Extending AutoCompressors via Surprisal-Based Dynamic Segmentation

Srivishnu Ramamurthi[*] **Richard Xu**[*]
Raine Ma **Dawson Park** **David Guo**
Charles Duong[†] **Vasu Sharma**[†] **Sean O'Brien**[†] **Kevin Zhu**[†]
Algoverse AI Research
srivishnur@ucla.edu, richardxu257@gmail.com, charlie@algoverseairesearch.org

Abstract

The long-context bottleneck of transformer-based language models can be addressed via context compression frameworks such as AutoCompressors, which distill tokens into **soft prompts** but silently assume uniform information density. We revisit this assumption and introduce dynamic segmentation by partitioning the input whenever the cumulative token-level **surprisal** exceeds a threshold τ, yielding segments with balanced information before **summary vector** generation. We show that dynamically adjusting segment boundaries based on surprisal enables better alignment between the original and soft prompts for prediction and inference. Experimental results show that our surprisal-based segmentation outperforms a pretrained baseline model and the randomized segmentation AutoCompressor baseline with regard to cross-entropy loss and in-context learning (ICL) accuracy. Code will be released upon publication.

1 Introduction

Context window limitations hinder long-context fine-tuning and inference in transformer-based language models (Vaswani et al., 2017) due to memory and compute constraints (Wang et al., 2024). To mitigate this, compression methods that reduce input complexity have been introduced, falling into two broad categories: either hard prompt or soft prompt techniques (Li et al., 2024).

Hard prompts are discrete natural language sequences consisting of tokens from a language model's vocabulary (Sennrich et al., 2016). While hard prompts are easily interpretable, they often fail to concisely express semantic intent. **Soft prompts** are vectors with the same dimensions as token embeddings in the language model's dictionary (Zhao et al., 2023). While soft prompts provide less interpretability compared to hard prompts, they capture semantic nuance more concisely.

Existing soft prompt methods such as GIST tokens (Mu et al., 2023), ICAE (Ge et al., 2024), and AutoCompressors (Chevalier et al., 2023) distill long inputs into soft tokens but assume uniform information density via constant token budgets or randomized segmentation—a flawed assumption given natural language's uneven semantic information distribution (Yu et al., 2016). Information-theoretic approaches such as LLMLingua (Jiang et al., 2023) and Selective Context (Li et al., 2023) have shown that token-level perplexity or self-information can effectively identify semantically important input regions, but apply only to hard prompts.

DAST (Chen et al., 2025) similarly implements dynamic allocation of soft tokens and may be considered parallel work, however, it is built on the Activation Beacon framework (Zhang et al., 2024) utilizing a different compression schema. We believe the implementation of DAST leaves room for improvement in method details and depth of experiments.

[*] Equal contribution
[†] Senior author

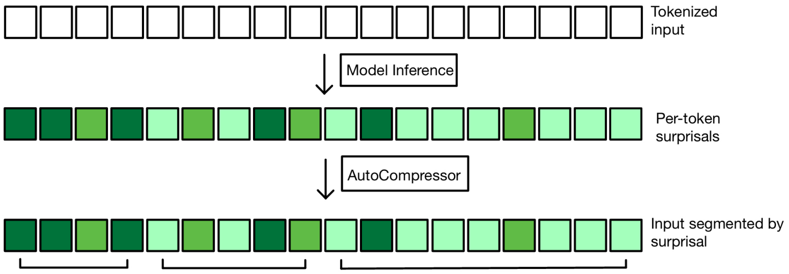

Figure 1: When fine-tuning AutoCompressors, we first quantify information density by obtaining per-token surprisals via baseline model inferencing on the tokenized input. We dynamically segment the input based on these surprisals, accumulating until a surprisal threshold τ is reached, resulting in variable-length segments and a variable number of segments per input sequence. These segments are then compressed into summary vectors, which are passed onto subsequent compression steps as soft prompts for the previous context.

We aim to bridge this gap by proposing a method that extends the AutoCompressor framework to incorporate surprisal-based dynamic segmentation. Specifically, we show that input segments of similar cumulative information produce more useful compressed representations when compressed into summary vectors, allowing for better language modeling performance on a variety of tasks.

Our main contribution is an extension of the AutoCompressor framework by introducing a new method for dynamic segmentation in soft prompt compression. Our method accounts for non-uniform information distribution across long contexts via the information-theoretic metric of surprisal, allowing for improved allocation of summary vectors as compared to allocation across randomized segments. We demonstrate the functionality of this methodology through cross-entropy loss and in-context learning (ICL) accuracy.

2 Related Work

We adopt the AutoCompressor framework from Chevalier et al. (2023), which builds on the recurrent memory transformer (RMT) architecture (Bulatov et al., 2022) to compress plain text into short soft prompts known as **summary vectors** (Lester et al., 2021). The tokenized input text is split into segments, with lengths randomly determined given a fixed hyperparameter of the number of segments. Segment lengths are guaranteed to be within the model's context window length. After generating summary vectors, the vectors are then prepended to all subsequent segments to recursively generate summary vectors over the entire segmented input.

Methods for extending a model's context window have been developed by previous work, such as RoPE-based scaling (Chen et al. (2023), Rozière et al. (2024), Ding et al. (2024), Zhu et al. (2025)) and utilizing different types of embeddings (Sun et al., 2023b). Non-transformer based architectures have also been proposed (Peng et al. (2023), Sun et al. (2023a)), allowing for extended context windows. However, these modifications do not perform well at longer scales or at a foundational level.

Other compression methods have been explored to tackle long-context input. Semantic Compression (Fei et al., 2024) utilizes graph-based chunking based on topic to dynamically compress context, but focuses on hard prompt compression through summarization techniques. DoDo (Qin et al., 2024) approaches compression architecturally by dynamically compressing the context via a trainable selector and compressor module to select and compress the most important hidden states in each layer to reduce computational intensity while maintaining model performance. Our work addresses the issue from the perspective of reducing input complexity via soft prompts.

3 Method

3.1 Framework

AutoCompressors are fine-tuned on base models and split long documents into a series of segments $S_1, \ldots, S_n$ with variable lengths constrained to fit within the model's context window. For each token x_t in a segment S_i with m_i tokens, the model is trained with the unsupervised objective of minimizing cross-entropy loss when conditioned on the previous tokens $x_1, \ldots, x_{t-1}$ and the previous summary vectors $\sigma_{<i}$:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{n} \sum_{t=1}^{m_i} \log p\left(x_t \mid x_1, \ldots, x_{t-1}, \sigma_{<i}\right)$$

over all segments and total number of tokens N. We follow this training objective when incorporating our method to fine-tune AutoCompressors.

3.2 Surprisal-Guided Segmentation

Our main contribution is implementing surprisal-based segmentation. The **surprisal** of a token x_t is the negative log probability of the token appearing given the preceding context (Ji et al., 2023):

$$\text{Surprisal}(x_t) = -\log P(x_t \mid x_{<t}).$$

Surprisal captures the model's uncertainty about the token in its generative context, thereby representing the information contained within the token; higher surprisal corresponds to more information, and lower surprisal corresponds to less information.

Token-level Segmentation. Given a tokenized input sequence $X = [x_1, x_2, \ldots, x_n]$, we define a segment $S_j = [x_{s_{j-1}+1}, \ldots, x_{s_j}]$ by computing per-token surprisal via baseline model inferencing and accumulating tokens until a fixed threshold τ is exceeded:

$$s_j := \min \left\{ k \in \{s_{j-1}+1, \ldots, n\} \;\middle|\; \sum_{i=s_{j-1}+1}^{k} \text{Surprisal}(x_i) \geq \tau \right\}$$

where $s_{j-1}+1$ and s_j are the start and end indices respectively of segment S_j. If the length of the segment exceeds the model's context window before the threshold is reached, we simply end the segment and begin a new segment. This procedure creates segments of roughly equal cumulative surprisal. Unlike the original methodology, which specifies a fixed number of segments for each training substep, we allow for a variable number based on the information distribution of the input, creating more flexibility in the fine-tuning process.

4 Experiments

4.1 Experimental Setup

We fine-tune an AutoCompressor model on a pre-trained OPT model (Zhang et al., 2022) with 1.3 billion parameters, fine-tuning on 6K-token sequences from the Wikipedia subdomain of the Pile dataset (Gao et al., 2020) with a surprisal threshold of $\tau = 1500$. This threshold was heuristically determined based on the total cumulative surprisal across sample input sequences.

Fine-tuning was done with 2-3 NVIDIA H100 GPUs each with 80 GB of memory, with one GPU solely dedicated to baseline model inferencing to obtain surprisal calculations. To

Model	Cross-Entropy	7	46	209	1071	4489	19972
OPT-1.3b	4.20	62.61	55.53	52.25	74.40	53.50	59.07
Baseline AC	2.66	67.16	63.50	60.46	71.30	55.30	**62.71**
Dynamic AC	**2.61**	**69.12**	**69.82**	**64.88**	**76.80**	**58.50**	62.70

Table 1: Evaluation results for a baseline OPT-1.3b model, baseline AutoCompressor (AC), and our dynamic AutoCompressor (AC). We evaluate cross-entropy loss on the Gutenberg subdomain and ICL accuracy on the AG News dataset with 6 different seeds.

ensure that the input would not exceed the base model's context window length during inferencing, we apply the extended full attention methodology introduced in the original work via extension of positional embeddings. Specifically, positional embeddings are reused beyond the model's context window length to allow for longer input sequences.

We evaluate our model by evaluating the out of domain cross-entropy loss on 6K-token sequences from the Gutenberg subdomain (consisting of various works of literature) of the Pile dataset. We split the input sequences into segments of 2,048 tokens except for the last segment, which has fewer tokens. Then, we compress all segments except the last, pass their summary vectors forward as soft prompts, and evaluate cross-entropy loss on the final segment.

We also evaluate in-context learning (ICL) accuracy on the AG News benchmark, which involves 4-way topic classification on news articles. Following the original implementation, we construct 10-shot prompts by sampling and concatenating 10 plain text training examples.

4.2 Results

We display our results in Table 1. We compare to a baseline OPT-1.3b model with extended full attention as well as an AutoCompressor model utilizing randomized segmentation as in the original methodology.

Our dynamic AutoCompressor achieves lower cross-entropy loss (2.61) compared to the baseline AutoCompressor (2.66) when evaluated on the out of domain Gutenberg dataset, demonstrating improved generative model prediction. On the AG News classification task, Dynamic AC outperforms the random segmentation baseline on most seeds, showing performance gain due to surprisal-aligned compression. For example, on Seed 7 and Seed 209, dynamic AutoCompressor improves accuracy in text classification by approximately +2% and +4.4%.

5 Discussion and Conclusion

We fine-tune an OPT model as an AutoCompressor using surprisal-based segmentation when partitioning input, determining segment boundaries by a surprisal threshold. We evaluate out of domain cross-entropy loss and ICL accuracy as compared to the pretrained baseline model and the randomized segmentation AutoCompressor, showing improved model performance. While gains are not uniform across all possible seeds or downstream tasks, future experiments may provide deeper insights.

We were unable to fine-tune larger models as AutoCompressors or fine-tune and evaluate on more subdomains due to budget and time constraints, potentially limiting generalizability. We were also unable to adjust the surprisal threshold τ as a hyperparameter for the same reasons. Future work should consider scaling to larger models and explore the effect of varying τ for optimization, as well as evaluation on a wider range of tasks. Furthermore, research is needed to consider the compression ratio presented by dynamic segmentation.

References

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11079–11091, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/47e288629a6996a17ce50b90a056a0e1-Paper-Conference.pdf.

Shaoshen Chen, Yangning Li, Zishan Xu, Yinghui Li, Xin Su, Zifei Shan, and Hai-tao Zheng. Dast: Context-aware compression in llms via dynamic allocation of soft tokens. *arXiv preprint arXiv:2502.11493*, 2025. URL https://arxiv.org/abs/2502.11493.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3829–3846, Singapore, 2023. doi: 10.18653/v1/2023.emnlp-main.232. URL https://aclanthology.org/2023.emnlp-main.232/.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

Weizhi Fei, Xueyan Niu, Pingyi Zhou, Lu Hou, Bo Bai, Lei Deng, and Wei Han. Extending context window of large language models via semantic compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5169–5181, Bangkok, Thailand, 2024. doi: 10.18653/v1/2024.findings-acl.306. URL https://aclanthology.org/2024.findings-acl.306/.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model, 2024. URL https://arxiv.org/abs/2307.06945.

Shaoxiong Ji, Wei Sun, and Pekka Marttinen. Content reduction, surprisal and information density estimation for long documents. *arXiv preprint arXiv:2309.06009*, 2023. URL https://arxiv.org/abs/2309.06009.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, Singapore, 2023. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Yucheng Li, Bo Dong, Frank Guérin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6342–6353, Singapore, 2023. doi: 10.18653/v1/2023.emnlp-main.391. URL https://aclanthology.org/2023.emnlp-main.391/.

Zongqian Li, Yinhong Liu, Yixuan Su, and Nigel Collier. Prompt compression for large language models: A survey. *arXiv preprint arXiv:2410.12388*, 2024. URL https://arxiv.org/abs/2410.12388.

Jesse Mu, Xiang Li, and Noah Goodman. Learning to compress prompts with gist tokens. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19327–19352, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3d77c6dcc7f143aa2154e7f4d5e22d68-Paper-Conference.pdf.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanislaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. Rwkv: Reinventing rnns for the transformer era, 2023. URL https://arxiv.org/abs/2305.13048.

Guanghui Qin, Corby Rosset, Ethan Chau, Nikhil Rao, and Benjamin Van Durme. Dodo: Dynamic contextual compression for decoder-only lms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 9961–9975, Bangkok, Thailand, 2024. doi: 10.18653/v1/2024.acl-long.536. URL https://aclanthology.org/2024.acl-long.536/.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024. URL https://arxiv.org/abs/2308.12950.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, 2016. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009/.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models, 2023a. URL https://arxiv.org/abs/2307.08621.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14590–14604, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.816. URL https://aclanthology.org/2023.acl-long.816/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2402.02244*, 2024. URL https://arxiv.org/abs/2402.02244.

Shuiyuan Yu, Jin Cong, Junying Liang, and Haitao Liu. The distribution of information content in english sentences. *CoRR*, abs/1609.07681, 2016. URL http://arxiv.org/abs/1609.07681.

Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. Long context compression with activation beacon. *arXiv preprint arXiv:2401.03462*, 2024. URL https://arxiv.org/abs/2401.03462.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

Wenbo Zhao, Arpit Gupta, Tagyoung Chung, and Jing Huang. SPC: Soft prompt construction for cross-domain generalization. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP)*, pp. 118–130, Toronto, Canada, 2023. doi: 10.18653/v1/2023.repl4nlp-1.10. URL https://aclanthology.org/2023.repl4nlp-1.10/.

Wenqiao Zhu, Chao Xu, Lulu Wang, and Jun Wu. Psc: Extending context window of large language models via phase shift calibration, 2025. URL https://arxiv.org/abs/2505.12423.