DcLM: Output Length Control of Large Language Models via Dynamic Length Markers

Zhe Chen[1], Jiaao Yu[1], Honglin Li[2,*]

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai, China
{zhechen666,yujiaao}@stu.ecnu.edu.cn
hlli@hsc.ecnu.edu.cn

Abstract

Length-controllable text generation (LCTG) is essential for tasks like text summarization and report generation. However, large language models (LLMs) have limited awareness of output length, so precise control over the length of generated text remains a significant challenge. Most existing methods focus on prompt-based frameworks, position encoding, and reinforcement learning for model training. These approaches may affect semantic quality, and struggle to maintain consistent length control across different models and tasks. In this paper, we propose DcLM, a model-agnostic approach that introduces dynamic length markers to guide length-controllable outputs. During training, the model leverages these markers as in-context information, without learning to generate them. At inference time, an external word counter and injected length information guide the model to produce outputs of accurate lengths. We evaluate our method across multiple datasets, and the experimental results demonstrate that DcLM significantly reduces length deviation, showcasing its robust generalization ability across various length scales and tasks.

1 Introduction

Large language models (LLMs) have achieved remarkable progress in recent years, exhibiting strong performance across a wide range of natural language generation tasks (Vaswani et al., 2017; Devlin et al., 2019). As model capacity and context length continue to expand, LLMs are increasingly deployed in real-world scenarios that require flexible and controllable text generation (Liang et al., 2024). In particular, many practical applications impose explicit constraints on output length, such as summarization, content compression, report generation, and so on (Hu et al., 2015; Chhun et al., 2022; Bai et al., 2024). Consequently, effective length-controllable text generation (LCTG) has become an important and practical requirement for modern LLM-based generation systems.

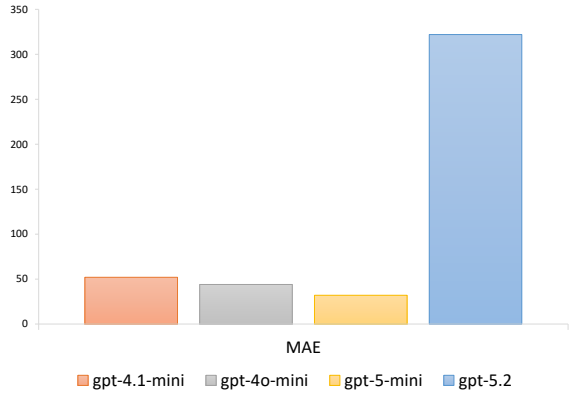

Figure 1: Length Control Performance of Closed-Source Models on HANNA dataset.

Despite their strong generative capabilities, most LLMs control output length only implicitly through learned distributions (Moon et al., 2025), making it difficult to reliably satisfy precise length requirements, especially for long-form generation. In length-controllable text generation, the model is expected to produce coherent text that adheres to a specified length constraint, while maintaining semantic relevance and fluency. However, length is not explicitly modeled as an optimization objective, and small deviations during early generation stages can accumulate over time. This challenge is further amplified in long texts, where maintaining global coherence under strict length constraints remains non-trivial. As shown in Figure 1, the length-controllable performance of the GPT series models on HANNA dataset also remains to be improved.".

Length-controllable text generation poses several inherent challenges that are not fully addressed by existing approaches. First, certain methods operate at the token level (Butcher et al., 2025), while length constraints in practical applications are usu-

Findings of the Association for Computational Linguistics: ACL 2026, pages 23117–23131

ally defined in terms of words or human-perceived units, resulting in a gap between controlled length and user expectations. Second, achieving precise length control remains difficult, as some approaches only provide coarse-grained constraints by guiding generation toward a target range rather than an exact length(Li et al., 2024). Third, some methods introduce explicit control markers and require the model to generate them during decoding (Song et al., 2024; Xie and yi Lee, 2025), which may disrupt semantic continuity and harm overall fluency. Fourth, training-free approaches often rely heavily on prompt engineering (Akinfaderin et al., 2025; Juseon-Do et al., 2024; Yuan et al., 2025a), and in some cases require repeated sampling or rejection to meet length constraints, resulting in high computational costs and limited efficiency. Finally, most existing studies focus on short texts, leaving length-controllable generation for long-form text largely unexplored, despite its importance in real-world applications.

Motivated by the aforementioned challenges, we propose DcLM, a novel framework for fine-grained length control in large language model generation based on dynamic length markers. Unlike the token-level approaches, DcLM operates at the word level, aligning length control more closely with human-perceived text length. Each dynamic length marker contains rich length information, including not only the target output length but also the length of text already generated and the remaining length to be produced. To minimize interference with semantic modeling, DcLM treats these markers as in-context information rather than generation targets, the model is not trained to explicitly generate dynamic length markers. During inference, an external length counter is employed to track the current generation length and insert dynamic length markers at appropriate positions, enabling precise and efficient length control. dynamic length markers are injected into the output sequence following a decay-interval strategy (Yuan et al., 2025a), which allows length signals to be provided throughout the generation process. Experimental results show that DcLM achieves strong performance on short-text tasks such as summarization and yields substantial improvements on long-form generation.

- We design dynamic length markers, which explicitly indicate the target length as well as the generated and remaining budgets at the word level, enabling fine-grained and human-aligned length control.
- We propose DcLM, a unified framework that integrates dynamic length markers with a masking-based training strategy and inference-time markers insertion, achieving precise length control without forcing the model to generate control markers.
- We conduct extensive experiments on multiple datasets across diverse generation tasks and length constraints, covering output lengths ranging from 10 to 6000 words, and demonstrate that DcLM consistently improves length accuracy and generation quality for both short- and long-form text.

2 Related Work

2.1 Training-Free Methods of LCTG

Length-controllable text generation aims to guide language models to produce outputs that satisfy predefined length constraints while maintaining semantic coherence and fluency. With the increasing capabilities of large language models, prompt-based approaches have been explored to achieve length control without modifying model parameters. Akinfaderin et al. (2025) propose Plan-and-Write, which decomposes text generation into two phases: explicit word-count planning followed by coherence-aware output verification through structured prompts. Xie and yi Lee (2025) introduces countdown markers into prompts to guide generation length, while Juseon-Do et al. (2024) proposes an instruction-based sentence compression method to control output length. Although prompt-based methods are intuitive and training-free, they still rely heavily on the model's inherent understanding of length constraints. Another line of work focuses on inference-time strategies for length control. MARKERGEN (Yuan et al., 2025a) addresses core errors in length-controllable generation and proposes a three-stage decoupled generation framework that balances length accuracy and text quality. Retkowski and Waibel (2025) improves length adherence through prompt-based automated revision and sample filtering, while Gu et al. (2025) applies importance sampling during decoding to steer output length.

2.2 Training-Based Methods of LCTG

Training-based methods modify model parameters to enable text generation under explicit length con-

straints. RULER (Li et al., 2024) enhances the instruction-following ability of large language models by introducing meta length tokens during training. Butcher et al. (2025) incorporates a length-difference positional encoding into input embeddings during fine-tuning to improve length adherence.

Other approaches inject position-related information into training data to encourage models to learn length control. PositionID (Wang et al., 2024) and Hansel (Song et al., 2024) train models on text augmented with explicit positional or marker information, enabling length-aware generation behavior. In addition, a line of work explores reinforcement learning–based methods for length control, Yuan et al. (2025b) applies Direct Preference Optimization on a curated dataset with length-constrained preferences, while another work designs length-based rule rewards and optimizes models using Proximal Policy Optimization (Jie et al., 2024).

3 Method

In this section, we introduce DcLM, a framework for length-controlled text generation with dynamic length markers. We first describe the construction of the DcLM dataset, and then present the DcLM framework for training and inference.

3.1 The DcLM Dataset

To enable length-aware generation, we construct the DcLM dataset by augmenting standard instruction–response pairs with dynamic length markers that explicitly indicate generation progress. Given a reference output with a target word length L, we inject structured markers into the output sequence at multiple positions, allowing the model to observe length information during training.

3.1.1 Dynamic Length Marker

Due to the inaccuracy of implicit length estimation in LLMs (Shin and Kaneko, 2024), we introduce explicit and human-aligned length control by combining external word counting with structured length information injection. Specifically, we design dynamic length markers at the word level that provide precise and interpretable guidance for length-controlled generation. Each dynamic length marker encodes rich length information, including the target output length, the amount of text generated so far, and the remaining length budget. Formally, a dynamic length marker is defined as a structured token sequence of the form:

$$\langle L \mid l \mid L - l \rangle$$

, where L denotes the target number of words, l represents the current word count up to the insertion position, and $L - l$ indicates the remaining length budget. To ensure accuracy and consistency, word counts are obtained using an external counting tool, avoiding reliance on the model's internal and often imprecise length estimation. Dynamic length markers are interleaved with the original reference text during training and do not replace or alter any semantic content. As a result, the augmented outputs preserve the original linguistic structure while providing explicit and dynamically updated length-related context at different stages of generation.

3.1.2 Data Construction

Inspired by prior marker-based approaches(Yuan et al., 2025a), we inject dynamic length markers into the reference outputs using a decaying insertion strategy that places markers sparsely at early positions and increasingly densely toward the end of the sequence. Formally, for a target length L, the insertion positions are defined as

$$k_i = \lfloor L \times (1 - \alpha^i) \rfloor \tag{1}$$

where $\alpha \in (0, 1)$ controls the decay rate and i indexes the insertion steps.

This strategy provides stronger guidance when the model begins generating the output by adding an initial marker $\langle L \mid 0 \mid L \rangle$, where length planning is most critical, while reducing interference with semantic coherence at later stages. At each insertion position k_i, a dynamic length marker encoding the current progress is inserted into the reference output.

Through this process, each training sample in the DcLM dataset consists of a source input and a reference output augmented with dynamically updated length markers. These marker-augmented sequences provide explicit conditioning signals for training the DcLM framework described in the following section.

3.2 The DcLM Framework

3.2.1 Training

Given the marker-augmented dataset $\mathcal{D}_{\text{DcLM}}$ constructed in Section 3.1, DcLM is trained using the standard causal language modeling objective. During training, dynamic length markers are included

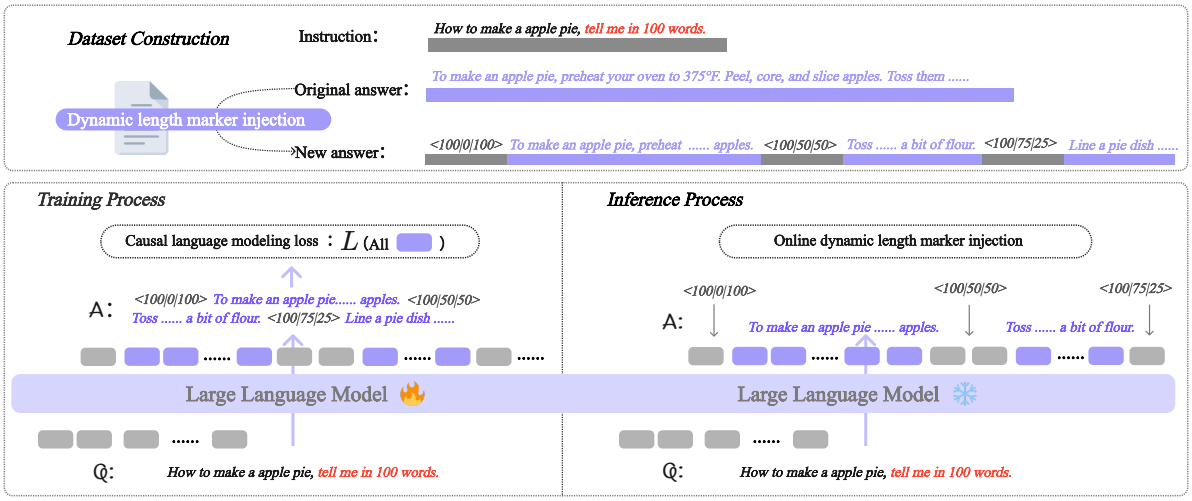

Figure 2: Overview of DcLM.

Algorithm 1 $\mathcal{D}_{\text{DcLM}}$ Data Creation

Require: Word count function $L(\cdot)$, decay factor α
Input: Initial dataset $\mathcal{D} = \{(x, y)\}$
Output: $\mathcal{D}_{\text{DcLM}}$

$\mathcal{D}_{\text{DcLM}} \leftarrow \{\}$
for each tuple (x, y) in $\mathcal{D}$ **do**
  $L \leftarrow L(y)$
  $\mathcal{K} \leftarrow$ unique_sorted($\{\lfloor L(1-\alpha^i) \rfloor \mid i = 1, 2, \ldots, i_{\max}\}$)
  $y' \leftarrow \langle L \mid 0 \mid L \rangle$
  $l \leftarrow 0$
  for each word w in y **do**
    Append w to y'
    $l \leftarrow l + 1$
    if $l \in \mathcal{K}$ **then**
      Append marker $\langle L \mid l \mid L - l \rangle$ to y'
    end if
  end for
  $\mathcal{D}_{\text{DcLM}} \leftarrow \mathcal{D}_{\text{DcLM}} \cup \{(x, y')\}$
end for
return $\mathcal{D}_{\text{DcLM}}$

in the input sequence and participate in the forward computation and attention. However, they are excluded from the loss computation by applying a loss mask, such that the model is not required to predict marker tokens. This design allows the model to leverage dynamic length markers as contextual signals for length control, while focusing the learning objective solely on generating semantic content. As a result, length information is incorporated as in-context guidance rather than an explicit generation target, reducing potential interference with language modeling.

Formally, let $\mathbf{z} = (z_1, \ldots, z_T)$ denote the token sequence of a marker-augmented output, and let $\mathcal{M} \subset \{1, \ldots, T\}$ denote the positions corresponding to dynamic length markers. The training objective minimizes the negative log-likelihood over non-marker positions:

$$\mathcal{L} = -\mathbb{E}_{(x,\mathbf{z})\sim\mathcal{D}_{\text{DcLM}}} \left[\sum_{t \notin \mathcal{M}} \log p(z_t \mid z_{<t}) \right]. \quad (2)$$

Through this masked training objective, DcLM learns to utilize dynamic length markers as contextual length signals without explicitly learning to generate marker tokens. This training strategy avoids exposing the model to marker prediction errors, which can accumulate and negatively affect the quality of generation at inference time.

3.2.2 Inference

At inference time, DcLM performs autoregressive decoding with online dynamic length marker injection. Given a user prompt specifying a target length L, an initial marker $\langle L \mid 0 \mid L \rangle$ is appended to the prompt before generation begins. During decoding, an external counter tracks the number of generated words, and dynamic length markers are inserted into the generation context at predefined positions determined by the same decaying insertion strategy used in dataset construction. Each inserted marker encodes the updated generation progress in the form $\langle L \mid l \mid L - l \rangle$, where l denotes the current word count. These markers are inserted dynamically into the generation context as the sequence

is decoded, guiding the model toward the target length L.

4 Experiments

To evaluate the effectiveness of DcLM, we conduct comprehensive experiments across multiple benchmarks covering a wide range of generation tasks and output lengths. Beyond standard evaluations, we further include additional studies to demonstrate the generalization ability of DcLM under different length constraints and text generation settings.

4.1 Dataset

For training, we construct a dynamic length–aware training set, denoted as D_{DcLM}, based on the RULER training dataset following 1. The training data are augmented with dynamic length markers, enabling the model to learn fine-grained length conditioning during generation. We evaluate DcLM on the following four benchmarks, which cover diverse domains and output length ranges, including short summaries, medium-length and long-form text generation:

- **CNN/DM** (Hermann et al., 2015) is a large-scale news summarization dataset constructed from articles published by Cable News Network (CNN) and Daily Mail (DM).
- **XSUM** (Narayan et al., 2018) is a highly abstractive BBC news summarization benchmark, where one-line article headlines are used as reference summaries.
- **HANNA** (Chhun et al., 2022) is a story generation benchmark designed to evaluate models' ability to produce coherent narrative text under flexible content and length requirements. The dataset consists of prompts paired with human-written stories, making it suitable for assessing length control in open-ended and creative generation settings.
- **LONGBENCH-WRITE** (Bai et al., 2024) is a benchmark specifically designed for long-form text generation under extended context lengths. It evaluates a model's ability to generate coherent and structured long outputs, making it particularly suitable for assessing length control and generation quality in long-text scenarios.

4.2 Metric

For evaluation, we adopt metrics that assess both length controllability and output quality, enabling a comprehensive comparison of different methods under length constraints.

- **Mean Absolute Error (MAE)** measures the average absolute difference between the generated length l_{gen} and the target length l_{target} across all samples, defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|l_{gen}^{(i)} - l_{target}^{(i)}\right|. \quad (3)$$

- **Length Deviation (LD)** measures the absolute deviation between the generated length l_{gen} and the target length l_{target}, defined as:

$$\mathrm{LD} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|l_{gen}^{(i)} - l_{target}^{(i)}\right|}{l_{target}^{(i)}}. \quad (4)$$

- **Hit@K** measures the proportion of generated outputs whose lengths fall within a tolerance window of $\pm k$ words around the target length. This metric captures the robustness of length adherence under relaxed constraints and complements LD by reflecting practical usability, defined as:

$$\mathrm{Hit@}k = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\left|l_{gen}^{(i)} - l_{target}^{(i)}\right| \le k\right). \quad (5)$$

- **Text Quality**. To evaluate generation quality, we adopt an LLM-as-a-Judge paradigm (Liu et al., 2023), where a strong language model evaluates generated texts in terms of fluency, relevance, coherence, and completeness under length constraints. We use gpt-4o as the judge model for all experiments. The evaluation prompt is detailed in Appendix A.4.

4.3 Main Results

Table 1 reports the main experimental results on four benchmarks covering diverse generation scenarios, ranging from news summaries to long-form generation. For the LongBench-Write dataset, we select English samples with target lengths no longer than 6,000 words. Overall, DcLM consistently achieves substantially better length controllability

Dataset	Model	MAE	LD	Hit@5	Hit@10	Hit@20	BertScore	Text Quality
CNN/DM	LLaMA-3.2	4.401	0.087	0.757	0.937	0.996	**0.618**	**4.671**
	w/ SFT	3.315	0.063	0.831	0.967	0.997	0.607	4.564
	w/ DcLM (Ours)	**0.431**	**0.008**	**0.995**	**0.997**	**0.997**	0.611	4.432
XSum	LLaMA-3.2	6.072	0.331	0.548	0.834	0.982	0.553	4.474
	w/ SFT	1.771	0.083	0.975	0.997	**0.999**	0.536	**4.515**
	w/ DcLM (Ours)	**0.225**	**0.011**	**0.997**	**0.998**	**0.999**	**0.558**	4.512
HANNA	LLaMA-3.2	88.510	0.212	0.083	0.135	0.281	0.490	4.638
	w/ SFT	59.041	0.109	0.093	0.145	0.385	**0.492**	4.326
	w/ DcLM (Ours)	**2.572**	**0.003**	**0.968**	**0.968**	**0.979**	0.488	**4.656**
Longbench	LLaMA-3.2	932.236	0.435	0.0169	0.067	0.145	-	**4.462**
	w/ SFT	947.909	0.418	0.054	0.054	0.127	-	4.123
	w/ DcLM (Ours)	**235.436**	**0.182**	**0.525**	**0.576**	**0.800**	-	4.174

Table 1: Experimental results on multiple datasets using the LLaMA-3.2-3B-Instruct model. The target length is set to the reference's length. All fine-tuned models are trained with the LoRA (Low-Rank Adaptation) method. The best results for each dataset and metric are highlighted in bold.

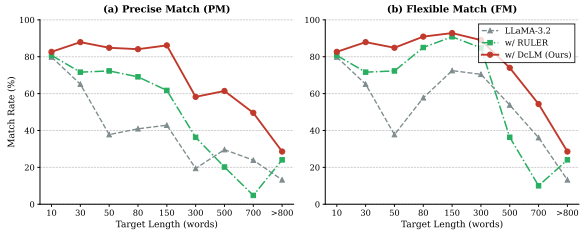

Figure 3: Comparison with RULER under the same evaluation protocol.

than all baselines across all datasets, while maintaining comparable text quality. On CNN/DM and XSum, which focus on summarization, the baseline LLaMA model and its SFT variant already exhibit certain length awareness. Nevertheless, DcLM significantly reduces the LD to below 0.011 and achieves near-perfect Hit@5 and Hit@10 scores, demonstrating its ability to precisely match target lengths even under strict constraints.

The advantage of DcLM becomes more pronounced on HANNA and LongBench, which require length control over long-form generation. Although all baselines suffer from severe length deviations and low Hit@k scores, DcLM improves length adherence, reducing LD from over 0.212 to 0.003 on HANNA and achieving over 0.968 Hit@5 precision. On LongBench, where generating text under a challenging length constraint of up to 6000 words is required, DcLM still delivers substantial gains over all baselines, underscoring its robustness in handling long-context settings. Importantly, DcLM achieves a strong balance between length control and generation quality. As shown in Table 1, DcLM maintains comparable BertScore and text quality to baseline models, suggesting that precise length control can be achieved without degrading fluency or semantic coherence.

4.4 Compared with Ruler

Following the experimental protocol of RULER (Li et al., 2024), we reproduce the results under the same experimental settings using the LLaMA-3.2-3B-Instruct model. RULER evaluates length controllability using two range-based metrics: Proportional Match (PM) and Fixed Match (FM), which measure whether the generated text length falls within predefined tolerance intervals around the target length. Specifically, PM and FM are defined as follows:

$$\mathrm{PM} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(\mathrm{lb}^{P}_{\mathrm{TL}_i} < L(c_i) \le \mathrm{ub}^{P}_{\mathrm{TL}_i}\big) \quad (6)$$

$$\mathrm{FM} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(\mathrm{lb}^{F}_{\mathrm{TL}_i} < L(c_i) \le \mathrm{ub}^{F}_{\mathrm{TL}_i}\big) \quad (7)$$

As shown in Figure 3, DcLM consistently outperforms RULER under the same evaluation protocol across most target lengths. While both methods achieve comparable performance for short outputs, DcLM demonstrates clear advantages as the target length increases. In particular, for medium and long target lengths, DcLM yields substantially higher PM and FM scores, indicating more robust length adherence within both proportional and

Model	LD ↓	Hit@5 ↑	Hit@10 ↑	Text Quality ↑
LLaMA	0.213	0.083	0.135	4.638
DcLM	**0.003**	**0.968**	**0.968**	**4.656**
w/o training	0.104	0.218	0.343	4.371
w/o loss mask	2.454	0.010	0.010	3.146

Table 2: Ablation study on the HANNA dataset. We compare inference-only markers insertion, training without masking, and the DcLM framework.

fixed tolerance ranges. These improvements are especially pronounced at longer lengths, where RULER's performance rapidly degrades, highlighting the effectiveness of dynamically updated length signals for long-form generation.

4.5 Ablation Studies

We conduct an ablation study on the HANNA dataset to examine the contribution of different components in the proposed framework. Table 2 reports results for three configurations based on the LLaMA-3.2-3B-Instruct model. Inserting dynamic length markers only at inference time improves length controllability compared with the vanilla model, as reflected by lower LD and higher Hit@k scores. This indicates that explicit length information can partially guide generation even without training. However, the improvement remains limited, showing that inference-only insertion is insufficient for precise length control. Training with dynamic length markers but without a loss mask leads to significant degradation in both length control and text quality. This suggests that forcing the model to generate length tokens interferes with semantic modeling and disrupts fluent generation. In contrast, the full DcLM framework achieves near-perfect length control, with an LD of 0.003 and Hit@5 and Hit@10 scores exceeding 0.96, while also achieving the highest text quality score. These results confirm that treating dynamic length markers as contextual signals rather than generation targets is critical for accurate length control without sacrificing generation quality.

4.6 Comparison of Insertion Strategies

Table 3 compares the performance of fixed-interval insertion strategies with different interval sizes against the DcLM dynamic length marker approach. Smaller interval values generally lead to better length control, as indicated by lower LD and higher Hit@K scores. For example, $n = 1$ achieves the lowest LD and near-perfect Hit@5 and Hit@10 scores among fixed-interval methods. However, text quality tends to be lower for very small intervals, suggesting a trade-off between strict length adherence and content richness. Increasing the interval reduces length accuracy but improves text quality, indicating that more relaxed insertion allows the model to produce richer and more coherent text. In contrast, the DcLM method achieves a strong balance between length control and text quality, with LD comparable to the strictest fixed-interval setting ($n = 1$) while maintaining the highest text quality among all evaluated strategies. This demonstrates that dynamic length markers can effectively guide the model to respect length constraints without sacrificing content quality.

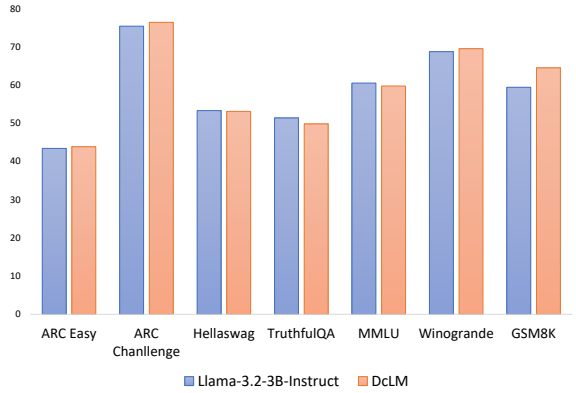

Figure 4: Generalization performance of DcLM across diverse benchmarks.

4.7 Generalizability

4.7.1 Across Diverse Benchmarks

To assess whether DcLM impacts the model's general capabilities beyond length-controlled generation, we evaluate it on a range of standard benchmarks (Gao et al., 2021), including ARC (Easy/Challenge) (Clark et al., 2018), HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), MMLU (Hendrycks et al., 2020), Winogrande (Sakaguchi et al., 2021), and GSM8K (Cobbe et al., 2021). These tasks involve commonsense reasoning, factual accuracy, and mathematical, without any explicit length constraints. We employ 5-shot in GSM8K and 0-shot for other tasks.

As shown in Figure 4, DcLM performs comparably to the LLaMA-3.2-3B-Instruct baseline across all benchmarks. Overall, these results indicate that incorporating dynamic length markers does not impair the model's inherent reasoning or task-solving abilities, demonstrating the robustness and general-

Interval	LD	Hit@5	Hit@10	Text Quality
$n = 1$	0.001	0.989	0.989	3.398
$n = 5$	0.012	0.875	0.895	4.398
$n = 10$	0.048	0.697	0.833	4.516
$n = 20$	0.032	0.468	0.687	4.565
DcLM	0.003	0.968	0.968	4.656

Table 3: Comparison of fixed-interval insertion strategies (with varying n) and the DcLM.

Dataset	Model	LD	Hit@5	Hit@10
CNN/DM	Mistral	0.442	0.101	0.228
	w/ DcLM	0.017	0.988	0.993
	Qwen	0.092	0.664	0.927
	w/ DcLM	0.015	0.986	0.992
XSum	Mistral	1.020	0.036	0.144
	w/ DcLM	0.022	0.992	0.996
	Qwen	0.144	0.864	0.993
	w/ DcLM	0.020	0.988	0.996
HANNA	Mistral	0.131	0.072	0.135
	w/ DcLM	0.003	0.937	0.937
	Qwen	0.075	0.104	0.270
	w/ DcLM	0.005	0.958	0.968

Table 4: Length controllability results across different backbone models. We employ Mistral-7B-Instruct-v0.3 and Qwen3-4B-Instruct-2507 in this experiment.

izability of the DcLM framework.

4.7.2 Across LLMs

We evaluate DcLM on different backbone models to assess its model-agnostic nature. As shown in Table 4, applying DcLM to Mistral-7B-Instruct-v0.3 results in consistent and substantial improvements in length controllability across the CNN/DM, XSum, and HANNA datasets. DcLM significantly reduces length deviation while achieving near-perfect Hit@k scores, without requiring any architecture-specific modifications. Experiments on Qwen3-4B-Instruct yield similar trends. These results indicate that DcLM generalizes well across model families and can be readily integrated with different pretrained LLMs.

4.7.3 Across Lingual

We trained the Qwen model with the DcLM framework on the RULER dataset, which consists of English data. To evaluate its cross-lingual generalization, we tested the model on LCSTS (Hu et al., 2015), a Chinese news summarization dataset. The outputs in LCSTS contain at most 24 characters, so we adopt MAE as well as Hit@5 and Hit@10 as evaluation metrics. As shown in Table 5, applying DcLM reduces the MAE from 5.10 to 2.21 while simultaneously improving Hit@5 and Hit@10 scores by 0.334 and 0.1, respectively. These results demonstrate that the DcLM framework effectively transfers precise length control capabilities across languages without the need for additional training on the target language, achieving significant improvements in both error reduction and accuracy.

Model	MAE	Hit@5	Hit@10
Qwen	5.108	0.619	0.888
w/ DcLM	2.214	0.953	0.988

Table 5: Cross-lingual length control evaluation of Qwen models on a Chinese news summarization dataset.

5 Conclusion

Controllable-length generation in large language models is increasingly important for applications such as text summarization and report generation. In this work, we proposed DcLM, a model-agnostic approach that introduces dynamic length markers to guide length-controllable outputs. During training, the model receives dynamic length markers as in-context information, without learning to generate them. At inference time, an external word counter and injected length information guide the model to produce outputs of accurate lengths. We evaluated our method across multiple length scales, demonstrating its effectiveness in achieving precise length control while maintaining high-quality text generation. These results highlight the potential of DcLM as a general framework for controlling output length in large language models across diverse tasks and languages.

Limitations

In this work, we propose DcLM, a framework that integrates dynamic length markers with a masking-based training strategy and inference-time markers insertion, achieving precise length control without forcing the model to generate control markers. While DcLM achieves impressive results in both length control and text quality, the model's ability to maintain high-quality generation while adhering to the target length may degrade as the target length increases. This limitation is likely due to the inherent capacity constraints of LLMs. As a result, using models with larger parameter sizes may be

necessary to achieve better performance at longer target lengths.

References

Adewale Akinfaderin, Shreyas Subramanian, and Akarsha Sehwag. 2025. Plan-and-write: Structure-guided length control for llms without model retraining. *Preprint*, arXiv:2511.01807.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *Preprint*, arXiv:2408.07055.

Bradley Butcher, Michael O'Keefe, and James Titchener. 2025. Precise length control for large language models. *Natural Language Processing Journal*, 11:100143.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, and 1 others. 2021. A framework for few-shot language model evaluation. *Zenodo*.

Yuxuan Gu, Wenjie Wang, Xiaocheng Feng, Weihong Zhong, Kun Zhu, Lei Huang, Ting Liu, Bing Qin, and Tat-Seng Chua. 2025. Length controlled generation for black-box LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16878–16895, Vienna, Austria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal.

Renlong Jie, Xiaojun Meng, Lifeng Shang, Xin Jiang, and Qun Liu. 2024. Prompt-based length controlled generation with multiple control types. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1067–1085, Bangkok, Thailand. Association for Computational Linguistics.

Juseon-Do, Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2024. InstructCMP: Length control in sentence compression through instruction-based large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8980–8996, Bangkok, Thailand. Association for Computational Linguistics.

Jiaming Li, Lei Zhang, Yunshui Li, Ziqiang Liu, Yuelin Bai, Run Luo, Longze Chen, and Min Yang. 2024. Ruler: A model-agnostic method to control generated length for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3042–3059, Miami, Florida, USA. Association for Computational Linguistics.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and 1 others. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Sangjun Moon, Dasom Choi, Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2025. Length representations in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19775–19793, Suzhou, China. Association for Computational Linguistics.

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

Fabian Retkowski and Alex Waibel. 2025. Zero-shot strategies for length-controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 551–572.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Andrew Shin and Kunitake Kaneko. 2024. Large language models lack understanding of character composition of words. *Preprint*, arXiv:2405.11357.

Seoha Song, Junhyun Lee, and Hyeonmok Ko. 2024. Hansel: Output length controlling framework for large language models. *Preprint*, arXiv:2412.14033.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Noah Wang, Feiyu Duan, Yibo Zhang, Wangchunshu Zhou, Ke Xu, Wenhao Huang, and Jie Fu. 2024. PositionID: LLMs can control lengths, copy and paste with explicit positional awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16877–16915, Miami, Florida, USA. Association for Computational Linguistics.

Juncheng Xie and Hung yi Lee. 2025. Prompt-based one-shot exact length-controlled generation with llms. *Preprint*, arXiv:2508.13805.

Peiwen Yuan, Chuyi Tan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Yueqi Zhang, Jiayi Shi, Boyuan Pan, Yao Hu, and Kan Li. 2025a. From sub-ability diagnosis to human-aligned generation: Bridging the gap for text length control via MarkerGen. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17370–17390, Vienna, Austria. Association for Computational Linguistics.

Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason E Weston, and Jing Xu. 2025b. Following length constraints in instructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24243–24254, Suzhou, China. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800.

A Details of Experiments

In this section, we present the experimental details of DcLM framework.

A.1 DcLM Dataset

As shown in Figure 5, the output text is enhanced through dynamic length markers.

A.2 Experiment Setting

In our experiments, we fine-tune LLaMA-3.2-3B-Instruct, which is obtained from the HuggingFace Transformers[1]. To enable parameter-efficient training, we adopt the LoRA approach. To improve memory efficiency and accelerate training, we further incorporate DeepSpeed ZeRO stage 2. All experiments are implemented based on the LLaMA-Factory framework[2]. We use the AdamW optimizer with a learning rate of 5e-5, a batch size of 4, and maximum token number to 8192 tokens. All experiments are conducted on two NVIDIA A100 GPUs with 80 GB memory.

A.3 Dataset Statistics of Test Set

We evaluate our method on a set of benchmark datasets covering summarization and story generation tasks. For the LongBench-Write dataset, we select English samples with target lengths no longer than 6,000 words. As shown in Table 6, these datasets span a wide range of target output lengths, from short summaries to long-form generation. Such diversity enables a comprehensive assessment of length controllability across both short- and long-text generation scenarios.

A.4 The Details of LLM-as-Judge

As shown in Figure 6, the evaluation prompt instructs the judge model to act as an expert evaluator for natural language generation systems. Given an input prompt and a generated response, the model is required to assess the output in a strict and objective manner without inferring missing information. The generated response is evaluated along four dimensions: fluency and readability, relevance to the input, semantic coherence, and content completeness under length constraints. Each dimension is scored on a 1–5 scale, where higher scores indicate better performance. The overall score is computed as the arithmetic mean of the four dimension scores and reported in a structured JSON format. This design ensures consistent, fine-grained, and reproducible quality assessment across all experiments.

[1]https://github.com/huggingface/transformers

[2]https://github.com/hiyouga/LLaMA-Factory

Dataset	Mean	Median	Min	Max
CNN/DM	52.90	49.0	8	524
XSum	21.73	22.0	2	75
Hanna	492.79	475.0	110	887
LongBench_Write	1591.82	800.0	100	6000

Table 6: Statistics of target output lengths across different datasets.

Table 7 reports the fine-grained LLM-as-Judge evaluation results across fluency, relevance, coherence, and completeness. Overall, DcLM achieves comparable text quality to the Instruct and SFT baselines across all datasets. While minor variations are observed on highly abstractive datasets such as XSum, DcLM consistently maintains strong fluency and relevance scores. Notably, on long-form generation tasks such as HANNA and LongBench-Write, DcLM preserves semantic coherence and completeness at a level similar to or slightly better than the baselines. These results indicate that our framework does not degrade the quality of the generation, even under strict length constraints.

B The Performance of Closed-Source Model

Table 9 presents the length control performance of several closed-source large language models. Among the evaluated models, gpt-5-mini achieves the best overall performance, exhibiting the lowest MAE and LD values and consistently high Hit@K scores, indicating strong accuracy and robustness in adhering to target lengths.

In contrast, gpt-4.1-mini and gpt-4o-mini show demonstrate length control capability, with relatively higher absolute and relative errors and lower Hit@5 scores, suggesting limited precision under strict length constraints. Notably, gpt-5.2 performs substantially worse across all metrics, with large length deviations and low Hit@K values, indicating unstable length adherence in the evaluated setting. These results highlight significant variability in length control behavior across closed-source models, even within the same model family.

C Length Control with Varying Target Lengths

In contrast to the experimental setup in the main section, we investigate the length control capabili-

```
{
   "id": 4,
   "source": "What causes the moon to appear yellow/orange?",
   "prompt": "Answer in 33 words.",
   "output": "Raleigh scattering. Same reason as why the sky is blue. Shorter wavelengths (blues) of light scatter easier and longer wavelengths (reds) pass through the atmosphere better, giving the object a red/yellow tint.",
   "length": 33,
   "output_with_tags": " <33|0|33> Raleigh scattering. Same reason as why the sky is blue. Shorter wavelengths (blues) of light scatter easier and longer wavelengths (reds) pass through <33|23|10> the atmosphere better, giving the object a <33|30|3> red/yellow <33|32|1> tint."
}
```


Figure 5: Example of DcLM Data.

ties of the DcLM method on the HANNA dataset with predefined target lengths. Specifically, we explore the performance of DcLM at target lengths of 100, 200, ..., 1000 words. The results demonstrate that, across the target length range from 100 to 1000, DcLM is far less affected by increasing target lengths compared to the baseline model, maintaining consistently high performance throughout the varying target lengths.

Prompt for Text Quality Evaluation

```
# Role
You are an expert evaluator for natural language generation systems.

# Task
Given the following:
- An input prompt or question
- A generated response
Your task is to evaluate the generated response across multiple quality dimensions.
Be objective, consistent, and strict.
Do NOT infer missing information and do NOT make assumptions beyond the given text.

-------------------------
Input Prompt:
{INPUT}

Generated Response:
{OUTPUT}
-------------------------

# Evaluation Criteria
Score each criterion on a scale from 1 to 5:
1 = very poor, 2 = poor, 3 = acceptable, 4 = good, 5 = excellent.

1. **Fluency and Readability**
Does the response read naturally and smoothly, with correct grammar and clear sentence structure?
Is it easy to follow without awkward or confusing phrasing?

2. **Relevance to the Input**
How well does the response address the input prompt or question?
Does it stay on-topic and directly respond to the user intent?

3. **Semantic Coherence**
Is the response logically consistent and well-organized as a whole?
Do ideas flow naturally without contradictions or abrupt jumps?

4. **Content Completeness**
Does the response include the most important and necessary information needed to address the input?
Are key points missing or insufficiently covered?

# Output Rules
- Output **ONLY** a valid JSON object.
- Do NOT include explanations, comments, or additional text.

# Output Format
{
  "fluency": <1-5>,
  "relevance": <1-5>,
  "coherence": <1-5>,
  "completeness": <1-5>,
}
```

Figure 6: Example Prompt of LLM-as-Judge.

Dataset	Model	Fluency	Relevance	Coherence	Completeness	Overall
CNN/DM	LLaMA-3.2-3B-Instruct	4.991	4.991	4.594	4.108	4.671
	w/ SFT	4.924	4.931	4.388	4.013	4.564
	w/ DcLM	4.599	4.902	4.301	3.927	4.432
XSum	LLaMA-3.2-3B-Instruct	4.881	4.791	4.352	3.872	4.474
	w/ SFT	4.888	4.813	4.453	3.907	4.515
	w/ DcLM	4.829	4.741	4.452	4.026	4.512
HANNA	LLaMA-3.2-3B-Instruct	4.990	4.854	4.479	4.229	4.638
	w/ SFT	4.677	4.635	4.094	3.896	4.326
	w/ DcLM	4.896	4.740	4.510	4.479	4.656
LongBench-Write	LLaMA-3.2-3B-Instruct	4.763	4.729	4.492	3.864	4.462
	w/ SFT	4.475	4.492	4.085	3.441	4.123
	w/ DcLM	4.373	4.492	4.068	3.763	4.174

Table 7: LLM-as-Judge Evaluation Results on Text Quality

Model	**MAE**	**LD**	**Hit@5**	**Hit@10**	**Hit@20**
gpt-4.1-mini	52.92	0.091	0.197	0.322	0.447
gpt-4o-mini	44.75	0.074	0.125	0.375	0.562
gpt-5-mini	23.94	0.039	0.562	0.750	0.937
gpt-5.2	328.08	0.498	0.083	0.116	0.166

Table 8: Length control performance of closed-source LLMs evaluated on HANNA.

Target Length	Model	MAE	LD	Hit@5	Hit@10	Hit@20
100	LLaMA	6.854	0.068	0.677	0.916	0.947
	w/ DcLM	0.614	0.006	0.989	0.989	0.989
200	LLaMA	25.760	0.128	0.156	0.281	0.427
	w/ DcLM	0.895	0.004	0.958	0.989	0.989
300	LLaMA	20.197	0.067	0.208	0.406	0.656
	w/ DcLM	3.291	0.011	0.979	0.979	0.979
400	LLaMA	39.052	0.097	0.114	0.177	0.364
	w/ DcLM	4.895	0.012	0.927	0.979	0.979
500	LLaMA	71.812	0.143	0.031	0.083	0.166
	w/ DcLM	0.687	0.001	0.979	0.989	0.989
600	LLaMA	63.270	0.105	0.072	0.145	0.250
	w/ DcLM	8.635	0.014	0.958	0.968	0.968
700	LLaMA	72.718	0.103	0.072	0.166	0.250
	w/ DcLM	1.458	0.002	0.958	0.979	0.979
800	LLaMA	83.020	0.103	0.052	0.083	0.177
	w/ DcLM	16.260	0.020	0.885	0.906	0.916
900	LLaMA	106.312	0.118	0.031	0.093	0.145
	w/ DcLM	7.687	0.008	0.927	0.947	0.968
1000	LLaMA	148.291	0.148	0.031	0.031	0.031
	w/ DcLM	23.479	0.023	0.822	0.833	0.875

Table 9: Length control performance of LLaMA-3.2-3B-Inistruct and DcLM methods evaluated on HANNA dataset with varying target lengths.