Dynamic Summarization Length Control via Model Merging

Anonymous authors

Abstract

Controlling the length of generated text is crucial in tasks like summarization, where different applications demand varying degrees of brevity. We propose a model-merging-based method to dynamically control the length of summarization during inference time, which avoids architectural changes, decoding interventions, and complex training setups. For this, we build two custom dataset divisions curated for short and long summarization lengths by combining the XSum and CNN/DM datasets and utilizing an LLM-based data augmentation pipeline. We then fine-tune two instances of an LLM on these two divisions to generate different lengths of summaries. Finally, we merge the weights of these models using linear interpolation to dynamically control the summarization length. We observe that length can be adjusted smoothly by varying interpolation weights. Moreover, the calculated ROUGE scores of the summaries show that the quality of summarization remains intact. Our analyses on ablating different merging techniques, dataset preparation considerations, and model size/family demonstrate that model merging can reliably enable length control across diverse models and configurations.

1 Introduction

Summarization is an important problem that has been studied extensively. Summarization to a specific length has also been studied (Liu et al., 2022; Miculicich et al., 2023; Anonymous, 2024; Kikuchi et al., 2016; Liu et al., 2018; Makino et al., 2019; Yu et al., 2021; See et al., 2017; Sarkhel et al., 2020; Saito et al., 2020) which is important in contexts requiring presentation of summaries on different mobile devices and websites, where space limit is a factor. Prior work on length-controllability of summarization focuses on generating summaries of fixed lengths, split into two categories: decoding early-stop and encoding information selection.

Decoding early-stop methods (Kikuchi et al., 2016; Liu et al., 2018; Makino et al., 2019; Yu et al., 2021) focus on controlling the timing of the EOS token, to conclude the generated summary. They ignore the content encoding step, which should be adjusted per length requirements, instead relying wholly on decoders. This results in unnatural summarizations that look truncated.

Encoding information selection methods (See et al., 2017; Sarkhel et al., 2020; Saito et al., 2020) encode information from the source text in the first stage. Then, this encoded information is summarized to target length. This results in noise from intermediate step and a longer summarization length, meaning weaker length control. Another method of note is LAAM (Liu et al., 2022), which is a mix of decoding and encoding methods. It utilizes attention scores between encoder and decoder to boost the attention scores of the tokens corresponding to the target length. This has the benefit of considering both encoding and decoding stages. However, the method requires a transformer seq2seq model, not applicable to decoder-only architecture.

In this paper, we develop **dynamic length-control** (DLC) for which we work with a decoder-only model (Llama 3.2 (Grattafiori et al., 2024)) and design a mechanism to dynamically obtain the generated summary length. For our main experiments, we first create two dataset divisions, one for a shorter length (< 128) and another for a longer length (> 256). For shorter length[1] summaries, we utilize an LLM (Fig. 2) to augment the dataset. We train two PEFT (Mangrulkar et al., 2022) checkpoints, corresponding to each divisions of shorter and longer lengths. Then, we interpolate the weights of checkpoints via a ratio w. We verify the summary length distributions (Fig. 4) and find that the lengths each satisfy a Gaussian distribution

[1] Our findings echo that of (Liu et al., 2022) - shorter length summaries are particularly rare.

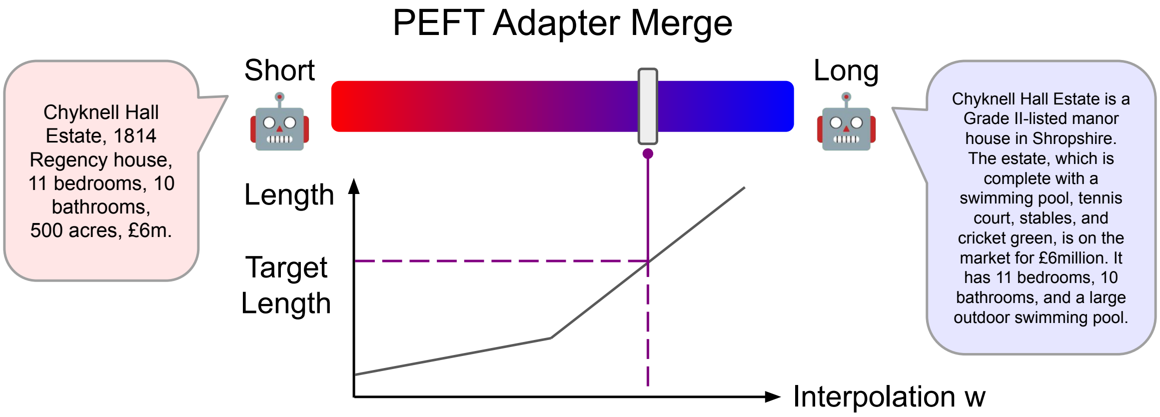


Figure 1: Our generation method. We perform dynamic merging of PEFT adapter weights to obtain arbitrary target lengths, which we measure and perform generations of.

with average pertaining to the combination ratio w of generative models with shorter and longer lengths. Moreover, the evolution of mean summary lengths exhibit a smooth trend curve, complemented with our ablations that the framework is applicable across various settings / models.

Our contributions are as follows:

1. We propose a new dynamic length-control (DLC) paradigm to generate summaries of desired lengths. The mechanism has the benefit of not requiring any additional training to generate summaries of arbitrary length.

2. We develop an LLM data augmentation pipeline to obtain summaries up to arbitrary length. We utilize the pipeline to obtain our 4 dataset divisions (Short, Long, Shortest, Long-Low-Variance).

3. We conduct ablation studies across different weight-merging methods and base models, accompanied by a detailed examination of summary length evolution and distribution. We analyze summary quality through ROUGE scores (Lin, 2004) and complemented with a qualitative analysis of representative outputs.

2 Related Work

Most prior length-control works have focused on stopping decoding at a given step. (Rush et al., 2015) generates end-of-sentence token by $-\infty$ assignment to vocabulary, generating a fixed number of words. LenEmb (Kikuchi et al., 2016) introduces length embeddings to LSTM decoders. LC (Liu et al., 2018) controls length in abstractive summarization using a convolutional neural network via providing length information in first layer of the encoder. GOLC (Makino et al., 2019) formalizes LenEmb and LC via introducing a loss function that has an overlength penalty. Our research differs in that we don't focus on stopping decoding, rather we train two generative models with different length-generating abilities.

Other approaches utilize intermediate summaries to obtain length-aware summaries. LPAS (Saito et al., 2020) obtains a word sequence of desired length and generate summary with a non-length-controlled model. MLS (Sarkhel et al., 2020) generates a general summary which is then input to a length-controlled model. Our method differs from above methods via not requiring an intermediate stage to obtain length-controlled outputs, instead requiring 1 inference of a weight-merged model.

LAAM and PtLAAM (Liu et al., 2022) introduces an attention mechanism that bridges encoder and decoder of seq2seq architecture. It boosts attention scores of tokens selectively based on length-aware attention mechanism. It differs from our method in that our method is applicable to decoder-only models such as GPT (Brown et al., 2020) and Llama (Grattafiori et al., 2024), which LAAM and PtLAAM aren't due to reliance on seq2seq architecture. They also introduce LBD (Length-Balanced Dataset), a heuristic-focused which is in line with our focus on datasets. However, our methods minimize reliance on heuristics, instead opting for LLM-based re-summarization pipeline.

3 Approach

We build upon prior work by first bucketing inputs based on their target summary lengths. However, our approach differs in the data curation pipeline,

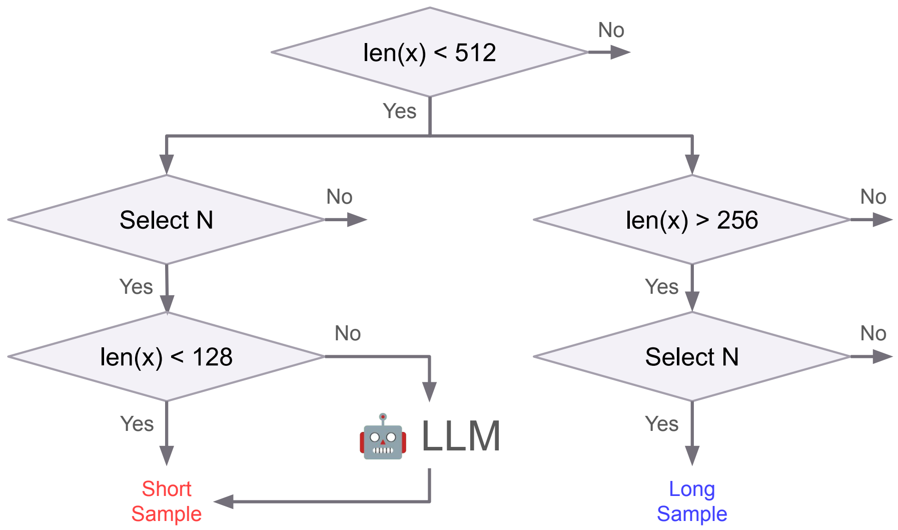


Figure 2: Data selection and LLM data augmentation pipeline. Length of the summary (len(x)) is the used criteria, and N samples, each for short and long, are left in the end. For train samples belonging to short division, N=2000 per each of XSUM and CNN/DM.

the use of a PEFT training scheme, and a focus on model merging—an aspect not explored in previous studies on summarization length.

3.1 Data Curation Pipeline

For a given dataset of inputs and target summaries, we curate two divisions with distinctive mean target summary lengths (shorter and longer) by following the pipeline introduced in Figure 2.

The pipeline first ignores all samples with length of summary larger than 512 as a filter for ensuring good quality. The pipeline then produces short samples via shortening the summary with an LLM (we use GPT-4o (OpenAI et al., 2024)). The prompt we use, parameterized by {length} and {text}, is the following:

```
Promptly summarize the following text to
under {length} characters.
Text: {text}
Print out the reduced summary only, which
should be under {length} characters.
```

As shown in Figure 3, the pipeline produces two divisions: (1) summary lengths in $[0, 128]$ (Short), and (2) summary lengths in $[256, 512]$ (Long). Those divisions are the ones we mainly experiment with. We further modify the pipeline to obtain 2 more divisions (also visible in Figure 3), for lengths $[0, 64]$ (Shortest, distribution, experimented with in Section 6.2) and $[256, 320]$ (Long (Low Var.), distribution experimented with Section 6.3). For the Shortest division we merely adjust the length criteria and follow same pipeline as Short. For the Long (Low Var.) on the other hand, we run GPT-4o on Long division by setting the max length to 320 and using the same prompt as Short. We repeat the runs up to 4 times in case generated outputs are shorter than 256.

3.2 PEFT

PEFT (Parameter-Efficient Fine-Tuning) (Mangrulkar et al., 2022) is a set of techniques designed for fine-tuning large models efficiently for specific tasks. Unlike traditional fine-tuning, PEFT optimizes a smaller subset of weights, and reduces both computational and storage requirements. In this work, we use the LoRA (Hu et al., 2022) technique, and fine-tune our model on the curated dataset divisions. Given model layer weights $W \in \mathbb{R}^{n \times m}$, LoRA approximates the weight updates $\Delta W \in \mathbb{R}^{n \times m}$ that would be done in full fine-tuning by decomposing them into the product of two low-rank matrices:

$$\Delta W = AB, \quad A \in \mathbb{R}^{n \times r}, \quad B \in \mathbb{R}^{r \times m}. \qquad (1)$$

Since the input/output dimensions (n, m) of layer weights are typically large in LLMs, choosing a small rank r (e.g., $r = 8$) significantly reduces the number of trainable parameters while still providing a reasonable approximation of full fine-tuning.

3.3 Merging PEFT Weights

Model merging combines multiple models into a unified model, aiming to preserve or enhance their

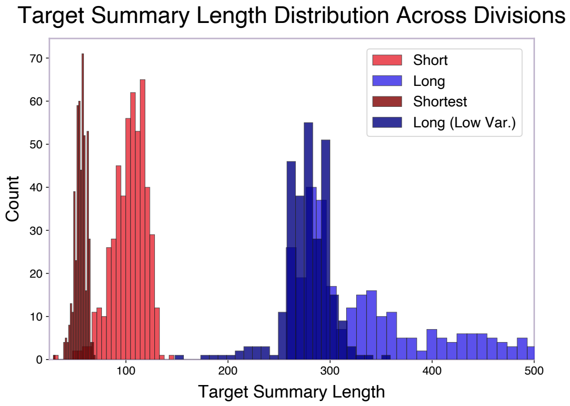


Figure 3: Summary lengths of all curated test divisions.

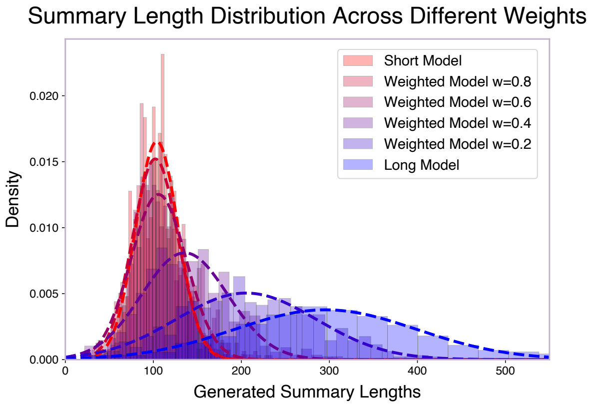


Figure 4: Generated summary lengths per model (from Short to Long, with interpolation weights of w).

performance. We perform weighted merging on the low-rank matrices from Short model (A_s, B_s) and Long model (A_l, B_l) by linearly combining the approximated weight update matrices ΔW_m using the concatenation-based merging method (Face, 2025a,b). Given coefficients w_s (for Short model) and w_l (for Long model), we formulate:

$$A_m = \begin{bmatrix} w_s A_s & w_l A_l \end{bmatrix}, \quad B_m = \begin{bmatrix} B_s \\ B_l \end{bmatrix}, \tag{2}$$
$$\Delta W_m = A_m B_m = w_s \Delta W_s + w_l \Delta W_l.$$

In our experiments, we set $w_s \in [0, 1]$ and $w_l = 1 - w_s$, making generation length a function of w_s. Thus, $w_s = 1$ matches the Short model behavior, and $w_s = 0$ matches the Long model.

4 Experiments

4.1 Data

We utilize the CNN/DM (Hermann et al., 2015) and XSUM (Narayan et al., 2018) datasets. CN-N/DM contains pairs of an article and a highlight, with 286,817 in training, 13,368 in validation, and 11,487 in test sets. XSUM contains pairs of an article and a single sentence summary, with 204,045 in training, 11,332 in validation, and 11,334 in test splits. We combine these datasets, and apply our data selection / augmentation pipeline (Figure 2), yielding two divisions: one with shorter and another with longer target summaries. Sample counts are (4000 / 500 / 500) and (2455 / 273 / 276) across train / validation / test splits, respectively (80% / 10% / 10%). The length distribution of target summaries in the test split per division is shown in Figure 3. Note that the distributions of train, validation and test splits are similar per division.

4.2 Training Details

We utilize PEFT (Mangrulkar et al., 2022) and train two LoRA adapters (models) per division on top of the "meta-llama/Llama-3.2-1B-Instruct" (Grattafiori et al., 2024) model with a single A100 GPU for our main experiments. We load the initial checkpoint in bf16, and for the PEFT, we set rank (of low-rank matrices) to 8, α (scaling factor of the low-rank matrices) to 32 and dropout to 0. We train the adapters for a single epoch with the effective batch size of 8, learning rate of $5e^{-5}$, warm-up ratio of 0.05, weight decay of 0.01, and the mixed-precision training. We train on the train splits, set the max token length to 2048 for the model inputs and generations, and save adapters every 50 steps.

4.3 Evaluation Method

Throughout the experiments, the generated summaries for each model are evaluated from both length and quality perspectives. We conduct the main evaluations by combining test sets of both divisions, and use the adapters saved at training step 300. For length, we analyze the mean, variance, and distribution. For quality, we report ROUGE scores (Lin, 2004) (ROUGE-1, ROUGE-2, and ROUGE-L). To enhance the reliability and smoothness of these measurements, we mitigate rare generation anomalies (e.g. cases where the model generation length blows up) by removing statistical outliers. Specifically, we filter out summaries with a length z-score greater than 4, as some examples shared in Appendix B.3. This filtering prevents extreme but rare cases from disproportionately affecting variance and mean estimates, while removing less than 1% of the samples. Addition-

Eval Set	ROUGE-1			ROUGE-2			ROUGE-L		
	Short Model	Long Model	Base Model	Short Model	Long Model	Base Model	Short Model	Long Model	Base Model
Our (Short + Long)	26.45	26.96	20.17	7.52	8.59	6.20	20.30	18.70	13.34
Our (Corresponding)	27.63	39.57	–	7.74	15.74	–	22.04	25.65	–
Gold (XSUM + CNN/DM)	25.54	33.33	26.04	7.50	12.10	8.71	18.68	22.09	16.46
Gold (XSUM)	26.10	21.96	16.59	7.11	4.65	3.89	19.95	15.09	11.47
Gold (CNN/DM)	25.20	39.88	31.28	7.74	16.41	11.37	17.97	26.11	19.19

Table 1: ROUGE Scores for Short model and Long model (Scaled to Percentage). Test inputs are from our data curation pipeline, with target sets of summaries varying by indicator: *Our* from our pipeline and *Gold* from original datasets. Each highlighted cell represents the best score per metric among models. We observe that our fine-tuned models outperform the Base model, indicating an effective summarization fine-tuning.

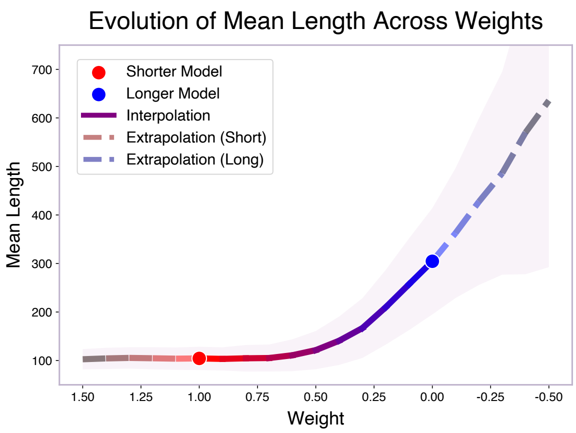


Figure 5: Summary length interpolation. A model with weight w blends Short (w) and Long ($1-w$) models. Faded purple area: generation variance (std). Weights outside $[0, 1]$ indicate extrapolation.

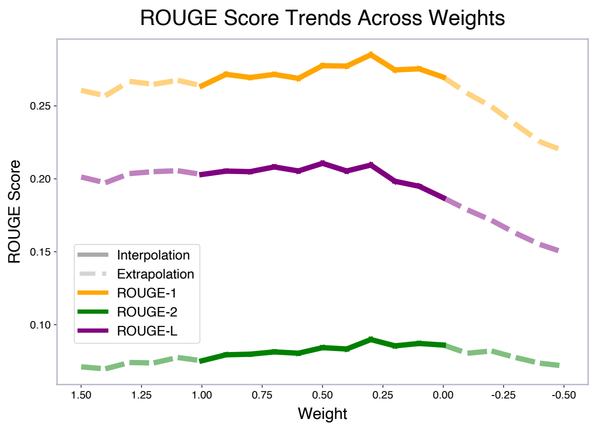


Figure 6: ROUGE score ablation across weighted models on the main evaluation set: *Our (short + long)* setting.

ally, we exclude test set inputs with more than $2048 - (margin = 256)$ tokens, where the margin is a heuristic to avoid truncation mid-output, ensuring alignment with the training setup. This step removes approximately 3% of the samples from the evaluation.

5 Results

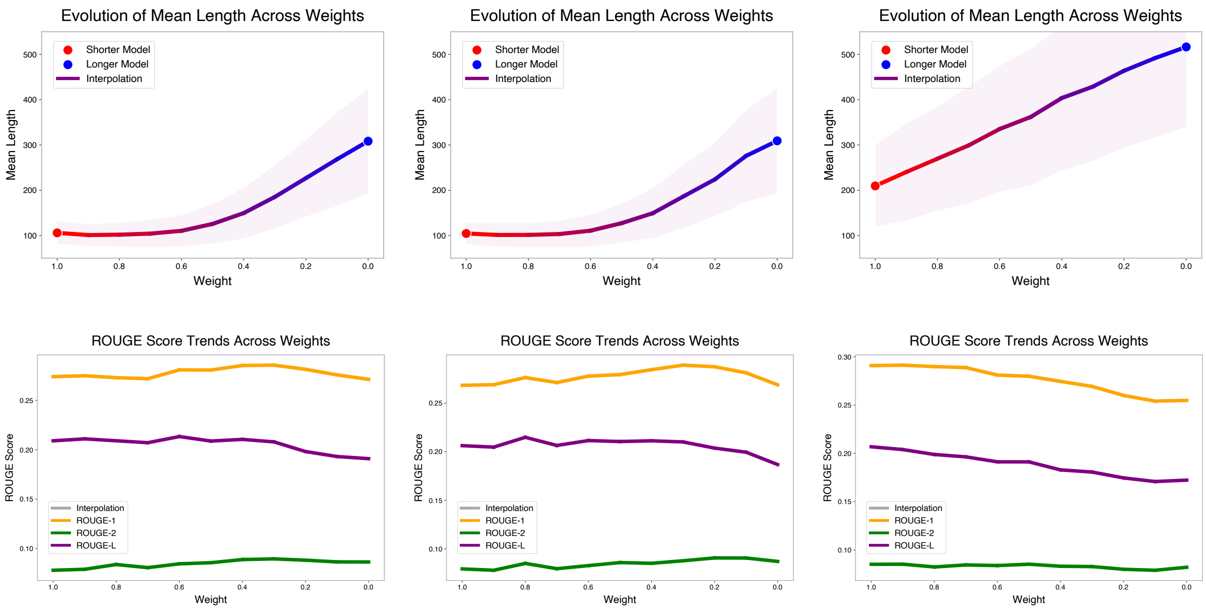
5.1 PEFT on Shorter and Longer Divisions

We evaluate whether our framework can adapt checkpoints to generate quality summaries of distinct lengths. As shown in Figure 4, our method yields two adapters: Short model and Long model, with clearly different mean generation lengths.

To assess quality, Table 1 reports ROUGE scores on multiple evaluation sets. The *Our* setting uses targets from our data augmentation pipeline, and we evaluate on both the combined test set (main) and length-specific subsets (short targets for Short model, long targets for Long model). The *Gold* setting on the other hand uses targets from the original datasets, and we evaluate on either the combined set or individual datasets. Across all evaluation sets, the highest score per metric (highlighted each row & block) is always noticeably achieved by either Short model or Long model, demonstrating that fine-tuning improves summarization quality compared to Base model.

5.2 Merging Short and Long Models

We explore a linear merge of Short and Long adapters by varying the weight w_s from 1 to 0 in steps of 0.1. The produced summary length distributions for the weights is shown in Figure 4. For $w \geq 0.4$, the generated summary lengths are centered around a similar mean value, whereas for $w \leq 0.4$, we observe a gradual shift in the distribution from shorter to longer summaries. Additionally, smaller values of w_s correspond to greater variance in the generated summary lengths. Notably, the length distribution per model approximately follows a normal distribution.



(a) Linear (b) DARE (Linear) (c) TIES

Figure 7: Merging method ablations.

Furthermore, Figure 5 presents an analysis of the relative summary lengths across models. For each model, we plot the mean and standard deviation of the generated summaries. The results align with Figure 4, where models with larger w_s values behave similarly to the Short adapter, while smaller w values gradually shift toward the behavior of the Long adapter (where $w_s \in [0, 1]$). The observed trend indicates a curved change, with the higher rate of change when the weight is lower.

Finally, Figure 6 illustrates the ROUGE score trends on the combined test set using targets generated by our pipeline. For $w_s \in [0, 1]$, ROUGE scores remain relatively consistent with minor variations, indicating that model merging does not compromise quality from the perspective of ROUGE scores.

5.3 Extrapolating Short and Long Models

We further investigate extrapolation beyond the nominal model merging region by considering values of $w_s \notin [0, 1]$, aiming to extend the generation-length trend in each direction. As in the interpolation setting, we maintain the constraint $w_l = 1 - w_s$.

In Figures 5 and 6, length extension is evident for Long model, whereas the generation lengths remain largely unchanged when Short model is assigned a dominant weight. ROUGE scores exhibit a gradual decline toward the longer end, which can be attributed to the reduced lexical overlap with the reference summaries as length increases.

6 Analysis

6.1 Impact of Merging Methods

We ablate merging methods to explore and evaluate potential alternatives (Face, 2025b). Specifically, we experiment with Linear (Ilharco et al., 2022), TIES (Yadav et al., 2023), and DARE (Yu et al., 2024) methods. Linear is a resource-efficient variant of the concatenation-based method, merging the low-rank matrices from two models (e.g. A_s and A_l) directly, approximating the merging of the weight-update matrix. Specifically, we apply the following formulation (Face, 2025b):

$$\begin{aligned} A_m &= \sqrt{w_s} A_s + \sqrt{w_l} A_l, \\ B_m &= \sqrt{w_s} B_s + \sqrt{w_l} B_l, \\ \Delta W_m &= A_m B_m \approx w_s \Delta W_s + w_l \Delta W_l. \end{aligned} \tag{3}$$

TIES (TrIm, Elect, and Merge) leverages parameter signs for effective model merging. DARE can be applied as a plug-in to both Linear and TIES methods (we report results with Linear), randomly dropping parameters and rescaling the remaining ones. For both DARE and TIES, we set the density value to 1.0, and apply the default setup in (Face, 2025b).

As shown in Figure 7, the length trajectories for the Linear and DARE methods largely mirror those of our concatenation-based approach. In contrast, TIES exhibits a distinct pattern where the

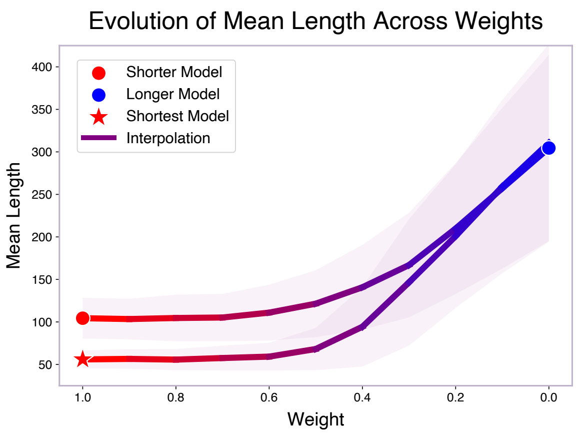


Figure 8: Length evolution from Short / Shortest to Long division fine-tuned models: length gap does not correspond to weight insensitivity for small w.

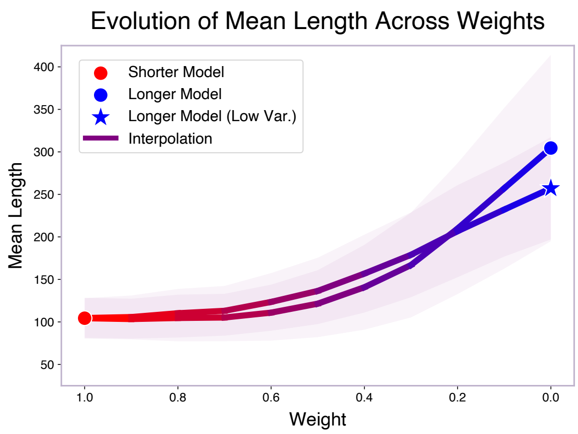


Figure 9: Length evolution from Short to Long / Long (Low-Var.) division fine-tuned models: lowering target variance impacts generation variance too

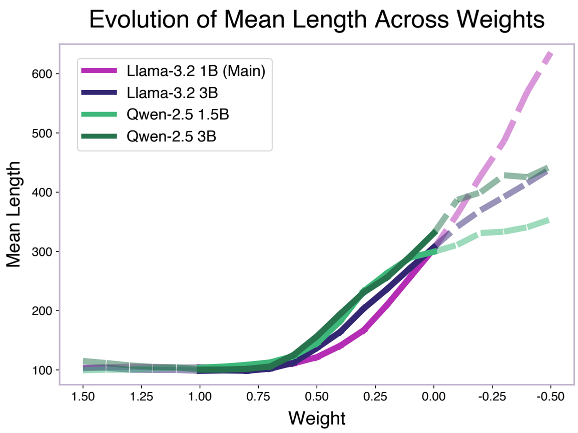


Figure 10: Mean generation length across weights for different base models, showing smooth evolution and generalizability of the proposed method across different models.

generation lengths scale in a linear correlation with weight but are consistently offset upward. Across all methods, ROUGE score trends remain comparable, indicating stable summarization quality over varying weights.

6.2 Length Gap and Sensitivity

Length ablation results in Figure 5 and Figure 7 indicate that summary length is largely insensitive to weights in the range $w_s \in [0.6, 1]$. We ask whether this could be due to the length gap between the Short and Long divisions and further augment the Short division by generating even shorter summaries, referred to as Shortest. This new division is generated following the process described in Section 3.1, with its length distribution shown in Figure 3.

As shown in Figure 8, the results suggest that the length gap may not explain the insensitivity of length change for smaller w_s values, implying that this behavior could be an inherent characteristic of the model or our merging pipeline. However, the model retains its ability to interpolate between different lengths, indicating that our method remains applicable across a broader range of available lengths.

6.3 Impact of Variance

We observe that variance increases as the length of generated summaries grows. To mitigate this effect, we construct a lower-variance dataset. In particular, we introduce another Long division, referred to as Long Low-Variance. This division is created by re-summarizing inputs into very long summaries, following the procedure described in Section 3.1. The resulting distribution of target summary lengths is shown in Figure 3.

Our results indicate that the variance of the merged models decreases significantly when using the Long Low-Variance division (Figure 9), suggesting that training on low-variance data may help control the variance in summary lengths, thereby improving the practicality of our method.

6.4 Ablation on Different Base Models

We ask whether our method generalizes across different Base models by re-running the analysis from Figure 5 on a larger model: Llama-3.2 3B; as well as models from a different family: Qwen-2.5 1.5B and Qwen-2.5 3B (Qwen et al., 2025).

The results in Figure 10 show that our method yields smooth generation length evolution across

Model	Len.	Generated Summary
w = 1.0	72	Johanna Johnson's 'Sirens' Call collection is a glamorous, opulent show.
w = 0.8	101	The Sirens' Call collection by Johanna Johnson features hand-beaded, mirrored, and feathered designs.
w = 0.6	104	Johanna Johnson's 'Sirens' Call' collection features hand-beaded dresses, fringe, feathers, and leather.
w = 0.4	138	Johanna Johnson's 'Sirens' Call' collection was a glamorous Hollywood-inspired show. The designer created the collection in under a month.
w = 0.2	244	Johanna Johnson's Sirens' Call collection featured glamorous Hollywood-inspired gowns. The designer created the collection in under a month. 2015 was a spectacular finale show by Johanna Johnson. The collection was inspired by the Great Gatsby.
w = 0.0	339	Johanna Johnson's 'Sirens' Call' collection featured mirrored embellishments, fringing, and feathers. The designer said she intended to 'bring a little bit of Hollywood to Sydney'. Johnson said she wanted to celebrate women who are 'independent strong women, everyone who is trying to juggle careers, motherhood and even just modern life'.
Gold target	344	Final show for MBFWA took place at 7pm at the Carriageworks venue in Sydney. Johanna Johnson's show, called Sirens' Call, was inspired by Hollywood glamour. The designer said the collection was 'a call to all independent strong women'Stand-out pieces from the show included a hand-beaded mirror gown, intricate wedding gowns and gold fringing.

Table 2: An example of a typical generation

all models, albeit with distinct curve characteristics: in Qwen models, generation length tends to saturate at lower weights, whereas in Llama models it increases steadily, with varying rates per model. These findings indicate that our method is broadly compatible across LLM architectures, requiring only minor and idiosyncratic adjustments.

Moreover, we further show in A.2 that empirically fitting a continuous and invertible function to these curves enables direct control over generation length for arbitrary targets (Figure 12), demonstrating the framework's generalizability beyond analysis to practical and user-specified length control.

6.5 Qualitative Analysis

We present a representative example of generated summaries in Table 2. Although strict standards for length-specific summaries are difficult to define, the outputs across lengths are generally fluent, grammatically accurate, and factually correct (see additional samples in Appendix B).

We notice in Table 2 that higher weights (closer to the Short model) yield concise, evaluative summaries (e.g. "glamorous, opulent"); mid-range weights add concrete attributes (e.g., "fringe/feathers," "under a month"); and lower weights (toward the Long model) introduce broader context and quotes (e.g. Hollywood theme). Thus, controlling length also shifts style and content selection. While longer outputs increase thematic coverage, they sometimes omit key slots (e.g. venue/time) or add extraneous details.

Another interesting phenomenon we observe is that shorter and longer summaries sometimes focus on complementary but different aspects of the original text. As illustrated in Appendix B.2, shorter summaries for *Prompt - 1* focus more on the funeral, while the longer ones focus more on the background information of the incident.

7 Conclusion

We present a novel method for generating summaries of varying lengths through dynamic interpolation of PEFT weights. Our pipeline effectively captures the length property of generation and yields smooth interpolation results. Importantly, the quality of the summaries remain stable, as shown by ROUGE scores and qualitative assessment. Extrapolation evaluations yield longer summaries but struggle with shorter ones. We further conduct ablations demonstrating that the method generalizes. Overall, our results establish DLC (Dynamic Length-Control) as a broadly applicable paradigm for length-controlled summarization.

Limitations

Our approach requires training two models with distinct target lengths, which can be more resource-intensive than single-model methods. While it supports arbitrary length control at inference without additional length-specific fine-tuning, extrapolation does not yield shorter summaries beyond the range of the trained shorter model.

Our quality assessment relies solely on ROUGE scores, which may not fully capture aspects such as coherence, fluency, or factual accuracy. Relatedly, by focusing solely on length control, we do not assess how other summary properties may shift as a side effect. While qualitative inspection suggests outputs remain reasonable across weights, we lack a systematic analysis of these emergent factors, which could affect the method's practical applicability. A broader examination of such factors, and their controllability, is left to future work. Finally, as our data sources are exclusively in English, assessing the framework's applicability in multilingual settings also remains for future study.

References

Anonymous. 2024. Controlling summarization length through eos token weighting.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Hugging Face. 2025a. Model merging. Accessed: 2025-03-13.

Hugging Face. 2025b. Peft welcomes new merging methods. Accessed: 2025-03-13.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

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.

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

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.

Yizhu Liu, Zhiyi Luo, and Kenny Zhu. 2018. Controlling length in abstractive summarization using a convolutional neural network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4110–4119, Brussels, Belgium. Association for Computational Linguistics.

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.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Lesly Miculicich, Yujia Xie, Song Wang, and Pengcheng He. 2023. Summarization with precise length control. *Preprint*, arXiv:2305.05171.

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. *arXiv preprint arXiv:1808.08745*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Itsumi Saito, Kyosuke Nishida, Kosuke Nishida, Atsushi Otsuka, Hisako Asano, Junji Tomita, Hiroyuki Shindo, and Yuji Matsumoto. 2020. Length-controllable abstractive summarization by guiding with summary prototype. *Preprint*, arXiv:2001.07331.

Ritesh Sarkhel, Moniba Keymanesh, Arnab Nandi, and Srinivasan Parthasarathy. 2020. Interpretable multi-headed attention for abstractive summarization at controllable lengths. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6871–6882, Barcelona, Spain (Online). International Committee on 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.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Preprint*, arXiv:2306.01708.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *Preprint*, arXiv:2311.03099.

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.

A Extended Analysis

A.1 Merging a Single Model to Base Model

In this experiment, we examine the case where a single fine-tuned model is merged with the Base model using the procedure introduced in Equation 2. Specifically, we set the weight update matrix of the Long model to zero, $\Delta W_l = 0$, effectively reducing it to the Base model. Given that the Base model exhibits a higher mean generation length than both the Short and Long models, we select the Short model for this analysis to provide a more expressive control range. Under this formulation, the merging operation is equivalent to scaling the weight update matrix of the Short model by w_s, enabling interpolation from no update to an amplified (positive) or deamplified (negative) update.

As shown in Figure 11 (Left), scaling a single model's weight update matrix yields a smooth and well-behaved length control curve, closely resembling those observed in our original two-model merging experiments. However, as illustrated in Figure 11 (Right), this approach exhibits a notable drawback: ROUGE scores decline sharply even within the interpolation range when the Short model's contribution is deamplified. This drop may not necessarily indicate degraded summary quality, as it could be due to the reduced lexical alignment as the model shifts away from the fine-tuned distribution. Nonetheless, it remains undesirable as our model merging methodology preserves both length control and measured quality more effectively.

A.2 Operationalizing Model Merging for Arbitrary Length Control

In this analysis, we serve two primary purposes. First, we aim to motivate our model-merging methodology by demonstrating the ineffectiveness of naive length control via prompting. Second, we introduce and evaluate an actionable framework that translates our experimental findings into a practical tool for generating summaries of an arbitrary length.

A.2.1 Baseline: In-Context Length Control

The baseline method attempts to control output length by directly instructing the Base model. We minimally modify the prompts used throughout our experiments by inserting a phrase specifying the target length, as highlighted below:

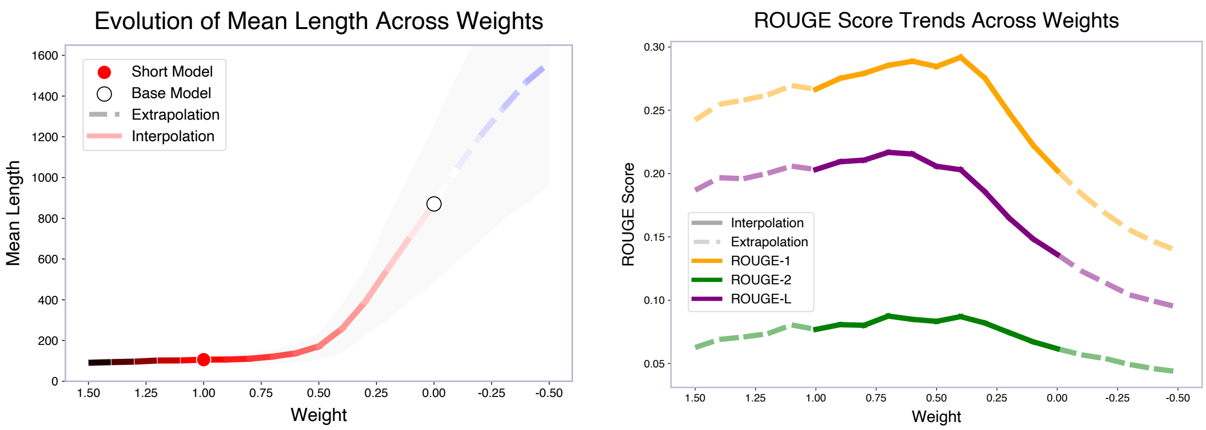


Figure 11: Using Short model and Base model as two models for merging, effectively examining impact of using a single model within our pipeline. While the length control is smooth, the rapid decline of ROUGE scores even within the interpolation range stands undesirable compared to merging two models.

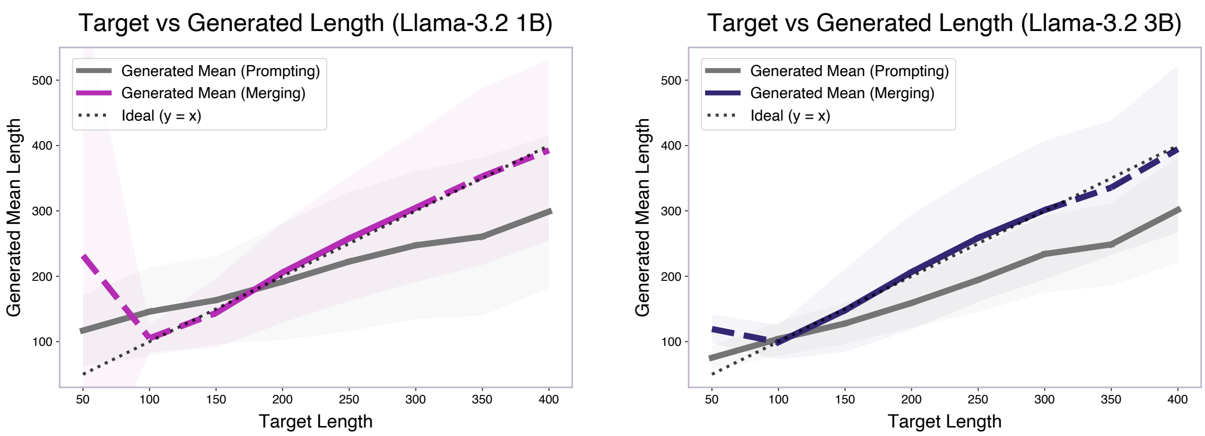


Figure 12: Evaluation of our proposed practical model-merging framework for arbitrary-length summarization against a basic prompting baseline. Our method operationalizes the smooth length control observed in earlier experiments into a practical setting, based on specified target length. Using Llama-3.2 1B (left) and 3B (right), our approach reliably produces summaries with mean lengths closely following the ideal $y = x$ line, whereas the prompting baseline often deviates substantially.

```
Summarize the following text in {length}
characters:
{text}
```

A.2.2 Proposed Practical Framework

Our proposed framework operationalizes the smooth length control observed in our experiments. The algorithm is as follows:

(1) Characterize the Length-Weight Relationship: First generate sufficient number of summaries across a set weights. In our case, we consider $w_s \in [-0.5, 1.5]$ by incorporating the extrapolation case as well. Then measure the mean generation length for each weight, providing empirical data that maps weights to output lengths.

(2) Fit a Continuous, Invertible Function: Based on the sigmoidal trend observed across various models (Figure 10), we fit the empirical data to a sigmoid function. This creates a continuous mapping from the weight, w_s, to the expected generation length, $L(w_s)$:

$$L(w_s) = L_{\min} + (L_{\max} - L_{\min})\sigma(s(w_s - w_0)) \tag{4}$$

where $\sigma(u) = \frac{1}{1+e^{-u}}$; and $L_{\min}$, $L_{\max}$, s (scale), and w_0 (center) are empirically fitted parameters. While this sigmoid function is a good fit for our experiments, different setups involving different models, architectures or curated datasets might require other functional forms.

(3) Deduce Weight from Target Length: As the fitted function $L(w_s)$ is invertible, we use its inverse, $w_s = L^{-1}(\text{target\_len})$, to calculate the precise weight needed to generate a summary of a desired length on expectation.

A.2.3 Evaluation and Results

We evaluate both the prompting baseline and our framework using Llama-3.2 1B and Llama-3.2 3B models, with target lengths from 50 to 400 characters increasing by 50 each step. As shown in Figure 12, our framework produces summaries whose mean length closely aligning with the ideal $y = x$ line. The prompting baseline, however, shows significant deviation, confirming it is an unreliable method.

Yet our method struggles at the shortest target length of 50 characters. This limitation arises because our main experiments fail to show that extrapolation toward the shorter end ($w_s > 1$) consistently reduces generation length, causing the inverse function's suggested weight to fail. A potential solution is to curate a new, even shorter dataset division (as in Figure 3) with our data augmentation pipeline.

B Model Summaries and Comparisons

B.1 Typically Generated Summaries

In this section, we share two examples of the typically generated summaries per model across weights from Short to Long model. The examples showcase the readability, fluency, conciseness and coverage of the typical generations.

Prompt - 1:

Summarize the following text:
A deskbound worker has used a GoPro camera to capture a unique view of his boring office job, turning even the most mundane day-to-day tasks into a high octane activities. Daniel Williams, from Toronto, filmed himself doing everything from sealing envelopes and sending emails to making cups and coffee and going to the bathroom, before editing all of the clips together to offer an action-packed perspective of his regular daily life. The video, which is titled 'Boring Office Job GoPro Commercial', has clearly struck a chord with people working in similar jobs around the globe, with the video already attracting an impressive three million views since being uploaded to YouTube last week. Office space: Daniel Williams, an office worker from Toronto, captured his entire day with a GoPro camera. Bright idea: Mr Williams came up with the idea after seeing other breathtaking GoPro videos online. The 90-second video makes a change from the usual, dramatic content normally caught by such cameras which, thanks to their durability, are often used to film some of the most breathtaking moments in life from base jumps to shark encounters. This time, however, viewers are treated to Mr WIlliams removing staples, photocopying, stuffing envelopes, drinking coffee, and even going to the bathroom. All set to a techno beat, the pacey clip turns a regular 9-to-5 office job into something far more exhilarating. Button pusher: In the 90-second video, Mr Williams filmed himself stapling, photocopying (pictured) and stuffing envelopes. Downtime: There are also shots of Mr Williams drinking coffee and going to the toilet - all set to a techno beat soundtrack. Stuffed: 'I wore my GoPro to work one day and decided to edit the footage to the style of a GoPro Commercial,' Mr Williams wrote. Alongside the video, the Toronto filmmaker wrote in the description: 'I wore my GoPro to work one day and decided to edit the footage to the style of a GoPro commercial.' He said that he decided to take on office temp work in order to pay the bills between documentary projects. And it was whilst he was working for a real estate company for six months that he came up with the idea for the unique video. 'The job was a lot of very mundane repetitive tasks. For a creative guy like me it kind of killed me,' Mr Williams told the New York Daily News.. Doodling: In between films and documentaries Mr Williams takes on temporary office jobs. It was during one such clerical job that he came up with the idea. Ready for launch! Mr Williams says he thought it would be a funny idea to record everything he does during the day including playing with his pen. '[People always use GoPro cameras to film] these action videos, but the reality is the people who buy these cameras have everyday jobs,' he added. The video has thus far earned a huge amount of praise from viewers on YouTube, many of whom have been quick to comment about how well Mr Williams has managed to capture the mundane repetition involved in a typical office job. 'So good. So accurate,' said YouTube user Ryan Bobbe. 'Sometimes I wish I had a job like this instead of dealing with a mass amount of idiotic people a day' wrote another.

Generated Summaries:

Model	Length	Generated Summary
w = 1.0 (Short)	107	Daniel Williams, a Toronto office worker, used a GoPro to film his mundane tasks and created a viral video.
w = 0.8	111	Daniel Williams filmed his mundane office tasks with a GoPro camera, capturing 90 seconds of his daily routine.
w = 0.6	116	Daniel Williams filmed his daily routine with a GoPro camera, turning mundane office tasks into a high-energy video.
w = 0.4	193	Daniel Williams filmed his mundane office job with a GoPro camera. He edited all clips together to create an action-packed video. It's been viewed 3 million times since uploading it to YouTube.
w = 0.2	219	Daniel Williams, a Toronto office worker, filmed his day-to-day tasks with a GoPro camera. The video, titled 'Boring Office Job GoPro Commercial', has attracted an impressive 3 million views on YouTube since its upload.
w = 0.0 (Long)	250	Daniel Williams from Toronto filmed himself doing everything from sealing envelopes to making coffee and going to the bathroom in his office. He captured his 90-second video with a GoPro camera, which has already attracted 3 million views on YouTube.
Gold Target	239	Daniel Williams' video shows a every aspect of a day in the life of an office worker. The 90 second clip has been viewed three million times on YouTube. GoPros are often used by adrenaline junkies to capture once-in-a-lifetime experiences.

Prompt - 2:

Summarize the following text:

A sprawling rural estate - complete with a Grade II-listed manor house and five cottages - is on the market for £6million. Set in 200 acres of land near Bridgnorth, Shropshire, the Chyknell Hall Estate also boasts a swimming pool, tennis court, stables - and even its own cricket green. Its centrepiece, the 11-bedroom Chyknell Hall, was built in 1814 and has only changed hands twice since. It is thought the Regency home could now attract A-list buyers as the secluded grounds offer residents complete privacy. Sprawling: Chyknell Hall Estate, which comes complete with a Grade II-listed house (pictured) and five cottages, is on the market for £6million. Grand: A sumptuous sitting room in the nine-bedroom Chyknell Hall, which was built in 1814 and has only changed hands twice since. Outdoor living: The property also boasts a large outdoor swimming pool and pool house (pictured), library, wine cellar and billiard room. Sports day: The current owner installed a cricket green, pictured, on the estate, and bought all the maintanence equipment needed. The owner will also be set up for entertaining with a wine cellar, four-oven Aga, loggia, library and billiard room. And a number of compartment gardens means guests will never be bored by the same view. The estate, described as the 'pinnicle of the residential tree', offers 57 acres of woodland and more than 500 acres of adjoining sporting rights, including game shooting. The land also planning permission for five more properties, should the new owner wish to expand. The ownership of Chyknell Hall Estate can be traced back to medieval times and this is only the third time in its history it has been on the market. It remained in the same family for generations before it was sold to a family in the 1930s. The current owners bought it four years ago. The estate is now being sold by estate agents Knight Frank for offers over £6million. Clive Hopkins, head of estates at Knight Frank, said: 'This estate is a package that ticks most boxes. 'You've got a listed Regency house that has been refurbished to a high standard but kept traditional features, set in some wonderful compartmentalised gardens which means as you walk round the house you get a different type of garden depending on what window you look out. Bright and airy: The Regency home has been refurbished to a high standard, but has kept original features. Above, one of the dining areas. Grand hallway: The ownership of Chyknell Hall Estate can be traced to medieval times and this is only the third time it has been on the market. Opulent: A roll-top bathtub takes centre stage in one of the 10 bathrooms in Chyknell Hall, which also boasts 11 bedrooms. Secluded: The sprawling manor, pictured, is set in traditional parkland which has no footpaths through it - offering complete privacy. 'That's then set in traditional parkland which is very private and secluded. There's no rights of way or footpaths through it, which is very rare for a house of that age. 'Rights of way originated from footpaths being made for staff working in a big country house to get back to their village so it's very rare for a house of this stature not to have a path. 'It's a really big draw in this day and age, owners want privacy but don't want to be too remote. 'This estate has no road noise, no light pollution and no pylons. It's a package which is very rarely found and is highly sought after. 'It would suit someone looking to buy the next stage of house, an estate like Chyknell is the pinnacle of the residential tree. 'The current owners are only moving because they have found an estate with more land. He is very keen on his cricket and put a cricket pitch in here so he could get his friends round for a game. It's not quite a Lord's, but it's a very high standard, he bought all the equipment, rollers and things. 'With this estate you get a lot of bricks and mortar for your money. Not many estates like this come up and when they do they are highly sought after.' Rural retreat: There are a number of smaller properties on the grounds, including three cottages housed in the red-brick block pictured above. Picturesque: The tennis court on the Chyknell House Estate, which offers residents no road noise, no light pollution and no pylons. Beautiful: There are compartment gardens on the estate, which means residents and guests will never be bored of the view outside. Elegant: The land also planning permission for five more properties, should the new owner wish to expand. Above, the estate's swimming pool.

Generated Summaries

Model	Length	Generated Summary
w = 1.0 (Short)	86	Chyknell Hall Estate, a Grade II-listed Regency home, is on the market for £6 million.
w = 0.8	153	Chyknell Hall Estate, a Grade II-listed manor house, is on the market for £6 million. The estate boasts a swimming pool, tennis court, and cricket green.
w = 0.6	160	Chyknell Hall Estate, a Grade II-listed manor house, is for sale for £6 million. It features a 10-bathroom Regency-style house, tennis court, and swimming pool.
w = 0.4	128	Chyknell Hall Estate, in Shropshire, is on the market for £6 million. It boasts a Grade II-listed manor house and five cottages.
w = 0.2	249	Chyknell Hall Estate, near Bridgnorth, Shropshire, is on the market for £6million. It boasts a Grade II-listed manor house and five cottages. The estate offers complete privacy and a secluded location, with no road noise, light pollution, or pylons.
w = 0.0 (Long)	236	Chyknell Hall Estate, which comes complete with a Grade II-listed house, five cottages, a swimming pool, tennis court and cricket green, is on the market for £6million. The estate is set in 200 acres of land near Bridgnorth, Shropshire.
Gold Target	343	Set on 200 acres near Bridgnorth, Shropshire, the Chyknell Hall Estate also boasts a tennis court and wine cellar. The Grade II-listed Regency home at the centre of the property offers 11 bedrooms, a library and a billiard room. It is thought it could attract A-list buyers as the secluded grounds and gardens offer residents complete privacy.

B.2 Summaries with Different Focus Across Weights

In this section, we share two example summaries where larger and smaller weighted models focus on different and mostly complementary aspects of the input text.

Prompt - 1:

Summarize the following text:

The service was held at the 22-year-old's school in Ohio, before his burial at a nearby cemetery.His death has heightened tensions between the US and North Korea.Pyongyang has called President Donald Trump a "psychopath", while he describes N Korea as a brutal regime.Pyongyang's official Rodong Sinmun newspaper claimed Mr Trump might launch a pre-emptive strike on North Korea to divert attention from his domestic difficulties, and warned South Korea "that following psychopath Trump... will only lead to disaster".Earlier in the week, after announcing Mr Warmbier's death, President Trump said he was determined to "prevent such tragedies from befalling innocent people at the hands of regimes that do not respect the rule of law or basic human decency".The thousands of funeral goers formed two lines as Mr Warmbier's coffin was carried into Wyoming High School for the 45-minute service, and then along the three-mile route to the cemetery.There was only space for 800 people inside the school's auditorium so most attendees watched the service on monitors in overflow rooms, CNN reports.Mr Warmbier's brother and sister, along with some friends, spoke during the service. Bagpipes played as mourners, wearing blue and white ribbons - the school's colours - to honour the former student, left the school.Thousands of people from his small hometown of Wyoming, Ohio, packed into Otto Warmbier's old high school for his funeral.For most, his casket was the first they had seen of Otto since he went to North Korea - awful confirmation of the shocking, surreal set of circumstances that had played out since early last year.Among several politicians in attendance, including officials from the state and defence departments was the Ohio Senator Rob Portman, who had been involved in the quest to bring him home.Mr Portman reflected a sense that while today was a time to talk about Mr Warmbier's life, as a young man with huge potential, the issue of responsibility had to be addressed.After the funeral, mourners stood to watch the cortege pass and disappear through the streets with blue and white ribbons tied to all the trees in his honour.As they reflect, there are many not just here but across the country, asking how North Korea will now pay for what many see as the murder of an American citizen.Otto Warmbier was returned by North Korea to the US last week, more than a year after he had been sentenced to 15 years' hard labour.He was in a coma and died a few days later. His parents say that though he never spoke or opened his eyes, "the countenance of his face changed - he was at peace. He was home and we believe he could sense that".US doctors say he suffered a "severe neurological injury", the most likely cause of which was a cardiopulmonary arrest that had cut the blood supply to the brain.A post-mortem examination was not carried out at the request of the family.Otto Warmbier, a student at the University of Virginia, was travelling with a tour group when he was arrested at Pyongyang airport in January 2016.He was accused of stealing a propaganda sign from a hotel and, two months later, was sentenced during a short trial.His family had no news of him until 13 June, when Pyongyang said he had been in a coma since his trial after contracting botulism and receiving a sleeping pill.US doctors say there is no evidence he suffered from botulism, and the Warmbier family believe he died as a result of "awful torturous mistreatment".

Generated Summaries:

Model	Length	Generated Summary
w = 1.0 (Short)	96	Funeral for Otto Warmbier, a US citizen, has been held at Ohio's school, where he was a student.
w = 0.8	95	Otto Warmbier's funeral drew thousands in Ohio; North Korea warned he'll only lead to disaster.
w = 0.6	76	Funeral for Otto Warmbier at Ohio high school, amid US-North Korea tensions.
w = 0.4	251	The US has called North Korea a "brutal regime" after Otto Warmbier's death. The US President had said he would "prevent such tragedies from befalling innocent people at the hands of regimes that do not respect the rule of law or basic human decency".
w = 0.2	243	Otto Warmbier, a 20-year-old student, died in a coma at Pyongyang hospital last year after being arrested and sentenced in a North Korean court. He was on a tour of the country. His death has heightened tensions between the US and North Korea.
w = 0.0 (Long)	238	Otto Warmbier, a 19-year-old University of Virginia student, died in a coma after being arrested in North Korea in January 2016. He was sentenced to 15 years' hard labour in a short trial. Warmbier was in a coma and died a few days later.
Gold Target	173	More than 2,500 family, friends and well-wishers gathered for the funeral of Otto Warmbier, the US student who died after falling into a coma while in prison in North Korea.

Prompt - 2:

Summarize the following text:

Javier Hernandez would have been in no doubt how important his decisive goal for Real Madrid was in the Champions League quarter-final against city rivals Atletico Madrid after the entrance the squad made to the Santiago Bernabeu. French defender Raphael Varane uploaded an astonishing video to Twitter of the fire-hot arrival Madrid's players were given from their supporters, who mobbed the team coach and banged on the windows in support and anticipation. 'Live from the bus before the match! A huge thank you for your support,' Varane wrote with the video. VIDEO - Scroll down to see Real Madrid's welcoming from their fans. Real Madrid defender Raphael Varane captured the team's entrance to the Santiago Bernabeu on Wednesday. Fans greeted the team coach as Real arrived for their Champions League second leg with Atletico Madrid. Supporters lit flares and banged on the windows to encourage their team against Atletico. Real defender Varane filmed the footage of the supporters as the team arrived to play Atletico. Javier Hernandez scored the winning goal for Madrid in the 88th minute against Atletico to win the tie. The atmosphere inside the Santiago Bernabeu meant the iconic stadium resembled a furnace upon kick-off, but Atletico provided stubborn resistance to Real. It wasn't until Hernandez struck in the 88th minute, converting a pass from Cristiano Ronaldo after the Portuguese star drove through Atletico's defence, that Madrid sealed a 1-0 aggregate triumph. Carlo Ancelotti's side could face Spanish rivals Barcelona, Serie A's runaway leaders Juventus or Pep Guardiola's Bayern Munich in the semi-finals. The draw takes place on Friday.

Generated Summaries:

Model	Length	Generated Summary
w = 1.0 (Short)	94	Real Madrid won 1-0 at the Santiago Bernabeu, with Javier Hernandez scoring the decisive goal.
w = 0.8	134	Real Madrid fans welcomed their team to the Santiago Bernabeu before the Champions League quarter-final match against Atletico Madrid.
w = 0.6	89	Real Madrid secured a 1-0 win over Atletico Madrid in the Champions League quarter-final.
w = 0.4	222	Real Madrid's players were mobbed by fans as they arrived at the Santiago Bernabeu. Raphael Varane uploaded a video to Twitter of the team's entrance. Javier Hernandez scored the winning goal for Madrid in the 88th minute.
w = 0.2	248	Real Madrid's players were given a warm welcome by the Santiago Bernabeu's fans after the team arrived for the Champions League second leg with Atletico Madrid. The fans lit flares and banged on the windows to encourage their team against Atletico.
w = 0.0 (Long)	374	Real Madrid defender Raphael Varane captured the team's entrance to the Santiago Bernabeu on Wednesday. Supporters mobbed the team coach and banged on the windows in support and anticipation. Real defender Varane filmed the footage of the supporters as the team arrived to play Atletico. Javier Hernandez scored the winning goal for Madrid in the 88th minute to win the tie.
Gold Target	345	Real Madrid eliminated Atletico Madrid from the Champions League. Javier Hernandez secured victory with 88th minute strike for Real Madrid. Raphael Varane filmed fans as Madrid arrived at the Santiago Bernabeu. READ: Barcelona vs Real Madrid is the dream Champions League final. Who will win the Champions League? Our reporters have their say...

B.3 Outlier Generations

In this section, we share an example outlier (length-wise) generation per model with their target pairs. Our results indicate that the outliers are extremely longer versions of the mean summary length per model.

Generated Summaries:

Model	Length	Generated Summary
w = 1.0 (Short)	203	Alpha Delta's brand is now banned; members can't get it without consent; pledge agreed to get it for pledge activities; fraternity has faced allegations of hazing, sexual assault, and excessive drinking.
Gold Target	309	College judicial committee found Alpha Delta responsible for causing harm to pledges and violating terms of suspension for alcohol violations. The 46-year-old fraternity has until next Monday to appeal the decision. Alpha Delta attorney previously said small group of members voluntarily chose to get brands.
w = 0.7	234	Service of process is an industry of its own, but most defendants avoid it, especially on Facebook. The law of service evolved to balance notice and access to court. Alternatives include nail and mail, and online service via Facebook.
Gold Target	122	A court allowed a wife to serve divorce papers via Facebook. Danny Cevallos: Why not let people be found via social media?
w = 0.5	297	The Triduum of Holy Week, the three days of Easter, have profound meaning. Good Friday is the day of the crucifixion, while Holy Saturday occupies the space between the darkness of the crucifixion and the bright hope of Easter. Easter is a time of transformation, a shift from one form to another.
Gold Target	268	Jay Parini: When religious identity, ethics, tolerance are roiling the culture, it's worth looking at message of Holy Week and Easter. He says ritual enactment of these three days is reminder that again and again the human condition moves through darkness into light.
w = 0.4	608	Third round of the EFL Cup: Bournemouth 2-3 Preston North End, Brighton & Hove Albion 1-2 Reading, Derby County 0-3 Liverpool, Everton 0-2 Norwich City, Leeds United 1-0 Blackburn Rovers, Leicester City 2-4 Chelsea, Newcastle United 2-0 Wolverhampton Wanderers, Nottingham Forest 0-4 Arsenal, Scottish League Cup quarter-finals: Greenock Morton 2-1 Dundee United, Rangers 5-0 Queen of the South, Fulham 1-2 Bristol City, Queens Park Rangers 1-2 Sunderland, Southampton 2-0 Crystal Palace, Swansea City 1-2 Manchester City, Stoke City 1-2 Hull City, Tottenham Hotspur 5-0 Gillingham, Celtic 2-0 Alloa Athletic
Gold Target	147	All the match reports for the midweek EFL Cup action, where Manchester United got back to winning ways and set up a Manchester derby in round four.
w = 0.2	598	The world's seventh-largest economy is under assault. The health and wealth of the ocean are assessed in a WWF report released Thursday, Reviving the Ocean Economy. The report is the result of a hard economic analysis performed by The Boston Consulting Group built on a foundation of the latest ocean science provided by the Global Change Institute of the University of Queensland. The report is the result of a hard economic analysis performed by The Boston Consulting Group built on a foundation of the latest ocean science provided by the Global Change Institute of the University of Queensland.
Gold Target	133	Ocean economic powerhouse valued at $24 trillion: WWF report. Marco Lambertini: Ocean plays direct role in livelihoods across globe.
w = 0.0 (Long)	891	The doctor who ploughed into an elderly couple in a 50mph head-on crash while rushing to work has escaped prosecution after the police officer investigating the case went off sick. The female driver in a Toyota Aygo pulled in front of Colin Kay as he was driving his Citroen Picasso on the A586 in Great Eccleston in Lancashire last year. The collision was captured on film as Mr Kay had fitted a camera to his dashboard and revealed how he had no time to react and avoid the crash. The incident happened on the A586 in Great Eccleston in Lancashire last year and Mr Kay provided Lancashire Police with the footage. After the incident last year, the doctor driving the Toyota also had to be cut from her vehicle and was taken to the Royal Preston Hospital by air ambulance. The 72-year-old was left with bruises after the crash and his wife Krysia suffered a whiplash injury to her shoulder.
Gold Target	297	Colin Kay was driving on the A586 when another car ploughed into his car. The whole incident was captured on film by a dashcam on his dashboard. He has been told the driver will not face any action over crash last year. This is because the police officer investigating the case had gone off sick.