, and Min-Yen Kan. 2018. CCSBench: Evaluating Compositional Controllability in LLMs for Scientific Document Summarization. In *Proceedings of KDD'25 SciSoc LLM Workshop: Large Language Models for Scientific and Societal Advances (KDD'25 SciSoc LLM Workshop).* ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding authors.
[1]Data is available at: https://huggingface.co/datasets/dyxohjl666/CCSBench.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD'25 SciSoc LLM Workshop, Toronto, ON, Canada
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

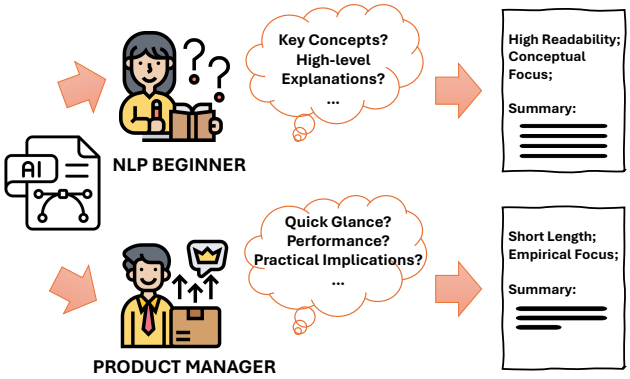

Figure 1: Compositional controllability allows scientific summaries to be tailored to diverse reader needs, addressing specific reader requirements.

1 Introduction

Compositional controllability – the ability to generate tailored summaries based on multiple attributes – is essential for scientific document summarization systems to effectively communicate research to diverse audiences. Scientific papers contain complex information, but different readers have distinct needs and preferences when engaging with this content. For instance, as illustrated in Figure 1, an NLP beginner may seek a clear, concept-focused explanation, while a product manager may prioritize a concise summary highlighting the model's empirical performance. By dynamically adjusting summaries to align with varying expertise levels and interests, compositional control ensures that scientific knowledge is more accessible, relevant, and engaging for a broader audience.

Despite the broad demand for compositional controllability in scientific summarization, this capability remains largely underexplored in controllable text generation research. Most current efforts focus on using large language models (LLMs) to generate non-academic texts such as movie reviews [23] and news articles [4], which are less complex than scientific documents adhering to strict academic conventions. Preliminary investigations into compositional scientific summarization [8] address only straightforward

control requirements such as length and keyword inclusion, overlooking more subtle user requirements such as focus on empirical aspects. Developing methods for real-world deployment requires benchmarks that not only incorporate more sophisticated control attributes but also capture their interplay in a truly compositional manner, a challenge that remains unresolved.

In this paper, we introduce CCSBench,the first evaluation benchmark for compositional scientific summarization that enables fine-grained control over multiple attributes. Drawing inspiration from Kahneman's cognitive theory [18], which distinguishes between the fast, intuitive "System 1" and the slow, deliberate "System 2" modes of thinking, CCSBench incorporates both **explicit** and **implicit** control attributes. Explicit attributes (e.g., length) are objective, easily quantifiable, and straightforward for both LLMs and humans to process. In contrast, implicit attributes (e.g., empirical focus) require deeper reasoning and human-like understanding, making them more challenging for LLMs to control effectively.

As illustrated in Figure 2, CCSBench is structured around two explicit attributes: (1) *length* and (2) *keyword inclusion*, and two implicit attributes: (3) *readability* and (4) *focus*. *Length* sets the target word count for the summary, while *keyword inclusion* ensures the presence of specific terms in the output. *Readability* controls language complexity and accessibility, tailoring the summary for technical or general audiences. *Focus*, short for "Empirical Focus Level", adjusts the emphasis on empirical evidence versus conceptual aspects, requiring deeper contextual understanding and abstract reasoning. By dynamically adjusting these attributes, we curate diverse scientific summaries – such as concise, highly readable conceptual explanations – that meet real-world needs and cater to diverse reader expectations.

We evaluate a diverse set of LLMs on CCSBench, including both proprietary LLMs and fine-tuned open-source variants. Results reveal significant limitations in managing trade-offs between control attributes, particularly due to inadequate abstract reasoning for implicit attributes. In parameter-efficient fine-tuning (PEFT), decoder-only models like LLaMA2 [32] and Mistral [17] struggle with long-term dependencies, while encoder-decoder models like Flan-T5 [5] adapt more effectively. Further analysis shows that standard task composition methods, such as AdapterFusion [28], are unsuitable due to oversimplified adapter aggregation. Our findings suggest the need for more targeted research on compositionality in scientific document summarization.

2 Related Work

2.1 Controllable Summarization

Controllable summarization has attracted significant research interest, with efforts focused on guiding summaries using attributes such as length, topic, keywords, and sentiment. Early work explored the impact of oracle-provided attributes on summary quality [10, 29]. More recent studies have aimed at controlling individual attributes like length or keywords [3, 13], but these methods often treat attributes independently, limiting their scalability and effectiveness in multi-attribute scenarios. MacSum [36] represents a step forward by supporting multi-attribute control in the news and dialogue domains, where the content is relatively less technical. In contrast, scientific texts pose unique challenges for controllable summarization due to their high degree of linguistic precision and domain-specific complexity. CocoScisum [8] extends control to both length and keywords within scientific papers but remains constrained to explicitly defined attributes. Our work, CCSBench, advances the field by enabling fine-grained control over both explicit and implicit attributes, offering a broader and more nuanced evaluation framework tailored to scientific summarization.

2.2 LLMs for Scientific Document Processing

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of scientific document processing tasks, including idea generation [30], literature review synthesis [34], research critique [9], and the production of scientific news articles [22]. These models also show promise in controllable text generation [33, 35, 37], though existing work primarily evaluates them under single-constraint settings (e.g., adjusting formality). Despite the growing need to tailor scientific content for diverse audiences, the ability of LLMs to perform compositional controllable generation – handling multiple constraints simultaneously – remains underexplored. To address this gap, we introduce the CCSBench benchmark, designed to systematically evaluate LLM performance under multi-attribute control. Our findings highlight the limitations of current models, particularly in handling implicit, conceptual attributes such as readability and aspect focus.

3 CCSBench Data Curation

3.1 Task Formulation

We formally introduce the task and notation used throughout the paper. Each example consists of a scientific document D accompanied by a set of control signals $C = \{len, kw, focus, read\}$, where len refers to length, kw to keywords, $focus$ to empirical focus level, and $read$ to readability. The objective is to generate a summary S that adheres to the constraints imposed by C while maintaining narrative coherence and preserving key information from the original document.

We define the four control attributes for CCSBench as follows:

Length. specifies the desired number of words in the summary, ranging from concise to detailed, and catering to different audience needs based on their familiarity with the topic or time constraints. Length is divided into five bins, each representing a 50-word range (e.g., Bin 0: 0-50 words; Bin 1: 51-100 words).

Keywords. ensures the inclusion of key terms from the source document, enabling readers to quickly grasp the core elements and assess the paper's relevance. Keywords can be left empty or set as pre-defined terms to be included in the summary.

Readability. adjusts syntactic complexity and vocabulary to match audience proficiency, making summaries more accessible or technical. This enhances comprehension and engagement. It has two levels: [normal, high].

Empirical Focus Level (Focus). adjusts emphasis between empirical results and theoretical contributions, with two levels: [low, high]. A high empirical focus highlights data-driven aspects like experiments and results, while a low empirical focus emphasizes

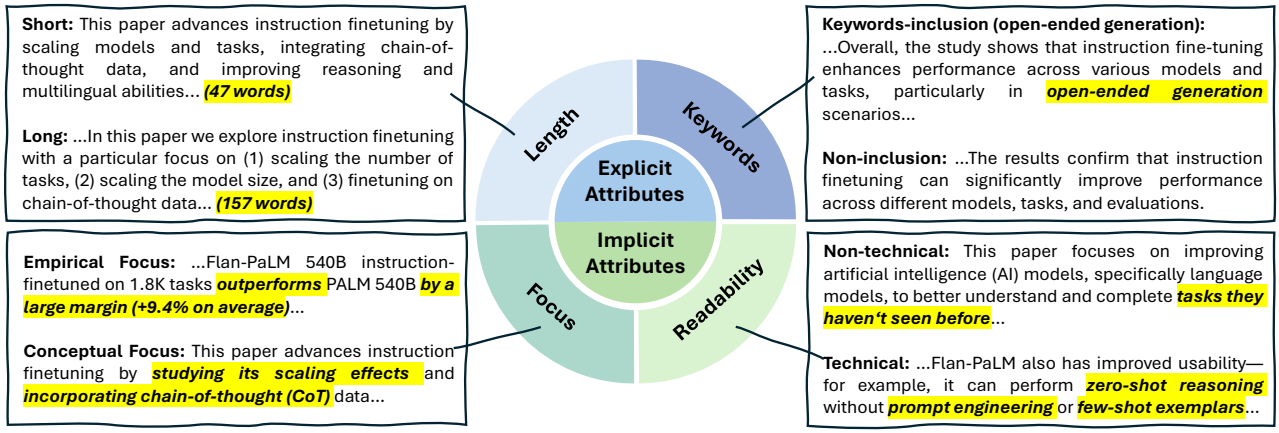

Figure 2: Illustration of explicit and implicit control attributes in CCSBench.

Table 1: CCSBench statistics.

Dataset	#Samples / #Docs		
	Train	Val	Test
Length	2,400 / 1,561	800 / 705	800 / 698
Keywords	2,029 / 2,029	677 / 677	677 / 677
Readability	2,400 / 1,687	800 / 724	800 / 728
Focus	2,332 / 1,659	779 / 709	768 / 703
Compositional	2,400 / 1,590	800 / 655	758 / 364

theories, frameworks, and broader implications. This control tailors summaries to readers seeking practical insights or conceptual understanding.

3.2 Dataset

We construct CCSBench based on the arXiv dataset [6], a comprehensive collection of scientific documents from the widely used online preprint repository, arXiv. CCSBench focuses on papers in Computer Science and Artificial Intelligence. To ensure high-quality input with minimal noise, we use the introduction and conclusion sections (**I+C**) as the input text, a practice validated by prior research [2, 25] and our empirical analysis in Section 7. Following established practices in scientific summarization research [7], we use abstracts as reference summaries, leveraging their widely recognized factuality and comprehensiveness [31].

The benchmark comprises four single-attribute control datasets and one compositional control dataset, each split into training, validation, and test sets using a 60/20/20 ratio. Dataset statistics are summarized in Table 1.

3.3 Dataset Construction

The construction of CCSBench involves two main steps: (1) construction of single-attribute control datasets, followed by (2) the creation of a compositional control dataset.

Single Control Dataset. We begin with KWX [11], a dataset that provides keywords for each scientific document sourced from arXiv. From this, we select 3,700 papers in the subjects of Computer Science and Artificial Intelligence to form the foundation of our controllable summarization dataset, using their abstracts as natural reference summaries.

The **Keywords** dataset, $\mathcal{D}_{kw}$, is created by filtering out any keywords from the KWX dataset that do not appear in the abstracts, ensuring high-quality keywords.

To address the challenge of automatically assessing readability and focus, we use GPT-4 [27] to assist in generating summaries related to these two control attributes. We sample 2,000 papers from our data source, and assign the original abstracts with "normal readability". GPT-4 rewrites each abstract into a more layman-friendly version while maintaining the original meaning. These rewritten abstracts are labeled as "high readability". Combining these two types of abstracts, we form the **Readability** dataset, $\mathcal{D}_{read}$.

For focus control, we task GPT-4 with identifying the conceptual and empirical components of each abstract, then rewriting them separately without altering their meaning. Summaries derived from conceptual components are labeled as having "low empirical focus," while those from empirical components are labeled as having "high empirical focus". These summaries form the **Focus** dataset, $\mathcal{D}_{focus}$.

To construct the **Length** dataset, $\mathcal{D}_{len}$, we sample 4,000 instances from $\mathcal{D}_{Read}$ and D_{focus}. Word counts are rounded to the nearest 50-word bin to create a length-controllable dataset.

Compositional Control Dataset. To manage the complexity of controlling all four attributes simultaneously, we construct a compositional control dataset using $\mathcal{D}_{read}$ and $\mathcal{D}_{focus}$, denoted as $\mathcal{D}_{CC}$. Since readability and empirical focus are already annotated, and keywords and length can be easily derived as described earlier, each instance in this dataset is tagged with 2-3 attributes. This allows us to create both training and validation sets with compositional control over multiple attributes.

While it is impractical to collect instances covering every attribute combination in the training set, we ensure comprehensive evaluation by constructing a test set that includes all four attributes. We sample 400 instances from D_{focus}, where the length, keywords, and empirical focus levels are pre-determined. Following

Table 2: Evaluation results validate the high quality of LLM-generated summaries in CCSBench.

Dataset	FAC	FLU	CTRL	ROUGE-1	ROUGE-2	ROUGE-L
Readability	0.98	0.93	0.92	53.88	21.81	41.31
Focus	0.91	0.93	0.93	51.25	34.23	41.41

the method used to generate D_{read}, we use GPT-4 to rewrite each summary for high readability, ensuring accessibility for a layman audience.

3.4 Data Quality Validation

We assess the quality of GPT-4-generated summaries in CCSBench through automated and human evaluations. First, we compute ROUGE scores to assess fluency and content accuracy by comparing generated summaries with their original counterparts. Next, we conduct a human evaluation to further verify the summaries' controllability and overall quality.

We sample 120 documents each from D_{read} and D_{focus} and recruit Amazon Mechanical Turk[2] annotators for binary assessments on three aspects: (1) **Factual Consistency (FAC)**: whether the summary accurately reflects key points, (2) **Fluency (FLU)**: whether it is natural and well-written, and (3) **Controllability (CTRL)**: for Readability, whether the summary is easier to read than the reference; for Focus, whether empirically-focused summaries emphasize data collection, experimental setup, and results over conceptual aspects (details in Appendix B).

Table 2 shows high accuracy scores (>0.9) across all metrics, confirming the quality of CCSBench 's summaries. Rigorous annotator selection and strong inter-annotator agreement (>0.85, Section 3.5) further validate these results.

3.5 Reliability of Human Evaluation

We implement rigorous measures to ensure the reliability of human evaluations. Annotators on Amazon Mechanical Turk are selected based on strict quality criteria, including a $\geq$ 98% HIT approval rate and a minimum of 500 completed HITs, ensuring experienced and reliable contributors.

To validate annotation reliability, we conduct a pilot study with 20 randomly sampled instances for readability- and focus-controlled summaries. Each summary is evaluated by two independent annotators, and inter-annotator agreement is calculated. As shown in Table 3, all agreement scores exceed 0.85, confirming evaluation consistency. Based on these reliable pilot results, we scale the evaluation to two larger sets of 120 randomly sampled instances, each assessed by a single annotator, as shown in Section 3.4.

Table 3: Inter-annotator agreement scores for assessment questions on readability (Appendix B.1) and empirical focus (Appendix B.2).

	Q1	Q2	Q3	Q4	Q5
Readability	1.00	0.85	0.95	0.90	-
Focus	0.95	1.00	0.85	0.90	0.95

[2]https://www.mturk.com/

4 Experimental Setup

4.1 Models

We evaluate a range of closed-source and open-source models for compositional controllable summarization, including GPT-3.5 [26], GPT-4 [27], GPT-4o [16], LLaMA2 [32], Mistral [17], and Flan-T5 [5]. For closed-source models, we conduct few-shot experiments by randomly selecting three demonstrations per test sample. For open-source models, we explore parameter-efficient fine-tuning (PEFT) using Low-Rank Adaptation (LoRA)[14] to adapt them to our dataset. Additionally, we revisit controllable summarization baselines from the news domain, including hard prompt (HP)[36] and soft prefix tuning (SP)[20]. Further implementation details, including training setups, are provided in Appendix A.

4.2 Evaluation Metrics

We evaluate all models based on both summarization quality and attribute controllability. We employ ROUGE [21] to evaluate the overall quality. To assess attribute controllability, we employ the following metrics tailored to each attribute.

Length Control. We measure length controllability using the Mean Absolute Deviation (MAD) [24], and the Pearson Correlation Coefficient (PCC) [24], to evaluate the distance between the target and generated lengths.

Keywords Control. Keyword control is evaluated using the Success Rate (SR) [13], which measures the fraction of specified keywords present in the generated summaries through exact matching after stemming.

Readability Control. For readability, we calculate the Flesch-Kincaid Grade Level (FKGL) [19], where a lower score indicates higher readability. The difference in FKGL scores (δ_{FKGL}) between the high readability setting (F_{high}) and the normal readability setting (F_{normal}) reflects the model's ability to control this attribute. A higher δ_{FKGL} is desirable.

Empirical Focus Control. For empirical focus, we use GPT-4 to classify summaries into high or low empirical focus levels. Validated with 93% accuracy on a manually labeled dataset, GPT-4 predictions are used to compute the F1 score for each category, assessing the model's focus control. Prompt details are provided in Appendix C.

5 Experiments

In this section, we benchmark and compare the compositional controllability of LLMs on CCSBench, focusing on four control attributes, which are categorized into explicit control (length, keywords) and implicit control (focus, readability). ***As we focus on evaluating LLM's capabilities for compositional control, all experiments are conducted on the Compositional dataset ($\mathcal{D}_{CC}$) of CCSBench, unless otherwise specified.***

Table 4: Performance comparison of representative language models on CCSBench, evaluating overall quality and four control attributes. Arch refers to model architectures, with ED for encoder-decoder and DO for decoder-only. Bold numbers indicate the best performance, while underlined numbers represent the second-best.

	Method	Arch	Quality	Explicit Attributes			Implicit Attributes			
				Length		Keywords	Readability			Focus
			ROUGE-L ↑	PCC ↑	MAD ↓	SR ↑	F_{normal}	F_{high}↓	δ_{FKGL} ↑	F1 ↑
Zero-Shot LLMs										
A1	Flan-T5-XL	ED	13.49	0.00	1.43	0.44	13.94	13.57	0.37	0.36
A2	Flan-T5-XXL	ED	14.25	0.00	1.39	0.46	12.66	12.68	-0.02	0.39
A3	LLaMA2-7B	DO	17.10	0.23	1.11	0.75	14.40	14.44	-0.04	0.51
A4	Mistral-7B	DO	17.67	0.13	1.16	0.75	12.08	12.44	-0.36	0.48
A5	GPT-3.5	DO	18.05	0.62	0.39	0.95	15.12	15.06	0.06	0.56
A6	GPT-4	DO	19.26	0.84	0.66	**0.99**	19.12	16.99	2.13	0.71
A7	GPT-4o	DO	**19.41**	**0.94**	**0.09**	**0.99**	16.26	15.38	0.88	0.72
Few-Shot LLMs w/ 3 Demonstrations										
A8	GPT-4	DO	19.22	0.77	0.69	0.98	18.77	17.70	1.07	0.68
A9	GPT-3.5	DO	19.05	0.54	0.41	0.77	16.07	16.17	-0.10	0.54
LLMs w/ Parameter-Efficient Fine-Tuning										
B1	Flan-T5-XL	ED	22.22	0.49	0.55	0.78	13.11	**9.37**	3.59	0.70
B2	Flan-T5-XXL	ED	**23.43**	0.78	0.27	0.85	14.40	10.78	**3.62**	**0.75**
B3	LLaMA2-7B	DO	17.75	0.20	1.44	0.77	14.12	13.80	0.32	0.61
B4	Mistral-7B	DO	18.21	0.29	1.76	0.83	13.10	12.42	0.68	0.55
Fully Fine-Tuned LLMs										
C1	Flan-T5-Large-HP	ED	21.36	0.56	0.53	0.86	13.71	10.40	3.31	0.63
C2	Flan-T5-Large-SP	ED	20.58	0.63	0.52	0.74	13.47	11.37	2.10	0.57
Two-Stage Modular Methods										
D1	Flan-T5-XL-LoraHub	ED	14.31	-0.05	2.59	0.55	13.69	13.92	-0.23	0.33
D2	Flan-T5-XL-AdapterFusion	ED	17.65	0.04	1.40	0.65	14.31	14.49	0.18	0.41

Implicit vs. Explicit Control. As shown in Table 4, implicit control proves more challenging than explicit control.

Most models struggle with implicit tasks. For readability, the δ_{FKGL} score remains below 1 across most models, indicating poor differentiation between readability levels. While GPT-4 achieves the highest δ_F of 2.13, its high FKGL score (∼17) suggests it generates overly complex summaries, failing to improve readability as intended. Focus control is similarly weak, with most models scoring around 0.5 F1, and even GPT-4o, the best performer, reaching only 0.72 F1, highlighting limited adaptability.

In contrast, closed-source models excel at explicit control. GPT-3.5, GPT-4, and GPT-4o achieve strong length and keyword control, with PCC scores up to 0.94, MAD scores as low as 0.09, and success rates consistently exceeding 95%. While open-source models underperform in length control, LLaMA2-7B and Mistral-7B still demonstrate strong keyword control.

Effect of Few-Shot Demonstrations. Comparing zero-shot and few-shot performance of GPT-3.5 and GPT-4, we observe that few-shot demonstrations do not improve LLM capabilities in this task. This is likely due to the challenges of processing long sequences, where the length of scientific documents limits the number of effective in-context examples. As a result, the context window becomes constrained, and overly lengthy prompts reduce model effectiveness.

Effect of PEFT. LoRA fine-tuning improves performance across all models, though decoder-only (DO) models show only minor gains in controllability. For example, δ_{FKGL} values for settings B3 and B4 remain below 1, indicating minimal differentiation in readability. Similarly, their length control remains weak, with PCC values below 0.3, reflecting a low correlation between generated and target lengths.

In contrast, encoder-decoder (ED) models B1 and B2 achieve results comparable to or surpassing GPT-4, despite weaker zero-shot performance. B2, in particular, exhibits strong control over length and keywords while significantly improving readability and focus—both implicit attributes. Under high readability constraints, its FKGL score drops below 11, aligning with a high school reading level. For focus control, while its F1 score of 0.75 leaves room for improvement, it still outperforms GPT-4.

Limitations of Decoder-Only Models in Compositional Controllable Summarization. Our comparison of encoder-decoder models (B1, B2) and decoder-only models (B3, B4) shows that LLaMA2 and Mistral underperform compared to Flan-T5 under fine-tuning. This disparity likely arises from architectural differences. Encoder-decoder models like Flan-T5 are optimized for sequence-to-sequence tasks, where the encoder processes input while the decoder generates output by attending to the encoded representations. In contrast, decoder-only models such as LLaMA2 and Mistral handle both

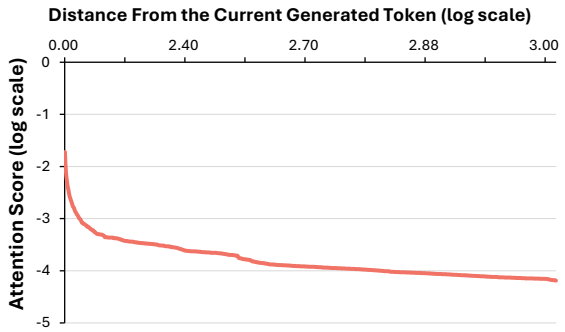

Figure 3: Average attention across all layers and heads for the last 10 tokens generated by LLaMA2 over the preceding 1,024 tokens, displayed on a log-log scale. The sharp decline in attention scores as token distance increases highlights the model's difficulty in maintaining focus on distant tokens within the sequence.

Table 5: Performance comparison of Flan-T5-XXL with LoRA on single and compositional control tasks.

Model	Readability	Length		Keywords	Focus
	δ_F ↑	PCC ↑	MAD ↓	SR ↑	F1 ↑
Single	4.66	0.89	0.23	0.88	0.92
Compositional	3.62	0.78	0.27	0.85	0.75

source and target sequences within the same unidirectional cross-attention mechanism [12]. As sequence length increases, attention becomes more diffused, weakening the model's ability to capture long-range dependencies—critical for controllable summarization.

To validate this, we analyze LLaMA2's attention patterns. Figure 3 shows a sharp decline in attention as token distance increases, indicating that the model prioritizes recent context while progressively disregarding earlier tokens. This degradation in long-range attention reduces the model's ability to maintain focus on control signals and relevant input, ultimately impairing summarization quality.

6 Discussion

Comparison of Single-Attribute vs. Compositional Control. The results in Section 5 reveal the limitations of LLMs in compositional control. To explore these limitations further, we compare the performance of Flan-T5-XXL fine-tuned on single-attribute tasks (i.e., fine-tuned separately on each of the four single-attribute datasets in CCSBench) with its performance on compositional control (i.e., fine-tuned on $\mathcal{D}_{CC}$ as outlined in Section 5).

As shown in Table 5, performance drops significantly when multiple attributes are controlled simultaneously, suggesting a conflict between the attributes that limits the model's compositional controllability.

Bias in LLMs to Prioritize Explicit Attributes. To better understand inter-attribute conflicts, we randomly select 300 samples from the test set of $\mathcal{D}_{CC}$ for each attribute. Keeping the other attribute settings unchanged, we modified the specified attribute (e.g., changing Readability from "normal" to "high") to create 300 new samples for each attribute accordingly. Then, we use the same model to generate the summary. For a given controllability metric, we calculate the change amplitude (CA) between the control ability before the attribute value change m and after the change m_{new}, defined as:

$$CA = |\frac{m_{new} - m}{m}|.$$

We illustrate the dependencies between attributes in Figure 4, where each row represents the attribute being changed and each column represents the attribute being affected. We observe that implicit attributes are more susceptible to the influence of other attributes compared to explicit attributes.

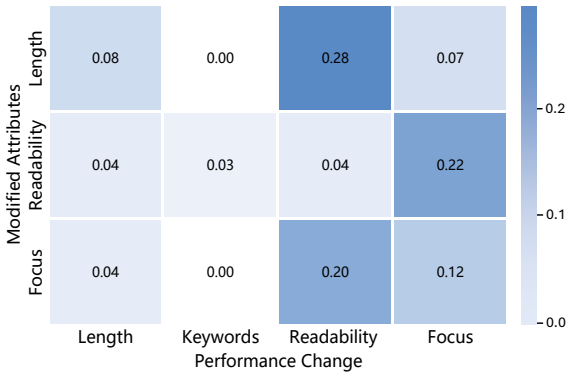

Figure 4: Attribute dependencies reflected by change amplitude (CA) on PCC (length), SR (keywords), δ_{FKGL} (readability), and F1 (focus). Each row represents the attribute being modified, and each column represents the affected attribute. Numerical values and color intensity denote the strength of the dependency.

Specifically, in the case of Flan-T5, readability shows a stronger dependency on length than vice versa. This indicates that when controlling both attributes simultaneously, the model tends to prioritize the length constraint. For example, in Case 1 of Table 6, to shorten the length of the summary, the model omits the original simple explanation of the term "feature" and replaces it with the more abstract concept of "descriptor", making the summary more concise but harder to understand.

While implicit control requires more abstract reasoning, explicit control involves clearer and more concrete constraints. This bias suggests a limitation in the model's ability to deeply understand and reason across multiple attributes.

Conflicts Between Readability and Focus. We observe a strong interaction between focus and readability, indicating that the model struggles to balance these two attributes. For example, in Case 2 of Table 6, when the summary emphasizes empirical content, it tends to use simpler sentence structures and phrases such as *"The results show.."* or *"something was tested.."* along with terms like *"effective"*. However, when readability requirements are reduced, the model

Table 6: Case studies on conflicts between readability and focus controllability. The left column presents unchanged attribute values, while the right column shows the impact of modifying a specific attribute. Blue text highlights content that satisfies the control conditions, while red text indicates control failures.

Case 1: Length Against Readability – Shorter Length Results in Unsatisfying Readability	
Readability: high Empirical Focus Level: low	**Length Bin: 1** - We are interested in how the human visual system recognizes objects. Inspired by this, we created a new set of features that include information about gradients in both horizontal and vertical directions, as well as multi-scale contrast maps. These features are designed to help identify pedestrians in complex scenes.
	Length Bin: 0 - We introduce a new set of features inspired by the human visual system, including multi-direction, multi-scale contrast and statistical multi-channel descriptors. These features are designed to enhance pedestrian detection performance in various conditions.

Case 2: Readability Affects Focus – Lower Readability Results in Less Empirical Focus	
Length Bin: 0 **Empirical Focus Level: high**	**Readability: high** - The effectiveness of the proposed human detection method was tested on several images. The results showed that the method was effective even when the human's face was not visible, when the head was partially obscured, or when the subject was far from the camera.
	Readability: normal - This paper presents a novel method for detecting human presence in surveillance videos by analyzing the OMEGA symbol. The proposed method involves identifying the unique signature of the human head-neck-shoulder region and classifying images as human or non-human based on this signature.

Case 3: Focus Affects Readability – Lower Empirical Focus Results in Lower Readability	
Length Bin: 0 **Readability: high**	**Empirical Focus Level: high** - We applied our model to a large set of product reviews and achieved an accuracy of 98.28%, which is comparable to the current state of the art in sentiment classification.
	Empirical Focus Level: low - This paper proposes a supervised sentiment classification model based on the Naive Bayes algorithm. It argues that a simple model can achieve comparable accuracy to state-of-the-art methods, highlighting the importance of selecting appropriate features and eliminating noise.

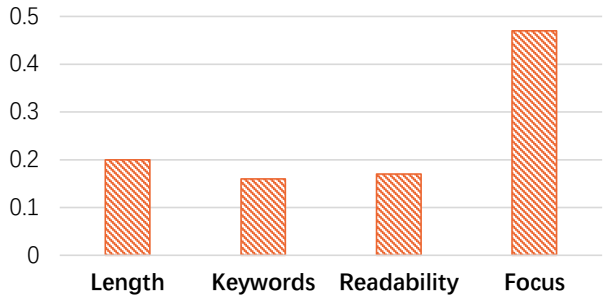

Figure 5: Average attention scores of the fusion layer for each adapter show a bias toward the focus adapter.

introduces more complex, technical vocabulary, shifting the focus toward conceptual content.

In practice, when controlling both readability and focus, we expect the model to prioritize identifying content that aligns with the specified focus, rather than introducing incorrect content just to incorporate more technical language. Similarly, in Case 3, when a lower empirical focus level is set, the model uses more conceptual terminology, making the summary harder for a lay audience to understand. This suggests that the model has difficulty maintaining simple language while describing content with a low empirical focus.

Can Two-Stage Modular Methods Address Inter-Attribute Conflicts? To address these trade-offs, we explore two-stage approaches where single-attribute control is learned in the first stage, and balancing multiple attributes is fine-tuned in the second stage. Specifically, we implement LoRAHub [15], which assigns a fixed weight to each LoRA module; and AdapterFusion [28], which utilizes an attention layer to learn how to focus on different attribute modules. However, results in Table 4 show that neither AdapterFusion nor LoRAHub achieves meaningful controllability, with both producing poor summaries, as reflected in their low ROUGE scores. We believe this failure stems from the overly simplistic integration mechanisms, which are insufficient for learning the complexities required for compositional control, resulting in incoherent outputs.

We further probe the limitations of AdapterFusion by analyzing the model's attention differences across various attributes in over the entire test set, as shown in Figure 5. The model exhibits a bias toward the focus module, suggesting that the attention mechanism struggles to adjust its emphasis on different attributes based on varying control requirements. This interference between modules ultimately degrades the model's overall performance.

7 Discussion on Input Choice for Compositional Controllable Scientific Document Summarization

Obtaining high-quality summaries is key to effective scientific document summarization. To this end, we investigate the effects of using only the introduction and conclusion sections (I+C) versus the full scientific document as input for compositional controllable scientific document summarization.

CCSBench adopts the I+C sections as input, a choice supported by prior research demonstrating its effectiveness over full-document

Table 7: Comparison of GPT-4o's performance on the first 200 samples of CCSBench. I+C: using introduction and conclusion sections as input; Full: using the retrieved full texts as input. Detailed descriptions of metrics are provided in Section 4.2.

Method	Quality	Explicit Attributes			Implicit Attributes				#Tokens
		Length		Keywords	Readability			Focus	
	ROUGE-L ↑	PCC ↑	MAD ↓	SR ↑	F_{normal}	F_{high}↓	δ_{FKGL} ↑	F1 ↑	
Full	**21.15**	**0.98**	**0.06**	0.98	16.55	16.10	0.45	0.67	6.78k
I+C (ours)	20.84	0.96	**0.06**	**0.99**	15.87	**15.37**	**0.50**	**0.72**	**1.61k**

summarization [2, 25]. This effectiveness likely stems from the challenges language models face in maintaining coherence and relevance over long contexts. As CCSBench aims to foster effective general scientific summarization rather than produce in-depth analyses, the I+C sections provide a concise yet comprehensive representation of a paper's key elements, including task definitions, prior limitations, methodologies, and main findings [31]. Thus, our CCSBench aligns with the established I+C practice in the field.

With recent LLMs such as GPT-4o [16] supporting extended context windows of up to 128k tokens, an open question arises: **Can these models overcome long-context challenges in processing full scientific documents for compositional controllable summarization?** To explore this, we conduct experiments on the first 200 samples of CCSBench, retrieving their full documents and comparing GPT-4o's summarization performance using introduction and conclusion sections (**I+C**) versus full-text input (**Full**).

The results, presented in Table 7, indicate that **I+C** outperforms **Full** in keyword preservation, readability, and focus while maintaining comparable performance on other metrics. Notably, **I+C attains comparable performance while consuming only 23.7% of the tokens required for Full, validating its effectiveness and practicality for compositional controllable scientific document summarization.**

8 Conclusion

We introduce CCSBench, the first benchmark for evaluating compositional controllable scientific summarization. Integrating both explicit (length, keyword inclusion) and implicit (readability, empirical focus) attributes, CCSBench provides a structured framework for assessing LLMs' ability to generate controlled, contextually appropriate scientific summaries. Our extensive experiments establish concrete LLM baselines and reveal significant limitations in compositional controllability, emphasizing the need for further research. We identify key challenges, including trade-offs between attributes and difficulties in handling implicit control, while also outlining potential directions for improvement. By providing actionable insights into LLMs' strengths and limitations in this challenging task, we offer a foundation for advancing the field.

Limitations and Future Work

Empirical results on CCSBench highlight key challenges in LLMs' ability to handle compositional controllable scientific summarization. While our findings show that fine-tuning smaller models (e.g., Flan-T5) can achieve performance comparable to GPT-4 across multiple metrics and provide deeper insights into failure cases, further research is needed to better understand and resolve observed interactions between control attributes.

Currently, CCSBench focuses on scientific documents in English. Expanding this framework to support multilingual scientific summarization is a promising direction for future work. Additionally, while CCSBench is designed for effective general communication, more granular summarization – addressing the need for in-depth analysis, such as detailed experimental setups and side observations – remains an important area for exploration. Specifically, extending our framework and methodology to full scientific documents using datasets like LimGen [9] could provide valuable insights into more comprehensive and nuanced summarization strategies.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[2] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme Summarization of Scientific Documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 4766–4777. doi:10.18653/v1/2020.findings-emnlp.428

[3] Hou Pong Chan, Lu Wang, and Irwin King. 2021. Controllable Summarization with Constrained Markov Decision Process. *Transactions of the Association for Computational Linguistics* 9 (2021), 1213–1232. doi:10.1162/tacl_a_00423

[4] Canyu Chen and Kai Shu. 2024. Can LLM-Generated Misinformation Be Detected?. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=ccxD4mtkTU

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. doi:10.48550/ARXIV.2210.11416

[6] Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the Use of ArXiv as a Dataset. arXiv:1905.00075 [cs.IR] https://arxiv.org/abs/1905.00075

[7] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 615–621. doi:10.18653/v1/N18-2097

[8] Yixi Ding, Yanxia Qin, Qian Liu, and Min-Yen Kan. 2023. CocoSciSum: A Scientific Summarization Toolkit with Compositional Controllability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 518–526.

[9] Abdur Rahman Bin Md Faizullah, Ashok Urlana, and Rahul Mishra. 2024. LimGen: Probing the LLMs for Generating Suggestive Limitations of Research Papers. arXiv:2403.15529 [cs.CL] https://arxiv.org/abs/2403.15529

[10] Angela Fan, David Grangier, and Michael Auli. 2018. Controllable Abstractive Summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Alexandra Birch, Andrew Finch, Thang Luong, Graham Neubig,

and Yusuke Oda (Eds.). Association for Computational Linguistics, Melbourne, Australia, 45–54. doi:10.18653/v1/W18-2706

[11] Xiao Fu. 2023. KWX: scholarly articles with keywords from arXiv. doi:10.5281/zenodo.7929439

[12] Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-Only or Encoder-Decoder? Interpreting Language Model as a Regularized Encoder-Decoder. arXiv:2304.04052 [cs.CL] https://arxiv.org/abs/2304.04052

[13] Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. CTRLsum: Towards Generic Controllable Text Summarization. arXiv:2012.04281 [cs.CL] https://arxiv.org/abs/2012.04281

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[15] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2024. LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition. arXiv:2307.13269 [cs.CL] https://arxiv.org/abs/2307.13269

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[17] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[18] Daniel Kahneman. 2013. Thinking, Fast and Slow. *Language* 21, 500 (2013), 25cm.

[19] Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. https://api.semanticscholar.org/CorpusID:61131325

[20] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4582–4597. doi:10.18653/v1/2021.acl-long.353

[21] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[22] Dongqi Liu, Yifan Wang, Jia E. Loy, and Vera Demberg. 2024. SciNews: From Scholarly Complexities to Public Narratives – a Dataset for Scientific News Report Generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italy, 14429–14444. https://aclanthology.org/2024.lrec-main.1258

[23] Yi Liu, Xiangyu Liu, Xiangrong Zhu, and Wei Hu. 2024. Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9231–9253.

[24] Yizhu Liu, Zhiyi Luo, and Kenny Zhu. 2018. Controlling Length in Abstractive Summarization Using a Convolutional Neural Network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4110–4119. doi:10.18653/v1/D18-1444

[25] Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing Structure into Summaries: a Faceted Summarization Dataset for Long Scientific Documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1080–1089. doi:10.18653/v1/2021.acl-short.137

[26] OpenAI. 2022. ChatGPT: Optimizing language models for dialogue.

[27] OpenAI, Josh Achiam, Steven Adler, et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[28] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 487–503.

[29] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. arXiv:1704.04368 [cs.CL] https://arxiv.org/abs/1704.04368

[30] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=M23dTGWCZy

[31] Stanford University. 2006. Tips for Writing Technical Papers. https://cs.stanford.edu/people/widom/paper-writing.html#intro

[32] Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL] https://arxiv.org/abs/2307.09288

[33] 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*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1603–1623. doi:10.18653/v1/2024.findings-acl.93

[34] Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. AutoSurvey: Large Language Models Can Automatically Write Surveys. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[35] Longfei Yun, Letian Peng, and Jingbo Shang. 2025. UltraGen: Extremely Fine-grained Controllable Generation via Attribute Reconstruction and Global Preference Optimization. *arXiv preprint arXiv:2502.12375* (2025).

[36] Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023. MACSum: Controllable Summarization with Mixed Attributes. *Transactions of the Association for Computational Linguistics* 11 (2023), 787–803. doi:10.1162/tacl_a_00575

[37] Shang Zhou, Feng Yao, Chengyu Dong, Zihan Wang, and Jingbo Shang. 2024. Evaluating the Smooth Control of Attribute Intensity in Text Generation with LLMs. In *Findings of the Association for Computational Linguistics ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 4348–4362. doi:10.18653/v1/2024.findings-acl.258

A Experimental Setup

A.1 Models

Recent advancements in large-scale generative language models based on transformers, such as GPT [1], have significantly improved performance on various natural language processing tasks. Given their widespread use and strong baseline performance, we select a range of large language models to evaluate their ability to perform compositional controlled summarization.

In our experiments, we focus on a variety of model architectures, including encoder-decoder models like the Flan-T5 series, decoder-only models such as LLaMA [32] and Mistral [17], and open-source large models like GPT-4 [27]. Additionally, we evaluate their parameter-efficient fine-tuning (PEFT) versions to assess whether parameter-efficient fine-tuning enhances their ability to handle controlled summarization tasks. Moreover, we also revisit previously established baselines in the news domain for controlled summarization, namely hard prompt (HP, 36) and soft prefix tuning (SP, 20) approaches on smaller models. We fine-tune each of these models using our compositional control dataset, specifically designed to test their ability to handle multiple attributes simultaneously. Across these models, we explore various objectives, yielding multiple model variants tailored for controlled summarization. These configurations allow us to comprehensively study the strengths and limitations of different architectures and tuning strategies when applied to compositional summarization tasks, as described in further detail below.

Zero-shot (ZS). In our simplest setting, we evaluate the compositional controllability already learned by large language models due to pretraining on large-scale corpora. In this setting, the models are not fine-tuned on any portion of our compositional control dataset. Instead, they must generate summaries for the test set based solely on the representations learned during pretraining.

At test time, the model receives the input document along with a set of control signals ($doc, len, focus, kw, read$) and generates the summary accordingly. The prompt template we used can be found in the Appendix C.

Parameter-efficient Fine-tuning (PEFT). Because the pretraining domains of large language models are broad and cover diverse textual sources, we investigate whether adapting these models to the specific data distribution of our compositional control dataset can enhance their ability to handle multi-attribute summarization for scientific documents. In this setting, we apply parameter-efficient fine-tuning (PEFT) to these models, which allows us to adapt these large models to our task without updating all of their parameters. In this setting, we leverage the widely-used PEFT technique LoRA (Low-Rank Adaptation, Hu et al. [14]), which inserts low-rank trainable matrices into the model's architecture. The document with the golden summary and four controlled attributes ($doc, summary, len, focus, kw, read$) is the input to the model.

Hard Prompt (HP). Following Zhang et al. [36], in the hard prompt setting, we provide the model with explicit instructions regarding the desired control attributes by using a structured prompt format. Each control attribute is formatted as "Attribute: Value", where "Attribute" refers to one of the specified control dimensions, such as "Length", "Focus", "Keywords", or "Readability", and "Value" represents the target value for that attribute (e.g., "Noramal", "High", etc.).

Soft Prefix (SP). In this setting, we prepend external trainable parameters-referred to as "prefixes"-to each layer of the summarization model to exert fine-grained control over the generation process. For controlling each attribute value, we assign m prefix embeddings for the respective attribute, where m is a hyperparameter representing the length of the prefix. For instance, for controlling summary readability with a "High" specification, we assign a series of embeddings $E_{Read:high} = [e^1_{Read:high}, ..., e^m_{Read:high}]$, where each e^j_i is a vector of the same dimensionality as the model's word embeddings. For implementation details, readers may refer to Li and Liang [20].

A.2 Implementation Details

We use PyTorch and the Huggingface library to implement all the models. The experiments are conducted on 2 A40 GPUs. We choose 1e-4 as the learning rate for all PEFT models and the max epoch is set to 20. For hard prompt and soft prefix tuning, we largely follow the same training and inference setups as in Zhang et al. [36]. In all of our experiments, we run each model once.

A.3 Evaluation Metrics

Here we introduce different metrics in detail.

Length Control. For length controllability evaluation, we adopt (1) the Mean of Absolute Deviation (MAD, 24) of length codes of system-generated summaries and the references, measuring their length distance; and (2) the Pearson Correlation Coefficient (PCC, 24) between the generated length and the input length bin.

Readability Control. For evaluating readability controllabiltiy, we calculate the Flesch-Kincaid Grade Level (FKGL, 19) for each document under both high and normal readability settings. A lower FKGL score indicate higher readability. We use δ_F to represent the difference in FKGL scores between the two categories. A larger δ_{FKGL} indicates a greater distinction in readability, which in turn reflects stronger control over readability by the model.

Keywords Control. For evaluating keyword controllability, we use the Success Rate (SR, 13), namely the fraction of keywords actually occurring in the output summaries. We calculate SR employing exact matching after stemming.

Empirical Focus Level Control. Since focus is difficult to distinguish using simple rules, we opt to use an LLM evaluator for identification. For the generated summaries, we instruct GPT-4o to determine whether they are more conceptual-focused or empirical-focused. We experiment with various instruction patterns and validate the approach on the manually annotated dataset, achieving a best accuracy of 93%. Next, we use the categories predicted by GPT-4o as the labels for the generated summaries and calculate the F1 score for each category. A higher F1 score indicates stronger control over the focus attribute. Prompt details can be found in Appendix C.

B Human Evaluation Details

As described in Section 3.3, we employed human evaluators to assess the quality of GPT-generated summaries (regarding Readability and Focus) in CCSBench.

The evaluations were conducted via Amazon Mechanical Turk (MTurk)[3], with participants compensated at a rate of USD $0.60 per Readability case and USD $0.75 per Focus case. We clearly stated during recruitment that the collected data would be used solely for research purposes. The human evaluation process has received IRB approval from the authors' institution.

B.1 Instructions for Readability Evaluation

Instructions to MTurk Annotators. Read the two scientific summaries and evaluate their facuality, readability and fluency.

[Summary 1]
[Summary 2]

Questions for MTurk Annotators. Q1: Do the two summaries agree on the main points? (Yes / No)

Q2: Which summary is more readable? (Summary 1 / Summary 2)

Q3: Is Summary 1 written in natural and fluent language? (Yes / No)

Q4: Is Summary 2 written in natural and fluent language? (Yes / No)

B.2 Instructions for Focus Evaluation

Instructions to MTurk Annotators. Read the two scientific summaries and the reference abstract, evaluate their factuality, focus and fluency.

Definition:

The empirical-focused summary emphasizes the experimental, data-driven aspects of a study. It highlights data collection, experimental settings, performance metrics and concrete results obtained from the research, especially when it comes to statistics and data.

The conceptual-focused summary emphasizes the theoretical and innovative aspects of a study. It highlights the motivation, the challenge, the underlying theoretical framework, the description of the proposed method or algorithm, and broader implications of the research.

Empirical-focused Example: This study evaluates the performance of a modified CRF-based POS tagging system for Manipuri, incorporating new features and the Reduplicated Multiword Expression (RMWE) feature. The experiment shows that the new CRF system achieves a Recall of 78.22%, Precision of 73.15%, and F-measure of 75.60%. With the inclusion of RMWE as a feature, the results improve to a Recall of 80.20%, Precision of 74.31%, and F-measure of 77.14%.

Conceptual-focused Example: This paper provides an in-depth overview of the updated feature selection approach in CRF for Manipuri POS tagging. It highlights the significance of optimal feature selection in enhancing CRF performance and discusses the introduction of new features, including the Reduplicated Multiword Expression (RMWE), which is crucial for accurately tagging Manipuri language POS due to its rich occurrence of RMWE.

[3]https://www.mturk.com/

[Abstract]
[Summary 1]
[Summary 2]

Questions for MTurk Annotators. Q1: Does Summary 1 agree with the reference Abstract on the main points? (Yes / No)

Q2: Does Summary 2 agree with the reference Abstract on the main points? (Yes / No)

Q3: Which summary is more empirically-focused? (Summary 1 / Summary 2)

Q4: Is Summary 1 written in natural and fluent language? (Yes / No)

Q5: Is Summary 2 written in natural and fluent language? (Yes / No)

C Prompt Templates

C.1 Readability Data Construction

"""

System

You are a NLP expert. Please help me paraphrase some scientific summaries according my requests.

Instruction

Please paraphrase this abstract for middle school students without using metaphors and without including information that cannot be obtained directly from the original abstract: {abstract}

"""

C.2 Focus Data Construction

"""

System

You are a NLP expert. Please help me analyze some scientific abstracts.

Instruction for focus identification

Can you identify the empirical content and conceptual content of the following abstract? If the abstract is obviously predominated by only one type of information, the other part can be 'None'.

Definition:

The empirical content emphasizes the experimental, data-driven aspects of a study. It highlights data collection, experimental settings, performance metrics and concrete results obtained from the research, especially when it comes to statistics and data.

The conceptual content emphasizes the theoretical and innovative aspects of a study. It highlights the underlying theoretical framework, the description of the proposed method or algorithm, and broader implications of the research.

Abstract: {abstract}

Output (Extracted by functions)

{"Empirical Content": output1, "Conceptual Content": output2}

Instruction for paraphrasing

If neither part is 'None', please paraphrase the empirical content and the conceptual content into a separated empirical-focused abstract and a separated conceptual-focused abstract respectively. Both abstracts should be coherent and fluent without including information that can't be inferred directly in the original abstract. If either part is 'None', take the input abstract and 'None' as the two output abstracts. Note that the difference between empirical focus and conceptual focus lies on the content rather than language expression.

Output (Extracted by functions)
{"Empirical Summary": output3, "Conceptual Summary": output4}
"""

C.3 Focus Evaluation

"""
System
You are a NLP expert. Please help me analyze some scientific abstracts.

Instruction
Help me decide whether this abstract is more empirically-focused or more conceptually-focused. The output should be 'empirical' or 'conceptual'.

Definition:
The empirically-focused abstract emphasizes the experimental, data-driven aspects of a study. It highlights data collection, experimental settings, performance metrics and concrete results obtained from the research, especially when it comes to statistics and data.
The conceptually-focused abstract emphasizes the theoretical and innovative aspects of a study. It highlights the motivation, the challenge, the underlying theoretical framework, the description of the proposed method or algorithm, and broader implications of the research.
Abstract: {abstract}

"""

C.4 Zero-Shot LLM Experiments

"""
System
You are an expert summarizer who can generate summaries with specific controls.

Instruction
Your task is to create a summary of the given scientific document with the following controls:

Length: The summary should fit in the specified word counts.
Keywords: Include the following keywords in the summary: [list of keywords]
Readability: Ensure the summary is either highly readable for laymen (high) or not specifically optimized for readability (normal).
Empirical Focus Level: Make the summary has high empirical focus level (high) by emphasizing the experimental, data-driven aspects of a study. It highlights data collection, experimental settings, performance metrics and concrete results obtained from the research, especially when it comes to statistics and data; or low empirical focus level (low) the theoretical and innovative aspects of a study. It highlights the motivation, the challenge, the underlying theoretical framework, the description of the proposed method or algorithm, and broader implications of the research.

Document: {doc}

Control signals
Length: 0-50 words
Keywords: "multi-scale"
Readability: high
Empirical Focus Level: high

Output
Summary:
"""

C.5 Few-Shot LLM Experiments

"""
System
You are an expert summarizer who can generate summaries with specific controls.

Instruction
Your task is to create a summary of the given scientific document with the following controls:

Length: The summary should fit in the specified word counts.
Keywords: Include the following keywords in the summary: [list of keywords]
Readability: Ensure the summary is either highly readable for laymen (high) or not specifically optimized for readability (normal).
Empirical Focus Level: Make the summary has high empirical focus level (high) by emphasizing the experimental, data-driven aspects of a study. It highlights data collection, experimental settings, performance metrics and concrete results obtained from the research, especially when it comes to statistics and data; or low empirical focus level (low) the theoretical and innovative aspects of a study. It highlights the motivation, the challenge, the underlying theoretical framework, the description of the proposed method or algorithm, and broader implications of the research.

Examples
Examples:

Document: {doc1}
{ctrl1}
Summary: {summary1}

Document: {doc2}
{ctrl2}

```
Summary: {summary2}

Document: {doc3}
{ctrl3}
Summary: {summary3}

# To generate
Document: {doc}

Length: 0-50 words
Keywords: "multi-scale"
Readability: high
Empirical Focus Level: high

# Output
Summary:
"""
```