Zero-Shot Strategies for Length-Controllable Summarization

Fabian Retkowski[1,3] **and Alexander Waibel**[1,2,3]
[1]Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Carnegie Mellon University, Pittsburgh PA, USA
[3]KIT Campus Transfer GmbH, Karlsruhe, Germany
{retkowski,waibel}@kit.edu

Abstract

Large language models (LLMs) struggle with precise length control, particularly in zero-shot settings. We conduct a comprehensive study evaluating LLMs' length control capabilities across multiple measures and propose practical methods to improve controllability. Our experiments with LLaMA 3 reveal stark differences in length adherence across measures and highlight inherent biases of the model. To address these challenges, we introduce a set of methods: length approximation, target adjustment, sample filtering, and automated revisions. By combining these methods, we demonstrate substantial improvements in length compliance while maintaining or enhancing summary quality, providing highly effective zero-shot strategies for precise length control without the need for model fine-tuning or architectural changes. With our work, we not only advance our understanding of LLM behavior in controlled text generation but also pave the way for more reliable and adaptable summarization systems in real-world applications.

1 Introduction

Text summarization has seen remarkable advancements with large language models (LLMs), which can now generate coherent, high-quality summaries that often rival human-written ones, particularly demonstrated in the extensively studied news domain (Pu et al., 2023; Retkowski, 2023; Zhang et al., 2024). Consequently, general abstractive summarization tasks are increasingly viewed as 'solved,' pushing the research frontier toward more specialized challenges such as *controllable summarization*, where summaries must adhere to constraints like length, style, or content focus.

Length control, in particular, represents a crucial capability for practical NLP systems, as real-world applications frequently demand outputs that meet strict length constraints—from social media posts to UI text, subtitles, and copywriting (Hori
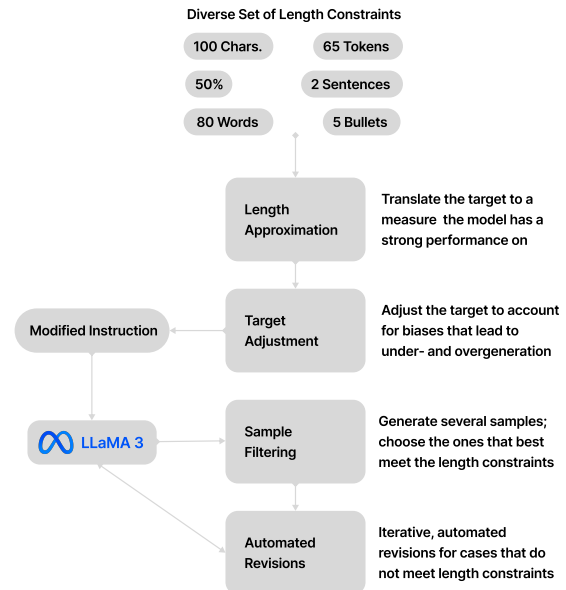

Figure 1: Overview of proposed methods for length controllability as part of a summarization inference pipeline

et al., 2002; Waibel et al., 2004; Waibel and Fuegen, 2012; Liang et al., 2024; Liu et al., 2024). However, while controllable generation performs well in flexible, user-driven tasks involving keywords or specific aspects (Goyal et al., 2022; Xiao et al., 2023; Yang et al., 2023), adherence to strict numerical or length constraints remains problematic (Sun et al., 2023). Studies indicate that even advanced models like Claude 3 Opus and GPT-4 struggle with rigid length control, particularly in zero-shot settings (Chen et al., 2024b; Yuan et al., 2024).

Studies on length controllability have generally been narrow, focusing on a single control scenario defined by a single length measure, such as word count (Juseon-Do et al., 2024; Yuan et al., 2024). Moreover, most approaches rely on fine-tuning models with additional training data or modifying model architectures—like length-aware attention mechanisms (Urlana et al., 2024). Although these methods can be effective, their dependence

on model internals and substantial computational resources limits their practicality in real-world settings where only inference APIs are accessible.

In this paper, we address these research gaps and challenges by conducting a comprehensive study on zero-shot length controllability in LLMs across multiple length measures. We propose and evaluate simple yet effective methods to enable LLMs to generate summaries with precise lengths. Our contributions are threefold:

- We systematically evaluate the inherent length control capabilities of LLMs across a multitude of measures and targets, identifying significant disparities in performance and offering insights into the model's behavior and biases.

- We propose practical methods, including length approximation, target adjustment, sample filtering, and automated revisions, to enhance length adherence in zero-shot settings while preserving or enhancing summary quality.

- We explore combinations of proposed methods yielding highly effective 'recipes' for length-controllable summarization (LCS) with LLMs.

2 Related Research

Traditionally, length-controllable generation has been achieved through decoding adjustments and learning-based methods, often requiring additional training data or model modifications (Kikuchi et al., 2016; Makino et al., 2019; Yu et al., 2021; Liu et al., 2022). However, they frequently either lead to degradation in quality or are not readily applicable to pre-trained language models.

Thus, recent methods have shifted towards fine-tuning using additional data on downstream tasks, including supervised fine-tuning, RLHF or DPO, and PEFT methods such as LoRA (Jie et al., 2024; Juseon-Do et al., 2024; Yuan et al., 2024). However, they often require substantial computational resources, access to model internals and sacrifice the generalizability of the model.

This critical gap in zero-shot approaches also becomes apparent in the recent survey about controllable generation by Urlana et al. (2024). To our knowledge, we are the first to propose and systematically explore zero-shot methods using LLMs for length-controllable generation. In addition, our comprehensive evaluation across multiple measures—including words, characters, tokens, sentences, and bullet points—sets our work apart

from prior studies that focus narrowly on controlling sentences (Goyal et al., 2022; Fonseca and Cohen, 2024) or words (Chen et al., 2024b; Juseon-Do et al., 2024; Yuan et al., 2024), providing deeper insights into the capabilities and limitations of LLMs in length-controllable summarization.

Sample Filtering. Our proposed sample filtering approach draws inspiration from similar methods used in code generation models like AlphaCode, where code samples are filtered based on passing unit tests (Li et al., 2022). More broadly, this methodology represents a form of re-ranking that has been effectively applied in various domains. For instance, Rao et al. (2023) employs it in machine translation and dubbing, ranking translations based on alignment with source audio duration.

Target Adjustment. An observation by Zhang et al. (2024) underscores the challenge of length bias in LLMs. For benchmarking, they targeted 50-word summaries but consistently received longer outputs, prompting them to compensate by setting the target to 25 words. Although anecdotal and not systematically analyzed, this illustrates the length deviation our *target adjustment* method addresses.

Automated Revisions. Moreover, revision and self-correction schemes have become an active area of research (Welleck et al., 2022; Madaan et al., 2023; Peng et al., 2023). These methods involve refining generated text with feedback provided, potentially in multiple passes. In domain-specific tasks like machine translation, post-editing with LLMs has shown significant promise (Chen et al., 2024a). Our automated revision method aligns with this line of research by enabling the LLM to adjust its outputs to better meet length constraints.

3 Methodology

3.1 Length Measures

In this study, we introduce the concept of *length measures* as a framework for specifying and measuring summary length in controllable text summarization. We propose a categorization of these measures into two fundamental groups:

1. **Structural Measures** define summary length using higher-level textual structures, such as *sentences*, *bullet points*, or *paragraphs*.

2. **Granular Measures** that define summary length using fine-grained linguistic or technical units such as *characters*, *tokens*, or *words*.

3.2 Length Approximation

LLMs exhibit varying degrees of proficiency in controlling different length measures, potentially excelling in one (e.g., words) while struggling with others (e.g., characters). To address this disparity, we introduce *length approximation* as a method that translates length constraints between different measures, leveraging the model's proficiency in one measure to improve its performance in others.

3.2.1 Character Length Approximation

We propose a method to approximate character-level control through word-level targets by employing a statistical mapping between character and word counts derived from generated summaries.

Let μ_w denote the mean word length in characters, determined as $\mu_w = 6.31 \pm 0.50$ from our generated summaries (Table 4). We define the character-to-word conversion factor $\alpha_{c \to w}$ as:

$$\alpha_{c \to w} = \frac{1}{\mu_w} \approx 0.158477 \quad (1)$$

The character measure is then approximated by substituting the original target character count C_{target} with a target word count W_{target}:

$$W_{\text{target}} = C_{\text{target}} \times \alpha_{c \to w} \quad (2)$$

3.2.2 Token Length Approximation

We apply a similar approach for token-level control. Based on our corpus, we observe a mean of $\mu_t = 0.798 \pm 0.0656$ tokens per word. The token-to-word conversion factor is thus defined as:

$$\alpha_{t \to w} = \frac{1}{\mu_t} \approx 0.798047 \quad (3)$$

Analogous to the character length approximation, we calculate the target word count from a given token count:

$$W_{\text{target}} = T_{\text{target}} \times \alpha_{t \to w} \quad (4)$$

3.3 Target Adjustment

LLMs often systematically deviate from specified target lengths within the same measure due to inherent biases, leading to systematic overestimation or underestimation. To correct this, we introduce a *target adjustment* method that refines the initial target length based on observed deviations.

3.3.1 Word Target Adjustment

For the word length measure, our study revealed quantifiable relationships between target length and length deviation. Visual inspection of the data indicated a straightforward nonlinear pattern, leading us to employ a third-degree polynomial model:

$$W'_{\text{target}} = a + bW_{\text{target}} + cW^2_{\text{target}} + dW^3_{\text{target}} \quad (5)$$

For word targets, the empirically determined coefficients (see Appendix E for details) are:

$$\begin{aligned} W'_{\text{target}} =\; & 23.7904 + 4.3 \times 10^{-5} W_{\text{target}} \\ & + 1.226 \times 10^{-2} W^2_{\text{target}} \\ & - 3.3 \times 10^{-5} W^3_{\text{target}} \end{aligned} \quad (6)$$

where W'_{target} is the adjusted word target and W_{target} is the original word target.

3.4 Sample Filtering

To enhance the precision of length control, we employ *sample filtering* that leverages the natural variability in LLM outputs. By generating multiple candidates from the search space of possible summaries and selecting the one that best complies with the specified length constraints, we improve adherence to target lengths without compromising semantic content.

3.4.1 Selection Process

For each input text, we generate a set of N candidate summaries $S_1, S_2, \ldots, S_N$ by sampling from the language model. We calculate the length $L(S_i)$ of each summary in the specified measure and compute its absolute deviation δ_i from the target length L_{target}:

$$\delta_i = \left| L(S_i) - L_{\text{target}} \right| \quad (7)$$

The summary that minimizes this deviation is selected as the best summary S^*:

$$S^* = \arg\min_{S_i} \delta_i \quad (8)$$

By selecting the summary with minimal length deviation, this method enhances length compliance while maintaining content quality.

3.5 Automated Revisions

We define an *automated revision* process that involves refining non-compliant summaries. Leveraging automatic length evaluation, we provide guidance to the LLM, steering it towards more length-compliant summaries. If an initial summary S_0

violates length constraints, we prompt the LLM to revise it accordingly, generating a revised summary S_1. A typical problem in refinement schemes is the quality of feedback (Madaan et al., 2023). We can avoid this pitfall since we can accurately determine length compliance and give precise feedback.

3.5.1 Revision Trigger

We define a compliance threshold ϵ as a percentage of the target length. A summary S is considered non-compliant if its length $L(S)$ deviates from the target length L_{target} by more than ϵ:

$$|L(S) - L_{\text{target}}| > \epsilon L_{\text{target}} \qquad (9)$$

3.5.2 Iterative Revision Process

The revision process can operate iteratively if the initial revised summary fails to meet the length constraint. In each iteration i, the LLM generates a new summary S_i based on the previous S_{i-1}, continuing until the summary complies with the length constraint or a maximum number of revisions N is reached. To prevent context length from exceeding the LLM's capacity, each revision step is treated independently, using the latest summary as input and applying a consistent prompt template.

$$S_i = \text{LLM}(T, S_{i-1}, L_{\text{target}}) \qquad (10)$$

where S_i is the i-th revision, T is the original text, and S_{i-1} is the previous summary.

3.6 Integrated Methodologies

In this section, we explore several combinations of the previously introduced methods.

Label	Approach
LA-SF	Length Approx. & Sample Filt.
LA-AR	Length Approx. & Automated Revisions
LA-TA	Length Approx. & Target Adjust.
LA-TA-SF	Length Approx., Target Adjust., & Sample Filt.
SF-AR	Sample Filt. & Automated Revisions
SR	Sampled Revisions

Table 1: Summary of Combined Approaches

LA-SF. This approach combines length approximation with sample filtering. For a target length L_{target} in measure M_1, we first approximate it in measure M_2 and generate N candidate summaries $\{S_1, \ldots, S_N\}$. However, the best summary $S*$ continues to be selected based on the original target length in M_1.

LA-TA. This method applies target adjustment after length approximation. Given a target length L_{target} in measure M_1, we first approximate it in measure M_2, then adjust the approximating target L'_{target} using the polynomial regression model as the final substituting target L''_{target}.

LA-TA-SF. This approach combines all three methods sequentially. We first apply length approximation and target adjustment as described above, then generate N candidate summaries based on the adjusted target length L''_{target} and filter according to the original target L_{target}.

LA-AR. This method combines length approximation with a cross-measure revision process. For a target length L_0 in measure M_0, we first approximate it in a more controllable measure M_1 to generate an initial summary. However, instead of revising toward the approximated target, we recontextualize the revision within the original measure M_0, presenting the initial summary as if it were generated under M_0's constraints. This measurement context manipulation allows subsequent revisions to directly optimize for the original target L_0.

SF-AR. This method combines sample filtering with the automated revision process sequentially. First, we generate N candidate summaries $S_1, \ldots, S_N$ and select the optimal summary S^* using the sample filtering criterion. If S^* is non-compliant, we initiate the automated revision process as described before using S^* as the starting point, potentially iteratively.

SR. This approach integrates sampling into each step of the revision process. For each revision step i, we generate N candidate summaries based on the revision prompt:

$$\{S_{i,1}, \ldots, S_{i,N}\} = \text{LLM}_N(T, S_{i-1}, L_{\text{target}}) \quad (11)$$

We select the best S_i^* and if S_i^* still exceeds the compliance threshold ϵ, it becomes the input for the next revision step.

4 Experimental Setup

We conducted a systematic evaluation of both the inherent length control capabilities of the `Llama-3-8B-Instruct` model (Dubey et al., 2024) and potential improvements through our proposed methods, using the YTSEG (Retkowski and Waibel, 2024) and CNN/DM (See et al., 2017) datasets. More than 6M summaries were generated during

the evaluation, and to support further analysis, we release all generated artifacts[1].

Dataset. The YTSEG dataset (19, 299 video transcripts) was selected for its diversity in length, structure, and domain. Transcript lengths in this dataset range from a few hundred to over ten thousand words. This variability ensured a robust assessment of length controllability across heterogeneous inputs. Additionally, we evaluate on the CNN/DM test set (11,490 articles) as a well-established benchmark in summarization.

Length Measures and Targets. Five length measures were examined with predefined targets, as shown in Table 2, and reference summary lengths from CNN/DM. This comprehensive set of targets enables a multifaceted assessment of length control across varying granularities.

Measure	Targets
Words	50, 100, 150, 200
Characters	150, 300, 500
Tokens	50, 100, 150, 200
Sentences	1, 2, 3, 5, 8
Bullet Points	3, 5, 8

Table 2: Length Measures and Targets

Qualitative Length Quantifiers. We also investigate the impact of qualitative quantifiers such as "short", "concise", "brief", "moderate", "medium-length", "comprehensive", "verbose", and "long" on summary length (Table 5).

Control Methods. We compare the proposed methods against a naïve prompting baseline, which does not employ any additional control mechanisms. Five primary methods were evaluated: (1) length approximation, (2) target adjustment, (3) sample filtering ($s \in [1, 8]$), and (4) automated revisions ($r \in [0, 5]$), (5) various integrated methods.

4.1 Experimental Conditions

Prompt Templates. A consistent prompting strategy is employed across all experiments, instructing the model to generate summaries matching the specified lengths. We utilize three different prompt templates:

1. **Baseline Prompt:** Directly instructing the model in the user message only (Figure A1).

2. **Prompt Prefilling:** Restating the instruction by prefilling the response to improve instruction-following (Figure A2).

3. **Revision Prompt:** An automated follow-up instruction to refine summaries that do not meet length constraints, with additional prompt prefilling (Figure A3)

Length Measurement. Summary lengths are measured using NLTK (Bird et al., 2009) for word and sentence counts. Token counts are determined based on the BPE tokenizer employed by LLaMA 3. Character counts are straightforwardly computed from the generated text; bullet point counts are by tallying bullet characters. Detailed counting functions are provided in Appendix B.

Metrics. To evaluate the performance of our length control methods, we employ seven metrics: Exact Match (EM), Length Compliance (LC), Length Deviation (LD), Compression Rate (CR), Perplexity (PPL), AlpacaEval, and ROUGE. EM quantifies the proportion of generated lengths that precisely match their targets, while LC measures adherence within a specified tolerance. LD quantifies the absolute difference between observed and target lengths, and CR measures data compression by comparing target and observed lengths. PPL helps assess model uncertainty, providing insight into generation quality. We employ ROUGE as a classic measure of summary quality. In line with recent work on LCS (Yuan et al., 2024), we also use the AlpacaEval protocol (Dubois et al., 2023) as an LLM-as-a-Judge metric to assess overall generation quality. For detailed definitions of these metrics, please refer to Appendix C.

5 Results

This section presents the empirical findings of our study on length controllability using LLaMA 3.

Influence of PP and Temperature. We first investigate how prompt prefilling and sampling temperature affect the model's adherence to length constraints. As shown in Table 3, incorporating PP significantly improves instruction-following and improves both LC and LD across different temperatures. For a target length of 50 words, LC improves from 75.6% without PP to 79.7% with PP at a temperature of 0.7. Lower sampling temperatures slightly enhance length control consistency, but the combination of PP and a temperature of

---

PP	Temp.	# Words	LC (↑)	LD (↓)
✓	0.3	49.6 ± 4.7	80.4%	3.41
✓	0.7	49.6 ± 4.9	79.7%	3.52
✓	1.0	49.6 ± 5.3	78.8%	3.64
✗	0.3	50.9 ± 11.3	79.2%	4.59
✗	0.7	52.7 ± 16.6	75.6%	6.41
✗	1.0	55.0 ± 24.2	71.7%	8.86

(a) Target Word Length of 50

PP	Temp.	# Sentences	EM (↑)	LD (↓)
✓	0.3	1.01 ± 0.10	99.3%	0.01
✓	0.7	1.01 ± 0.11	99.2%	0.01
✓	1.0	1.01 ± 0.10	99.3%	0.01
✗	0.3	1.02 ± 0.10	98.8%	0.02
✗	0.7	1.05 ± 0.41	97.5%	0.05
✗	1.0	1.11 ± 0.69	95.2%	0.11

(b) Target Sentence Length of 1

Table 3: Influence of Prompt Prefilling (PP) and Sampling Temperature (Temp.) on Length Adherence on YTSEG

(a) Length deviation as a function of target length

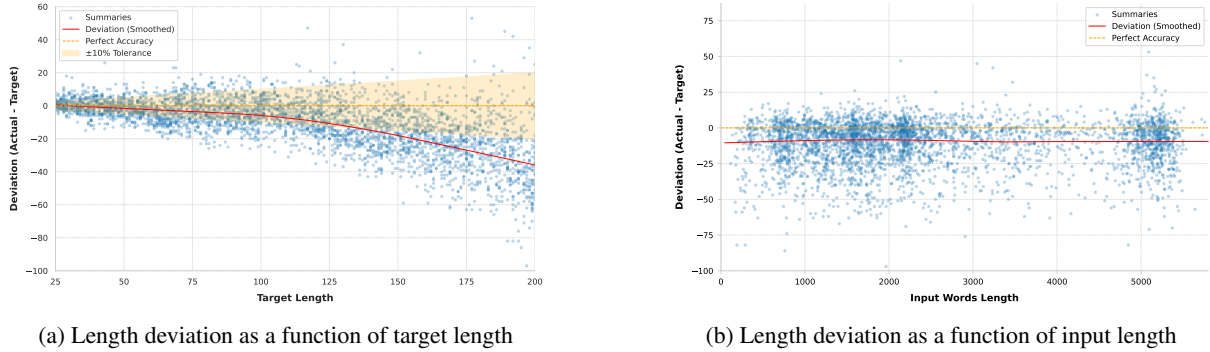(b) Length deviation as a function of input length

Figure 2: Analysis of length deviation across different parameters on YTSEG with data downsampled for visualization. Target summary lengths range from 25 to 200 words. Smoothed trend line using local regression (LOESS).

0.7 provides a good balance between adherence and output diversity. Therefore, we use PP and a temperature of 0.7 in subsequent experiments.

Varying Performance Across Measures. We assessed the model's ability to control summary length across measures by generating summaries with specified length constraints (see Table 2). Our results, presented in Table 4, reveal significant differences in length controllability across measures. Notably, the model demonstrates near-perfect compliance for structural measures, achieving high exact match rates for both sentence and bullet point counts. For granular measures, the performance varies: the model achieves relatively good control over word counts but exhibits significantly lower compliance rates and greater deviations for token and character counts. These varying performances across measures motivate our *length approximation* method, which exploits these differences.

Content Density. Notably, the model adapts its content density based on structural constraints: as seen in Table 4, average sentence length decreases from 50.3 words for one-sentence summaries to 22.0 words for eight sentences. This pattern is mirrored in bullet point summaries, where the average length per bullet point decreases non-linearly from 51.6 words with three bullet points to 27.7 with eight. These trends indicate that when constrained to fewer sentences or bullet points, the

model packs more content into each, rather than simplifying. This contrasts with the relatively consistent sentence lengths observed in word, token, and character-controlled or qualitatively controlled summaries (see Table 4 and 5).

Lexical Complexity. Relatedly, both token-to-word and character-to-word ratios vary with summary length. For example, the token-to-word ratio decreases from 1.28 for 50-word summaries to 1.21 for 200-word summaries, while the character-to-word ratio drops from 6.39 to 6.15 (Table 4). These parallel trends provide strong evidence for shifts in lexical complexity as summary length increases.

Length Deviation Analysis. Figure 2 illustrates a nonlinear relationship between target length and LD, with adherence decreasing for longer summaries. A systematic bias towards under-generation emerges for targets exceeding ~125 words. This trend is further supported by Table 4, indicating an inverse correlation between target length and adherence, with compliance decreasing for longer targets (e.g., 8 sentences, 150–200 words, 8 BPs). Conversely, input length shows minimal correlation with LD, indicating robust length control across diverse document sizes. Scattered data points demonstrate flexibility in handling non-standard word targets. These findings highlight the model's efficacy in producing specified-length summaries while identifying a length-dependent

Measure	Target	EM (%)	LC (%)	LD	# Words	# Tokens	# Chars.	# Units[1]	Unit Len.[1]	C/W Ratio	T/W Ratio
Sentence Length	1 Sentence	99.2	–	–	50.5	63.4	325	1.01	50.3	6.44	1.26
	2 Sentences	99.0	–	–	70.6	86.5	444	2.01	35.2	6.29	1.23
	3 Sentences	98.6	–	–	91.5	111.3	571	3.01	30.5	6.24	1.22
	5 Sentences	96.1	–	–	126.4	152.9	779	4.99	25.4	6.16	1.21
	8 Sentences	84.6	–	–	173.5	209.3	1057	7.88	22.0	6.09	1.21
Word Length	50 Words	–	79.7	3.52	49.6	63.4	317	2.60	20.1	6.39	1.28
	100 Words	–	84.0	6.00	97.9	120.6	611	4.96	20.3	6.24	1.23
	150 Words	–	55.6	15.54	136.9	166.3	847	6.52	21.5	6.19	1.21
	200 Words	–	26.8	33.41	169.8	204.9	1045	7.84	22.2	6.15	1.21
Token Length	50 Tokens	–	0.8	86.48	109.4	136.4	679	6.24	18.5	6.21	1.25
	100 Tokens	–	8.1	63.42	131.4	162.1	814	6.94	19.8	6.20	1.23
	150 Tokens	–	21.7	60.70	166.1	202.4	1024	8.30	20.7	6.16	1.22
	200 Tokens	–	24.1	53.43	179.5	218.0	1105	8.76	21.1	6.15	1.21
Character Length	150 Chars.	–	7.0	120.53	42.0	55.1	269	2.21	20.6	6.40	1.31
	300 Chars.	–	23.8	110.68	61.1	78.2	386	3.28	19.7	6.32	1.28
	500 Chars.	–	26.4	116.45	76.3	96.5	479	4.08	19.7	6.28	1.26
Bullet Point Count	3 Bullet Points	99.8	–	–	154.5	188.9	958	3.00	51.6	6.20	1.22
	5 Bullet Points	99.8	–	–	200.5	246.0	1238	5.00	40.2	6.18	1.23
	8 Bullet Points	98.8	–	–	220.7	275.0	1359	8.00	27.7	6.16	1.25

[1] Number of sentences / bullet points if applicable. Length of a unit is the avg. number of words per unit.

Table 4: Empirical Analysis of Length Controllability Across Measures on YTSEG

bias, motivating targeted improvements in handling longer summaries such as *target adjustment*.

Impact of Qualitative Quantifiers. Analysis of qualitative length quantifiers reveals the model's attempt to adjust summary lengths accordingly (Table 5). These quantifiers yield a range of compression rates and output lengths, with "short" summaries averaging 127.2 ± 44.9 words ($7.0 \pm 4.8\%$ CR) and "long" summaries reaching 343.0 ± 83.9 words ($19.8 \pm 13.5\%$ CR). Notably, heteroscedasticity in output metrics across quantifiers indicates inconsistent control. Figure 3 further illustrates non-linear trends and variability in summary length and CR relative to input length. This inconsistency, combined with heteroscedastic results, suggests that qualitative quantifiers provide imprecise control, limiting their effectiveness for LCS tasks.
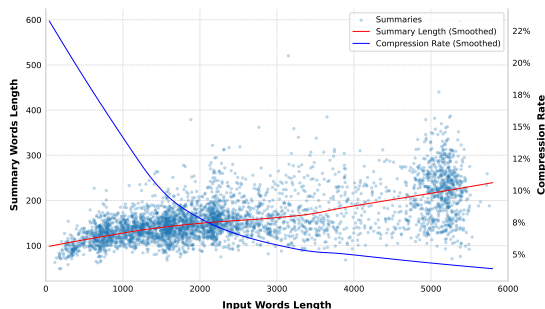

Figure 3: Summary length and CR vs. input length for summaries generated using the "moderate" quantifier.

Cross-Model Generalizability. Additional results in Appendix J provide insight into length controllability and biases across various models. Our analysis compares different LLaMA versions and other LLMs in zero-shot scenarios, highlight-ing that observed patterns of under- and over-generation and preference for certain measures are broadly consistent across architectures. This suggests that the length control challenges and model behaviors identified in our study are not exclusive to a single model configuration, reinforcing the relevance of our findings for a broad range of LLMs.

5.1 Effectiveness of Proposed Methods

The proposed methods were systematically evaluated to assess their impact on length controllability across different measures. Table 6 summarizes the performance of each method.

Length Approximation. The *length approximation* method significantly improved LC in measures where the model underperforms. For instance, in the character length measure, the baseline LC for a target of 150 characters was merely 7.0%. By approximating character counts using word counts, LC increased to 52.1% (Table 6). Similar enhancements were observed in the token length measure, where LC rose from 0.8% to 51.8% for a target of 50 tokens. However, we observe diminishing returns for longer targets as length adherence declines for longer summaries in the word measure.

Transferability of LA Coefficients. The results in Table A2 demonstrate that the statistical mappings used for length approximation are robust and can be effectively applied to datasets beyond the one from which they were originally derived. Specifically, the LA method with coefficients derived from the YTSEG dataset significantly and similarly improved LC for the CNN/DM dataset. For example, in the character length measure, the LC increased from 24.0% in the baseline to 65.6%

Quantifier	CR (%)	# Sentences	# Words	# Tokens	# Chars.	Sent. Len.
<none>	8.8 ± 5.6	7.52 ± 3.92	162.2 ± 58.8	197.0 ± 71.4	1000 ± 360	23.4 ± 8.4
short	7.0 ± 4.8	5.64 ± 2.56	127.2 ± 44.9	154.5 ± 53.0	788 ± 274	23.7 ± 7.6
concise	7.2 ± 4.5	6.30 ± 3.44	133.4 ± 48.1	163.9 ± 59.1	831 ± 298	23.4 ± 10.6
brief	7.4 ± 4.8	6.11 ± 3.00	136.1 ± 50.1	165.4 ± 59.8	843 ± 307	23.7 ± 8.0
moderate	9.0 ± 6.0	7.21 ± 2.83	161.8 ± 52.3	195.5 ± 62.1	1001 ± 322	23.3 ± 5.8
medium-length	10.4 ± 7.3	8.04 ± 2.62	185.2 ± 54.8	223.0 ± 64.6	1140 ± 337	23.5 ± 4.5
comprehensive	15.1 ± 9.2	12.79 ± 5.34	272.7 ± 72.2	329.7 ± 88.7	1677 ± 440	22.9 ± 7.1
verbose	19.7 ± 15.0	13.50 ± 4.48	335.6 ± 92.3	400.0 ± 111.4	2078 ± 552	25.5 ± 4.2
long	19.8 ± 13.5	14.81 ± 4.65	343.0 ± 83.9	409.4 ± 102.0	2090 ± 501	23.9 ± 4.6

Table 5: Influence of Quantifiers on Summary Length on YTSEG

	150 Ch.		300 Ch.		500 Ch.		50 Tok.		100 Tok.		150 Tok.		200 Tok.		50 Wrd.		100 Wrd.		150 Wrd.		200 Wrd.	
	LC	LD	LC	LD	LC	LD	LC	LD	LC	LD	LC	LD	LC	LD	LC	LD	LC	LD	LC	LD	LC	LD
Baseline	7.0	120.5	23.8	110.7	26.4	116.5	0.8	86.5	8.1	63.4	21.7	60.7	24.1	53.4	79.7	3.5	84.0	6.0	55.6	15.5	26.8	33.4
LA	52.1	18.2	66.2	26.1	61.6	46.0	51.8	6.5	47.5	12.2	43.1	18.3	26.9	31.9	–	–	–	–	–	–	–	–
LA-TA	4.3	59.9	66.3	25.9	65.0	43.2	51.5	6.6	61.1	9.8	61.0	14.8	57.0	21.5	–	–	–	–	–	–	–	–
SF	17.4	71.8	51.3	53.1	56.0	59.4	2.1	58.8	20.0	35.6	47.0	28.5	54.0	25.6	98.5	1.4	98.6	2.6	83.5	8.3	53.1	21.2
AR	18.6	57.9	50.2	50.5	47.1	93.7	8.6	25.1	42.2	17.9	54.7	20.6	60.8	22.7	97.6	2.3	98.6	4.4	92.6	8.3	57.0	20.4
SR	46.8	27.9	84.3	20.8	78.3	41.3	26.9	13.6	79.5	7.6	86.8	9.2	87.5	11.8	100.0	1.3	100.0	2.4	99.7	5.6	87.8	11.8
SF-AR	33.0	41.8	71.4	32.5	67.3	65.6	14.1	20.0	57.7	13.1	71.4	14.9	73.9	18.0	99.9	1.3	99.9	2.5	97.2	6.3	74.6	15.5
LA-SF	83.7	8.6	94.1	11.3	92.0	21.0	84.4	3.0	78.8	6.4	71.1	11.0	49.9	21.5	–	–	–	–	–	–	–	–
LA-AR	67.3	23.3	74.8	40.5	69.9	66.7	65.8	10.7	54.1	23.8	51.2	14.9	54.4	28.5	–	–	–	–	–	–	–	–
LA-SF-AR	90.1	10.1	95.6	14.1	93.7	26.3	89.5	4.2	81.2	11.4	74.7	16.7	67.8	22.2	–	–	–	–	–	–	–	–
LA-TA-SF	10.8	42.2	94.1	11.2	93.8	19.4	84.8	3.0	91.4	4.5	89.4	6.9	87.4	10.0	–	–	–	–	–	–	–	–
TA	–	–	–	–	–	–	–	–	–	–	–	–	–	–	79.7	3.5	61.7	9.8	49.3	20.2	49.5	24.8
TA-SF	–	–	–	–	–	–	–	–	–	–	–	–	–	–	98.5	1.4	90.1	4.5	81.4	9.2	83.2	11.4

Table 6: Performance of different methods across character, token, and word length measures on YTSEG

using LA for a target of 150 characters. This improvement suggests that the approximation coefficients generalize well to another dataset with differences in content, style, and structure.

Target Adjustment. The *target adjustment* method showed notable effectiveness for longer targets, particularly where the model's bias towards undergeneration is pronounced. For instance, for a 200-word target, LC improved meaningfully from 26.8% to 49.5% after upward adjustment (Table 6). However, for shorter targets, such as 100 words, LC dropped from 84.0% to 61.7%. This suggests that refining the conversion model or introducing a threshold to activate adjustments only under substantial undergeneration bias could help avoid performance degradation on shorter outputs.

Sample Filtering. *Sample filtering* works best when the model generates at least some compliant summaries, requiring the output distribution to cover the target length with nontrivial density. Table 1 shows that for $N = 3$ samples, notable improvements are achieved if the model initially generates diverse enough outputs. However, when the model has a strong bias toward noncompliance—such as extreme under- or overgeneration—filtering adds little benefit. Figure 4a further shows that LC consistently improves as the number of generated samples (N) increases, though the gains diminish after a certain point.

Automated Revision. The *automated revision* method demonstrates particular effectiveness for longer target lengths across all measures, where initial length compliance is generally lower. In Table 1, we report the results after a single revision, highlighting substantial improvements even when initial compliance is extremely poor—scenarios where sample filtering often fails. For iterative revisions, Figure 4b shows compliance trends over multiple revisions. Notably, the statistics for 200-word summaries reveal that most summaries improve with each revision, with only a small percentage showing any degradation in LD. Moreover, unlike sample filtering, iterative revisions show fewer signs of diminishing returns, enhancing length compliance effectively with each iteration.

Integrated Methods. The integration of different methods yields highly effective strategies for improving length adherence. *Sampled revisions* with $N = 1$ has proven very effective, particularly when the initial performance is reasonable. For cases involving iterative sampled revisions, near-perfect compliance can be achieved across all measures, given sufficient iterations (see Appendix I). In contrast, when the base performance is extremely poor, a different combination proves more effective: length approximation paired with sample filtering. Length approximation helps ensure that the model generates at least some outputs close to the target, thereby making sample filtering significantly more impactful. Additionally, target
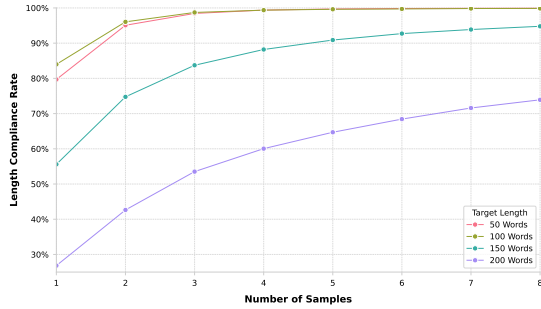
(a) Length compliance vs. number of samples using Sample Filtering approach
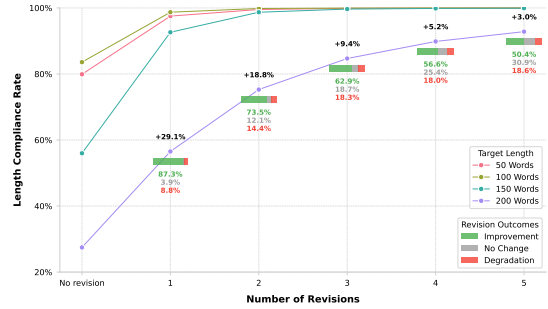
(b) LC across revision steps, with proportions shown of summaries improved (■), unchanged (■), or degraded (■).

Figure 4: Length Compliance with the Number of Revisions or Samples on YTSEG

Method	Characters						Tokens						Words				
	LC	LD	R1	R2	RL		LC	LD	R1	R2	RL		LC	LD	R1	R2	RL
Baseline	17.7	93.6	28.6	9.0	24.6		3.3	60.3	30.8	10.2	25.1		67.7	4.4	30.4	9.4	25.5
LA	58.3	31.2	30.5	9.5	25.7		61.9	6.3	31.1	9.8	26.1		–	–	–	–	–
AR	47.1	50.4	29.2	8.6	24.3		18.4	21.2	31.3	9.8	25.7		92.7	2.8	30.2	9.2	25.3
SF	39.2	62.9	29.0	8.8	24.6		9.1	39.0	31.3	10.1	25.6		94.0	2.0	30.5	9.3	25.5
SF-AR	64.4	37.0	29.5	8.9	24.7		29.8	16.2	31.5	10.0	26.0		98.9	1.7	30.5	9.3	25.5
SR	79.4	23.4	29.4	8.7	24.5		46.2	10.7	31.2	9.8	25.8		99.7	1.6	30.5	9.3	25.5
LA-AR	71.7	33.9	30.0	9.0	25.0		70.3	14.0	30.6	9.4	25.5		–	–	–	–	–
LA-SF	89.0	15.3	30.1	9.1	25.2		89.5	3.1	30.7	9.5	25.6		–	–	–	–	–
LA-SF-AR	93.1	16.0	30.1	9.1	25.2		92.4	4.6	30.7	9.5	25.7		–	–	–	–	–
LA-SR	95.3	13.7	30.1	9.1	25.2		94.2	3.8	30.6	9.5	25.6		–	–	–	–	–

Table 7: Performance of the proposed methods on the CNN/DM dataset, with reference length as target length

adjustment can further improve compliance, especially with longer summaries.

	Characters				Tokens					Words				Avg.
	150	300	500		50	100	150	200		50	100	150	200	
LA	66.0	62.0	58.0		55.0	75.0	59.0	58.0		–	–	–	–	61.9
SF	52.0	44.0	50.0		55.0	51.0	43.0	51.0		44.0	50.0	59.0	54.0	50.3
AR	42.0	47.0	58.0		44.0	49.0	55.0	50.0		53.0	56.0	54.0	51.0	50.8
TA	–	–	–		–	–	–	–		50.0	56.0	46.0	62.0	53.5

Table 8: Win Rates (%; ↑) against the baseline for different measures and target lengths on YTSEG. Evaluating 100 pairs per condition, using GPT-4o as judge.

Summary Quality. Our methods preserve—and in some cases enhance—summary quality. Since they do not modify model parameters, significantly alter prompts, or impose restrictive decoding constraints, we expected minimal impact on quality. This expectation is supported by multiple lines of evidence. First, our perplexity analysis in Table A1 shows no degradation, suggesting that key content quality features such as coherence and fluency are preserved. The only notable difference is a lower perplexity in the LA approach, implying that summaries generated via word measure may achieve higher quality than those based on finer-grained measures like tokens or characters. This finding is further validated through AlpacaEval (see Table 8), where GPT-4o as a judge ranks our controlled summaries comparable to baseline outputs (SF: 50.3%,

AR: 50.8%, TA: 53.5%), with LA again achieving a notably higher win rate of 61.9%. Lastly, the ROUGE scores in Table 7 largely remain stable across all methods, with some approaches slightly improving the performance and achieving substantially better length control. These results confirm that our interventions maintain or enhance content quality while providing superior length control.

6 Conclusion

This study presents a comprehensive analysis of length-controllable summarization using LLMs in zero-shot settings. Our results reveal that LLMs excel in structural measures (e.g., sentences, bullet points) but struggle with granular length control, particularly for character and token counts. To address these challenges, we proposed and evaluated several methods: length approximation, target adjustment, sample filtering, and automated revisions. These techniques significantly enhance length compliance across various measures without compromising summary quality, with integrated approaches yielding the most substantial improvements and achieving near-perfect compliance in some cases. Our findings advance the understanding of LLM behavior in controlled text generation and provide practical strategies for implementing precise length control.

Limitations

While this study aims to be comprehensive, examining multiple length measures, targets, and control methods across a diverse dataset, and several key findings were validated through cross-model and cross-dataset experiments, some limitations remain. Additional research could extend these findings across an even broader spectrum of models and domains. The proposed methods demonstrate significant improvements in length controllability, with approaches like length approximation and target adjustment being virtually cost-free. However, certain techniques, particularly those involving multiple samples or iterations, increase computational resources and inference time. While the study primarily addresses length control in summarization tasks, the effectiveness of these methods in other text generation scenarios remains to be explored.

Acknowledgements

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*, 1st edition. O'Reilly, Beijing ; Cambridge [Mass.]. OCLC: ocn301885973.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024a. Iterative Translation Refinement with Large Language Models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

Yihan Chen, Benfeng Xu, Quan Wang, Yi Liu, and Zhendong Mao. 2024b. Benchmarking Large Language Models on Controllable Generation under Diversified Instructions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17808–17816. Number: 16.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, ..., and Zhiwei Zhao. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S. Liang, and Tatsunori B. Hashimoto. 2023. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069.

Marcio Fonseca and Shay Cohen. 2024. Can Large Language Model Summarizers Adapt to Diverse Scientific Communication Goals? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8599–8618, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. *arXiv preprint*. ArXiv:2209.12356 [cs].

Chiori Hori, Sadaoki Furui, Rob Malkin, Hua Yu, and Alex Waibel. 2002. Automatic summarization of english broadcast news speech. In *Proceedings of the second international conference on Human Language Technology Research*, pages 241–246.

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 and virtual meeting. Association for Computational Linguistics.

Juseon-Do 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 and virtual meeting. Association for Computational Linguistics.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling Output Length in Neural Encoder-Decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, ..., and Oriol Vinyals. 2022. Competition-Level Code Generation with AlphaCode. *Science*, 378(6624):1092–1097. ArXiv:2203.07814 [cs].

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable Text Generation for Large Language Models: A Survey. *arXiv preprint*. ArXiv:2408.12599 [cs].

Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. 2024. "We Need Structured Output": Towards User-centered Constraints on Large Language Model Output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 10:1–10:9. ACM.

Yizhu Liu, Qi Jia, and Kenny Zhu. 2022. Length Control in Abstractive Summarization by Pretraining Information Selection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6885–6895, Dublin, Ireland. Association for Computational Linguistics.

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

Takuya Makino, Tomoya Iwakura, Hiroya Takamura, and Manabu Okumura. 2019. Global Optimization under Length Constraint for Neural Text Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1039–1048, Florence, Italy. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. *arXiv preprint*. ArXiv:2302.12813.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (Almost) Dead. *arXiv preprint*. ArXiv:2309.09558 [cs].

Zhiqiang Rao, Hengchao Shang, Jinlong Yang, Daimeng Wei, Zongyao Li, Jiaxin Guo, Shaojun Li, Zhengzhe Yu, Zhanglin Wu, Yuhao Xie, Bin Wei, Jiawei Zheng, Lizhi Lei, and Hao Yang. 2023. Length-Aware NMT and Adaptive Duration for Automatic Dubbing. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 138–143, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Fabian Retkowski. 2023. The current state of summarization. In Andreas Sudmann, Anna Echterhölter, Markus Ramsauer, Fabian Retkowski, Jens Schröter, and Alexander Waibel, editors, *Beyond Quantity. Research with Subsymbolic AI*, 1 edition, pages 291–312. transcript Verlag, Bielefeld, Germany.

Fabian Retkowski and Alexander Waibel. 2024. From Text Segmentation to Smart Chaptering: A Novel Benchmark for Structuring Video Transcriptions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 406–419. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating Large Language Models on Controlled Generation Tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Ashok Urlana, Pruthwik Mishra, Tathagato Roy, and Rahul Mishra. 2024. Controllable Text Summarization: Unraveling Challenges, Approaches, and Prospects - A Survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1603–1623, Bangkok, Thailand. Association for Computational Linguistics.

Alex Waibel, Hartwig Steusloff, Rainer Stiefelhagen, and others. 2004. CHIL: Computers in the human interaction loop. In *Proceedings of the 5th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2004*. Place: Lisboa, Portugal.

Alexander Waibel and Christian Fuegen. 2012. Simultaneous translation of open domain lectures and speeches. Published: US Patent No. 8,090,570.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating Sequences by Learning to Self-Correct. *arXiv preprint*. ArXiv:2211.00053.

Wen Xiao, Yujia Xie, Giuseppe Carenini, and Pengcheng He. 2023. ChatGPT-steered Editing Instructor for Customization of Abstractive Summarization. *arXiv preprint*. ArXiv:2305.02483 [cs].

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization. *arXiv preprint*. ArXiv:2302.08081 [cs].

Zhongyi Yu, Zhenghao Wu, Hao Zheng, Zhe XuanYuan, Jefferson Fong, and Weifeng Su. 2021. LenAtten: An Effective Length Controlling Unit For Text Summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 363–370, Online. Association for Computational Linguistics.

Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. 2024. Following Length Constraints in Instructions.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57. Place: Cambridge, MA Publisher: MIT Press.

A Prompt Templates

System

You are an assistant who replies with a summary to every message.

User

Summarize the following text in `{length}` words: \n\n `{input}`
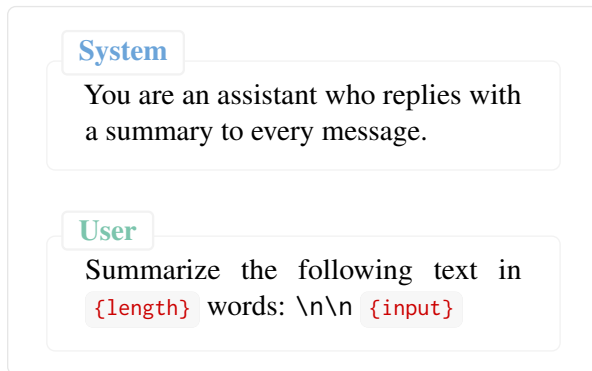
Figure A1: Simple Prompting

System

You are an assistant who replies with a summary to every message.

User

Summarize the following text in `{length}` words: \n\n `{input}`

Assistant

Sure! Here is a summary of the text in `{length}` words: \n\n
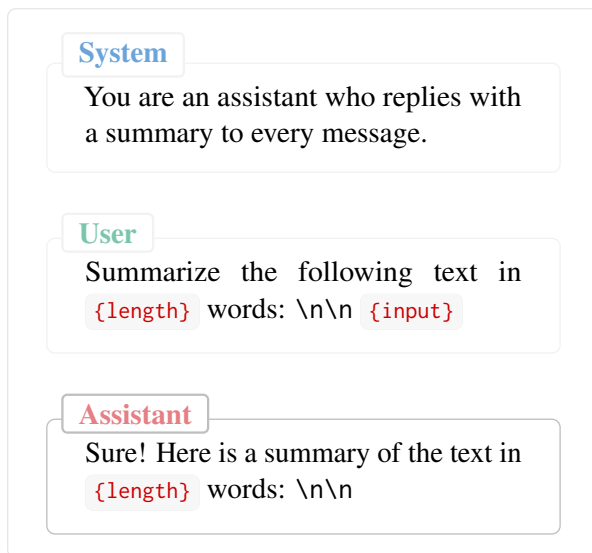
Figure A2: Prompting with **Prompt Prefilling**

System

You are an assistant who replies with a summary to every message.

User

Summarize the following text in `{length}` words: \n\n `{input}`

Assistant

Sure! Here is a summary of the text in `{length}` words: \n\n `{summary}`

User

Your summary has `{summary_length}` words which is `{length_difference}` words `{more_less}` than the requested length. Please revise the summary to be closer to the requested length of `{length}` words.

Assistant

Sure! Here is a revised summary that is closer to the requested length of `{length}` words: \n\n

Figure A3: Prompting for **Automated Revisions**

In Figures A1, A2, and A3, we present the different prompt templates using the word measure as an example. For other measures, such as character or token-length-controlled summaries, the word measure can be substituted accordingly. Additionally, for the bullet point measure, we introduce the bullet symbol (• or \u2022) as an additional prefix in the response to implicitly define the typographical symbol to be used for the bullet-point format. The same symbol is then used in the counting function (see Listing A5).

B Counting Functions

```python
from nltk.tokenize import RegexpTokenizer

def count_words(text) -> int:
    tokenizer = RegexpTokenizer(r'\w+')
    return len(tokenizer.tokenize(text))
```

Listing A1: Word Count Function

```python
from transformers import AutoTokenizer

model_id = "meta-llama/Meta-Llama-3-8B-Instruct"
tokenizer = AutoTokenizer.from_pretrained(model_id, use_fast=True)

def count_tokens(text) -> int:
    tokens = tokenizer(text, truncation=False, return_tensors="pt")
    return len(tokens["input_ids"][0])
```

Listing A2: Token Count Function

```python
def count_characters(text) -> int:
    return len(text)
```

Listing A3: Character Count Function

```python
from nltk.tokenize import sent_tokenize

def count_sents(text) -> int:
    return len(sent_tokenize(text))
```

Listing A4: Sentence Count Function

```python
def count_bps(text) -> int:
    return text.count('\u2022')
```

Listing A5: Bullet Point Count Function

C Length Metrics

In the following sections, N denotes the number of instances, $L_{y,i}$ represents the observed length for the i-th instance, and $L_{t,i}$ signifies the target length.

C.1 Exact Match

Exact Match (EM) quantifies the proportion of observed lengths $L_{y,i}$ that exactly match their target lengths $L_{t,i}$, defined as

$$\text{EM} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(L_{y,i} = L_{t,i}\right), \tag{12}$$

EM provides a normalized measure of the frequency with which observed lengths precisely adhere to their target values, particularly useful for structural length measures.

C.2 Length Compliance

Length Compliance (LC) quantifies the proportion of observed lengths $L_{y,i}$ that are within a tolerance T of their target lengths $L_{t,i}$, defined as

$$\text{LC} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(\left|L_{y,i} - L_{t,i}\right| \leq TL_{t,i}\right), \tag{13}$$

LC provides a normalized measure of adherence within a specified relative tolerance, offering a more flexible assessment of granular measures than EM by allowing slight deviations.

C.3 Length Deviation

Length Deviation (LD) measures the average absolute difference between observed lengths $L_{y,i}$ and target lengths $L_{t,i}$, defined as

$$\text{LD} = \frac{1}{N} \sum_{i=1}^{N} \left|L_{y,i} - L_{t,i}\right|, \tag{14}$$

LD quantifies the average deviation from target lengths. Unlike normalized measures, it reflects absolute errors, making it scale-sensitive and limiting comparability across varying length scales.

C.4 Compression Rate

Compression Rate (CR) quantifies the average ratio of target lengths $L_{t,i}$ to observed lengths $L_{y,i}$, defined as

$$\text{CR} = \frac{1}{N} \sum_{i=1}^{N} \frac{L_{t,i}}{L_{y,i}}, \tag{15}$$

CR provides insight into the relative compression of target data by comparing target and observed lengths. It serves as a descriptive statistic rather than an evaluation metric.

D   Perplexity Scores

	Characters			Tokens				Words			
	150	300	500	50	100	150	200	50	100	150	200
Baseline	27.38	14.77	9.87	19.51	10.13	8.15	6.99	14.08	8.78	7.16	6.28
LA	24.89	13.11	9.35	14.90	8.89	7.14	6.31	–	–	–	–
LA-TA	19.98	13.08	9.20	15.49	9.11	7.42	6.53	–	–	–	–
SF	28.20	14.40	9.79	18.09	9.93	7.98	7.02	13.51	8.64	7.16	6.36
AR	31.26	15.50	10.38	20.76	10.96	8.35	7.09	14.08	8.81	7.12	6.37
SR	31.10	15.19	10.25	20.21	10.60	8.07	7.00	13.50	8.64	7.09	6.34
SF-AR	30.05	14.97	10.05	20.12	10.57	8.07	7.03	13.49	8.63	7.12	6.35
LA-SF	27.12	13.74	9.73	15.46	9.27	7.44	6.55	–	–	–	–
LA-AR	29.03	14.11	9.93	15.83	9.19	7.36	6.49	–	–	–	–
LA-SF-AR	27.40	13.69	9.77	15.51	9.30	7.48	6.61	–	–	–	–
LA-TA-SF	21.73	13.72	9.59	16.01	9.37	7.52	6.66	–	–	–	–
TA	–	–	–	–	–	–	–	14.08	8.88	7.48	6.59
TA-SF	–	–	–	–	–	–	–	13.51	8.82	7.37	6.58

Table A1: Perplexity (PPL $\downarrow$) of various methods across character, token, and word-length measures. Average perplexity is calculated solely for length-compliant samples to minimize potential length bias and ensure fair comparison within a target. We use the pretrained `Llama-3-8B` language model to calculate the perplexity.

E   Additional Experimental Details

Revision Trigger.   In the automated revision process, we set the revision trigger threshold ϵ equal to the tolerance T used in the LC metric ($\pm 10\%$ of target length). However, it's worth noting that in practice, this threshold can be adjusted to limit revisions for summaries that excessively deviate from the length target, balancing accuracy with computational efficiency.

Prompt Prefilling and Temperature.   The impact of prompt prefilling and sampling temperature $\tau \in \{0.3, 0.7, 1.0\}$ on length adherence was initially investigated. For subsequent experiments, we assumed the usage of PP and a temperature of 0.7, based on their performance in preliminary tests.

Context Length.   The limited maximum context length of LLaMA 3 (8,192 tokens) imposes constraints on the input prompt. Since we reserve at least 1,024 tokens for generation, only 7,168 tokens remain available for the input, which includes both the user instruction and the source document. Given that the dataset contains numerous lengthy documents, many of which exceed this limit, any necessary truncation is applied to the original source text to ensure it fits within the remaining context window.

AlpacaEval Evaluation.   We used GPT-4o (`gpt-4o-2024-05-13`) to evaluate model outputs following the AlpacaEval protocol (Dubois et al., 2023) and prompt[2]. For each method and target length combination, we selected the 100 summary pairs where baseline and method outputs had the least length difference to minimize length bias. The total cost for evaluation was $51.73. The numbers indicate win rates in percentage against the baseline model. Given the sample size per condition, individual cell values should be interpreted with consideration for sampling uncertainty, while averaged results across conditions provide reliable indicators of overall performance.

Length Approximation and Target Adjustment Coefficients.   To calculate the length approximation coefficients, we aggregated summaries from the LLaMA 3 baseline experiments on YTSEG targeting word counts and character counts (see Table 4). We computed the average word length in characters and the average number of tokens per word across these summaries. For the target adjustment method, we derived a polynomial regression based on generated summaries of the YTSEG dataset with random word targets between 25 and 300 words.

---

[2]https://github.com/tatsu-lab/alpaca_eval/blob/0b4af760/src/alpaca_eval/evaluators_configs/chatgpt/basic_prompt.txt

F Length Approximation and Target Adjustment

F.1 Results on CNN/DM

		150 Ch.		300 Ch.		500 Ch.		50 Tok.		100 Tok.		150 Tok.		200 Tok.	
		LC	LD	LC	LD	LC	LD	LC	LD	LC	LD	LC	LD	LC	LD
YTSEG	Baseline	7.0	120.5	23.8	110.7	26.4	116.5	0.8	86.5	8.1	63.4	21.7	60.7	24.1	53.4
	LA	52.1	18.2	66.2	26.1	61.6	46.0	51.8	6.5	47.5	12.2	43.1	18.3	26.9	31.9
CNN/DM	Baseline	24.0	44.6	30.2	57.0	4.7	182.3	3.1	49.5	18.1	33.2	40.1	24.9	28.4	37.8
	LA	65.6	13.2	66.3	25.6	56.6	49.5	53.7	6.0	57.1	10.2	53.5	15.4	29.6	29.4

Table A2: Performance comparison of Baseline and Length Approximation methods on character and token targets for YTSEG and CNN/DM, using LA coefficients derived from YTSEG generations

		50 Words		100 Words		150 Words		200 Words	
		LC	LD	LC	LD	LC	LD	LC	LD
YTSEG	Baseline	79.7	3.5	84.0	6.0	55.6	15.5	26.8	33.4
	TA	79.7	3.5	61.7	9.8	49.3	20.2	49.5	24.8
CNN/DM	Baseline	86.7	2.9	87.9	5.4	58.5	14.7	9.7	39.9
	TA	86.7	2.9	61.2	9.4	62.4	13.8	54.6	21.7

Table A3: Performance comparison of Baseline and Target Adjustment methods on word targets for YTSEG and CNN/DM, using the TA regression model trained on YTSEG generations

F.2 Additional Data Statistics

Target Len.	Approx. Measure	# Words	# Chars.	LC (↑)	LD (↓)
150 Chars.	24 Words	24.3 ± 2.9	162 ± 20	52.1%	18.16
300 Chars.	48 Words	45.6 ± 5.2	293 ± 33	66.2%	26.09
500 Chars.	79 Words	76.7 ± 9.1	482 ± 55	61.6%	46.04

Table A4: Character Length Approximation

Target Len.	Approx. Measure	# Words	# Tokens	LC (↑)	LD (↓)
50 Tokens	40 Words	40.5 ± 5.8	52.2 ± 8.2	51.8%	6.48
100 Tokens	80 Words	73.6 ± 8.9	91.7 ± 12.2	47.5%	12.22
150 Tokens	120 Words	109.5 ± 10.2	134.1 ± 14.3	43.1%	18.28
200 Tokens	160 Words	139.3 ± 14.3	169.3 ± 18.9	26.9%	31.90

Table A5: Token Length Approximation

Target Len.	Adjust. Target	# Words	LC (↑)	LD (↓)
50 Words	50 Words	49.6 ± 4.9	79.7%	3.52
100 Words	113 Words	107.3 ± 10.1	61.7%	9.78
150 Words	188 Words	164.4 ± 22.8	49.3%	20.17
200 Words	250 Words	209.3 ± 29.9	49.5%	24.77

Table A6: Target Adjustment for Word Measure

G Results on Sample Filtering

Number of Samples	150 Chars.		300 Chars.		500 Chars.	
	LC	LD	LC	LD	LC	LD
1 Sample	7.2%	121.27	24.5%	109.62	26.2%	117.78
2 Samples	12.7%	87.27	40.3%	70.41	43.7%	76.78
3 Samples	17.4%	71.97	51.3%	53.11	56.0%	59.44
4 Samples	21.4%	63.36	58.8%	44.34	63.9%	50.06
5 Samples	24.9%	56.44	65.0%	37.58	70.2%	43.17
6 Samples	28.4%	50.98	69.5%	32.81	74.8%	38.24
7 Samples	31.0%	47.19	73.1%	29.20	78.1%	34.54
8 Samples	33.4%	44.16	75.8%	26.49	80.7%	31.72

Table A7: Performance at different sample counts on the character length measure

Number of Samples	50 Tokens		100 Tokens		150 Tokens		200 Tokens	
	LC	LD	LC	LD	LC	LD	LC	LD
1 Sample	0.7%	90.89	7.0%	68.80	19.8%	68.07	23.8%	57.70
2 Samples	1.4%	68.20	14.1%	44.99	35.7%	39.20	41.4%	34.47
3 Samples	2.1%	58.87	19.9%	35.66	47.0%	28.55	53.9%	25.63
4 Samples	2.6%	54.31	23.7%	31.46	54.0%	24.30	62.2%	21.41
5 Samples	3.3%	49.81	27.9%	27.47	60.6%	20.21	68.8%	18.07
6 Samples	4.2%	46.36	31.6%	24.62	65.6%	17.48	73.8%	15.70
7 Samples	4.7%	44.03	34.8%	22.67	69.5%	15.49	77.7%	14.00
8 Samples	5.4%	41.99	37.8%	21.07	72.7%	13.91	80.7%	12.64

Table A8: Performance at different sample counts on the token length measure

Number of Samples	50 Words		100 Words		150 Words		200 Words	
	LC	LD	LC	LD	LC	LD	LC	LD
1 Sample	80.0%	3.49	84.3%	5.95	57.3%	14.99	28.7%	32.66
2 Samples	95.0%	1.93	96.0%	3.54	75.3%	10.45	43.7%	24.85
3 Samples	98.4%	1.36	98.6%	2.58	83.6%	8.30	53.1%	21.07
4 Samples	99.4%	1.06	99.4%	2.03	88.5%	6.96	60.0%	18.52
5 Samples	99.7%	0.85	99.7%	1.67	91.0%	6.11	65.3%	16.75
6 Samples	99.8%	0.71	99.8%	1.43	92.7%	5.48	68.8%	15.54
7 Samples	99.9%	0.61	99.8%	1.25	93.8%	4.99	72.0%	14.40
8 Samples	99.9%	0.53	99.9%	1.11	94.6%	4.60	74.2%	13.58

Table A9: Performance at different sample counts on the word length measure

H Results on Iterative Automated Revisions

Number of Revisions	150 Chars.		300 Chars.		500 Chars.	
	LC	LD	LC	LD	LC	LD
No Revision	6.7%	121.15	23.7%	110.25	26.3%	115.75
Revision 1	18.5%	57.74	50.1%	51.00	46.9%	93.70
Revision 2	32.6%	39.51	71.1%	32.15	68.2%	57.09
Revision 3	44.8%	31.43	82.5%	26.46	78.0%	50.63
Revision 4	54.1%	26.98	89.0%	21.98	85.6%	39.86
Revision 5	61.7%	23.71	92.9%	19.57	90.1%	36.73

Table A10: Performance at different revision steps on the character length measure

Number of Revisions	50 Tokens		100 Tokens		150 Tokens		200 Tokens	
	LC	LD	LC	LD	LC	LD	LC	LD
No Revision	0.7%	86.42	7.9%	63.60	22.1%	60.61	24.7%	54.50
Revision 1	8.6%	25.27	43.0%	17.95	55.1%	20.17	60.7%	22.52
Revision 2	29.3%	13.51	67.8%	12.39	72.9%	16.49	72.4%	20.93
Revision 3	44.7%	10.52	81.2%	9.86	83.0%	13.36	82.7%	16.50
Revision 4	55.3%	8.89	88.6%	7.86	88.4%	11.58	87.7%	14.63
Revision 5	63.3%	7.62	93.0%	6.67	92.0%	9.94	91.9%	12.97

Table A11: Performance at different revision steps on the token length measure

Number of Revisions	50 Words		100 Words		150 Words		200 Words	
	LC	LD	LC	LD	LC	LD	LC	LD
No Revision	79.9%	3.50	83.6%	6.07	56.0%	15.36	27.5%	33.50
Revision 1	97.5%	2.32	98.7%	4.38	92.6%	8.17	56.5%	20.57
Revision 2	99.5%	2.21	99.8%	4.28	98.7%	7.34	75.3%	16.20
Revision 3	99.9%	2.19	100.0%	4.26	99.7%	7.21	84.7%	14.42
Revision 4	100.0%	2.18	100.0%	4.26	99.9%	7.19	89.8%	13.45
Revision 5	100.0%	2.18	100.0%	4.26	99.9%	7.18	92.8%	12.94

Table A12: Performance at different revision steps on the word length measure

I   Results on Sampled Revisions

Number of Revisions	150 Chars.		300 Chars.		500 Chars.	
	LC	LD	LC	LD	LC	LD
Initial Sample Filt.	18.0%	71.42	51.2%	53.41	57.0%	58.81
Sampled Rev. 1	46.8%	27.86	84.3%	20.75	78.3%	41.28
Sampled Rev. 2	68.9%	16.38	95.6%	14.56	92.8%	26.92
Sampled Rev. 3	81.1%	12.29	98.7%	13.13	96.4%	24.63
Sampled Rev. 4	88.0%	10.28	99.5%	12.70	98.4%	22.81
Sampled Rev. 5	91.8%	9.23	99.8%	12.55	99.2%	22.29

Table A13: Performance at different revision steps on the character length measure

| Number of Revisions	50 Tokens		100 Tokens		150 Tokens		200 Tokens	
	LC	LD	LC	LD	LC	LD	LC	LD
Initial Sample Filt.	2.2%	58.64	19.8%	35.56	48.1%	28.23	53.0%	26.00
Sampled Rev. 1	26.9%	13.63	79.5%	7.60	86.8%	9.24	87.5%	11.75
Sampled Rev. 2	62.8%	6.25	93.8%	5.42	94.1%	7.75	92.9%	10.85
Sampled Rev. 3	77.4%	4.69	98.1%	4.51	97.4%	6.89	97.1%	9.50
Sampled Rev. 4	84.6%	3.91	99.1%	4.31	98.5%	6.66	98.5%	9.23
Sampled Rev. 5	88.8%	3.52	99.6%	4.22	99.2%	6.49	99.3%	8.99

Table A14: Performance at different revision steps on the token length measure

Number of Revisions	50 Words		100 Words		150 Words		200 Words	
	LC	LD	LC	LD	LC	LD	LC	LD
Initial Sample Filt.	98.3%	1.35	98.5%	2.62	83.0%	8.47	52.9%	21.22
Sampled Rev. 1	100.0%	1.25	100.0%	2.44	99.7%	5.57	87.8%	11.76
Sampled Rev. 2	100.0%	1.25	100.0%	2.44	100.0%	5.54	96.4%	10.31
Sampled Rev. 3	100.0%	1.25	100.0%	2.44	100.0%	5.54	98.6%	9.99
Sampled Rev. 4	100.0%	1.25	100.0%	2.44	100.0%	5.54	99.4%	9.90
Sampled Rev. 5	100.0%	1.25	100.0%	2.44	100.0%	5.54	99.6%	9.86

Table A15: Performance at different revision steps on the word length measure

J Cross-Model Experiments

J.1 Length Deviation Analysis
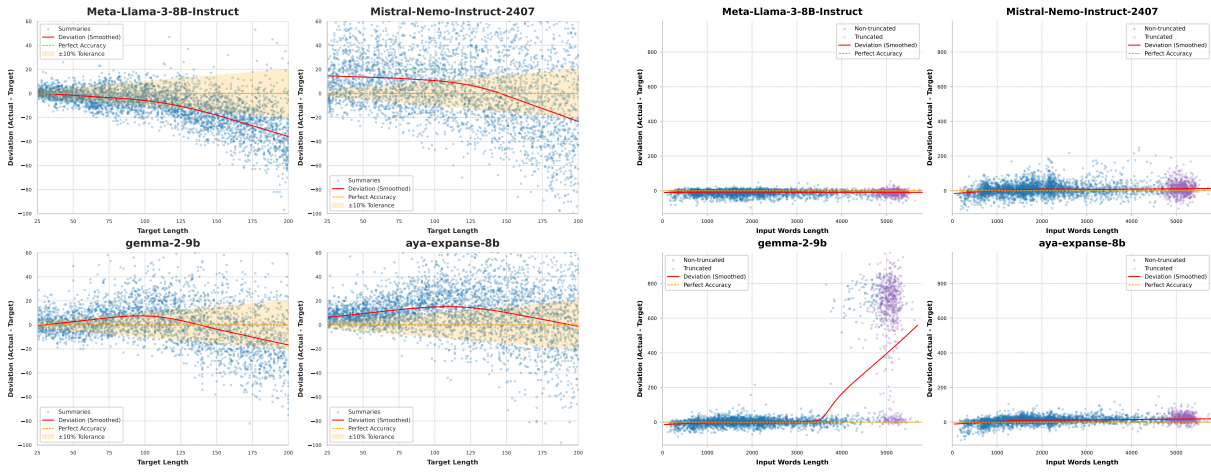
(a) Length deviation as a function of target length

(b) Length deviation as a function of input length

Figure A4: Analysis of length deviation across different parameters and models with data downsampled for visualization. Target summary lengths range from 25 to 200 words. Smoothed trend line using local regression.

J.2 Baseline Performance Comparison

Model	150 Chars.		300 Chars.		500 Chars.	
	LC	LD	LC	LD	LC	LD
Llama-3-8B-Instruct	7.0%	120.6	23.8%	110.7	26.4%	116.5
Llama-3.1-8B-Instruct	28.8%	37.4	29.3%	71.2	24.4%	119.0
Llama-3.2-11B-Vision-Instruct	35.0%	31.6	26.5%	69.3	20.5%	129.4
Mistral-Nemo-Instruct-2407	8.7%	146.9	11.9%	205.9	21.9%	204.2
gemma-2-9b-it	9.6%	769.7	15.3%	805.7	34.5%	812.5
aya-expanse-8b	15.4%	78.0	24.7%	120.6	21.9%	204.1

Table A16: Character length control performance comparison across different models

Model	50 Tokens		100 Tokens		150 Tokens		200 Tokens	
	LC	LD	LC	LD	LC	LD	LC	LD
Llama-3-8B-Instruct	0.8%	86.5	8.1%	63.4	21.7%	60.7	24.1%	53.4
Llama-3.1-8B-Instruct	0.1%	141.3	2.0%	99.6	8.9%	97.0	21.6%	82.0
Llama-3.2-11B-Vision-Instruct	0.1%	127.7	3.4%	88.1	10.0%	93.8	20.6%	80.0
Mistral-Nemo-Instruct-2407	0.7%	66.9	13.3%	64.7	23.0%	60.5	25.9%	61.1
gemma-2-9b-it	15.3%	186.6	22.2%	196.5	35.5%	186.8	34.0%	186.8
aya-expanse-8b	1.1%	64.2	15.7%	62.9	22.6%	61.0	25.3%	62.7

Table A17: Token length control performance comparison across different models

Model	50 Words		100 Words		150 Words		200 Words	
	LC	LD	LC	LD	LC	LD	LC	LD
Llama-3-8B-Instruct	79.7%	3.5	84.0%	6.0	55.6%	15.5	26.8%	33.4
Llama-3.1-8B-Instruct	91.7%	2.6	96.2%	3.8	86.7%	8.7	67.2%	17.5
Llama-3.2-11B-Vision-Instruct	93.1%	2.4	96.4%	3.7	88.1%	8.2	72.7%	16.1
Mistral-Nemo-Instruct-2407	14.5%	26.7	29.8%	31.8	34.1%	42.5	28.1%	54.7
gemma-2-9b-it	35.2%	151.5	50.3%	149.7	24.4%	153.5	16.4%	155.2
aya-expanse-8b	33.8%	10.4	38.4%	20.4	46.5%	25.0	48.6%	29.7

Table A18: Word length control performance comparison across different models

J.3 Detailed Cross-Model Analysis

Our comparative analysis across different model architectures reveals several key patterns in length controllability:

Measure-Specific Performance. All models demonstrate their strongest performance when controlling output length at the word level. LLaMA-based models, in particular, show remarkable accuracy; for instance, LLaMA-3.2-11B-Vision-Instruct achieves up to 96.4% compliance with a 100-word target. In contrast, character-level and token-level controls result in significantly lower compliance rates, highlighting that finer-grained measures are more challenging to master.

Input Length Sensitivity. For most models, the length of the input document has minimal impact on the precision of their length control. Notably, however, gemma-2-9b-it proves unstable when input exceeds about 3000 words. The model enters a generation loop regardless of whether the input is truncated, underscoring a general and severe limitation in handling long inputs.

Target Length Effects. Non-LLaMA models (Mistral-Nemo-Instruct-2407, gemma-2-9b-it, and aya-expanse-8b) tend to overgenerate content for shorter targets. For example, at 50-word targets, their compliance can be as low as 14.5%, with some summaries substantially exceeding the requested length. However, as target lengths increase beyond approximately 125 words, these models—like LLaMA variants—begin to undergenerate, often undershooting the requested length by tens of words. This pattern suggests a fundamental and broad-based difficulty for current language models: they struggle to maintain precise length control as target length increases.

LLaMA Model Evolution. More recent LLaMA generations (3.1, 3.2) consistently outperform their predecessors, especially for the word measure. Not only do they achieve near-perfect compliance for shorter word targets, but they also exhibit more stable performance as target length grows.

Model-Specific Behaviors. While non-LLaMA models generally lag behind in length control, some display distinctive strengths. For instance, gemma-2-9b-it shows relatively strong token-level compliance (up to 35.5%), even though it struggles with longer input documents. These model-specific quirks, combined with the broader patterns observed, reinforce our overall findings and add nuance to our understanding of how different architectures tackle the challenge of controlling output length.