DAST: Context-Aware Compression in LLMs via Dynamic Allocation of Soft Tokens

Shaoshen Chen[1*], **Yangning Li**[1,2*], **Zishan Xu**[1*],
Yongqin Zeng[1], **Shunlong Wu**[1], **Xinshuo Hu**[3],
Zifei Shan[4], **Xin Su**[4], **Jiwei Tang**[1], **Yinghui Li**[1], **Hai-Tao Zheng**[1,2†]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Peng Cheng Laboratory, [3]Harbin Institute of Technology (Shenzhen)
[4]WeChat,Tencent
css24@mails.tsinghua.edu.cn
zheng.haitao@sz.tsinghua.edu.cn

Abstract

Large Language Models (LLMs) face computational inefficiencies and redundant processing when handling long context inputs, prompting a focus on compression techniques. While existing semantic vector-based compression methods achieve promising performance, these methods fail to account for the intrinsic information density variations between context chunks, instead allocating soft tokens uniformly across context chunks. This uniform distribution inevitably diminishes allocation to information-critical regions. To address this, we propose Dynamic Allocation of Soft Tokens (DAST), a simple yet effective method that leverages the LLM's intrinsic understanding of contextual relevance to guide compression. DAST combines perplexity-based local information with attention-driven global information to dynamically allocate soft tokens to the informative-rich chunks, enabling effective, context-aware compression. Experimental results across multiple benchmarks demonstrate that DAST surpasses state-of-the-art methods.[1]

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Xu et al., 2025; Zhang et al., 2025b; Li et al., 2024c,b; Liu et al., 2024; Yu et al., 2024; Huang et al., 2024; Li et al., 2023, 2025a; Kuang et al., 2024; Li et al., 2025c) have demonstrated remarkable performance on long context tasks (Li et al., 2022c, 2024d,a; Ma et al., 2022; Li et al., 2022b; Du et al., 2024), excelling capturing complex dependencies and generating coherent responses over extended contexts (Li et al., 2025b; Ye et al., 2023b,a; Huang et al., 2023). Nevertheless, processing long contexts incurs high computational cost, making the development of efficient **context compression** methods that preserve semantic integrity while reducing input length crucial.

Early approaches to context compression primarily relied on context pruning or summarization (Dong et al., 2023; Jiang et al., 2023; Pan et al., 2024; Tang et al., 2025a), which reduced input length through content removal or rephrasing. However, these methods often compromise semantic integrity through direct modification of the input sequence. Recent semantic vector-based methods (Li et al., 2022a; Liu et al., 2022; Cheng et al., 2024; Zhang et al., 2025a; Tang et al., 2025b) address this limitation by replacing the original context of length n with m compressed soft tokens ($m \ll n$), preserving essential information in a more compact representation. Although effective, these methods typically append soft tokens at the context terminus or distribute them uniformly, overlooking uneven information density across context chunks. This uniform distribution prevents optimal allocation of compression capacity to information-rich regions.

Notably, text-pruning-based approaches like LongLLMLingua (Jiang et al., 2024) attempt dynamic pruning using external models to estimate tokens importance. However, this external guidance fails to capture the LLM's intrinsic understanding of information relevance, creating incompatibility with vector-based methods.

This raises a key research question: **How can we dynamically allocate compression tokens based on the LLM's inherent understanding of contextual information density?**

To address this, we propose **D**ynamic **A**llocation of **S**oft **T**okens (**DAST**), a simple yet effective approach to soft tokens compression that fully leverages the LLM's internal capabilities without requiring external models. DAST utilizes perplexity to assess local importance and attention mechanisms to capture global relevance, dynamically allocating

[*]indicates equal contribution.
[†]Corresponding author. Yangning is the project leader.
[1]Our code can be available at: https://github.com/chenchen chen77/DAST

Findings of the Association for Computational Linguistics: ACL 2025, pages 20544–20552
July 27 - August 1, 2025

soft tokens based on intrinsic information density. This enables more efficient and context-aware compression, improving both compression quality and model performance compared to prior methods.

2 Method

2.1 Compression Background

Traditional methods for compressing long context sequences typically employ chunk-based decomposition. Given an input sequence $X = \{X^{\text{que}}, X^{\text{doc}}\}$, where X^{que} denotes a query or instruction and X^{doc} represents a lengthy document, the sequence is segmented into N contiguous chunks of fixed length $|\boldsymbol{X}_i| = L$. During compression, each chunk of length L is condensed into a fixed number m of soft tokens, where $m \ll L$.

Existing compression strategies, as illustrated in Figure 1(a) and (b), can be broadly categorized into two paradigms. The first, termed ***Single-chunk Compression***, processes the entire sequence as a single chunk and appends all fixed m soft tokens after the full sequence. To enhance granularity, methods such as AutoCompress (Chevalier et al., 2023) and Beacon (Zhang et al., 2025a) introduced ***Multi-chunks Compression***, which assigns a compression constraint (divisible by the chunk length L) stochastically during training and evenly distributes a fixed m soft tokens across all chunks during inference.

However, a major limitation of these methods is their **fixed tokens allocation scheme**, which implicitly assumes uniform information density across the entire context. This assumption introduces the risk that regions with high information density receive fewer soft tokens, while regions with low information density are allocated more soft tokens. To address this issue, we propose a *Dynamic Allocation of Soft Tokens* method, which adaptively assigns a **dynamic number of soft tokens** d_i to each chunk $\boldsymbol{X}_i$, where d_i is determined by localized and global information density, as shown in Figure 1(c).

2.2 Overall Framework

Our method dynamically determines the number of soft tokens assigned to each chunk, as described in the next section. Given the i-th chunk, the compressed representation is constructed as:

$$\boldsymbol{C}_i = \{\langle \boldsymbol{ct} \rangle_1, \ldots, \langle \boldsymbol{ct} \rangle_{i-1}, \boldsymbol{X}_i, \langle \boldsymbol{ct} \rangle_i\}, \quad (1)$$

where $\langle \boldsymbol{ct} \rangle_j \in \mathbb{R}^{d_j}$ represents the compressed soft tokens of the j-th chunk ($1 \leq j < i$). The compressed tokens $\langle \boldsymbol{ct} \rangle_i \in \mathbb{R}^{d_i}$ capture essential information from the current chunk through contextual interactions, which are facilitated by the cross-attention mechanism:

$$\text{CrossAttn.}(\boldsymbol{C}_i; \text{Mask}). \quad (2)$$

2.3 Dynamic Allocation

In this section, To effectively allocate soft tokens, we consider both **local importance** (i.e., within each chunk) and **global importance** (i.e., across the entire sequence). Specifically, we employ **Perplexity (PPL)** to estimate local importance and **Attention (Attn)** to capture global importance. These two metrics are then combined to determine the number of soft tokens assigned to each chunk.

PPL: Perplexity is a widely used metric for evaluating contextual informativeness. A lower perplexity signifies greater relevance of the current context information (Jiang et al., 2023). Since each chunk is visible within its own local context during compression, we compute the perplexity for each chunk separately. The resulting perplexity scores thus embody the local importance information of each respective chunk. The PPL of i-th chunk as follows:

$$P_i = -\sum_{l=1}^{L} q(x_l) \log p(x_l | x_{<l}), \quad (3)$$

where $q(x_l)$ represents the probability distribution of the ground truth.

Attn: Once the sequence has been compressed, the importance of each chunk's compressed representation can be inferred from attention weights. Intuitively, chunks with more crucial information have higher attention weights, reflecting their contribution to global understanding. Hence, we utilized global attention to measure the global information, which has been demonstrated to be an effective method (Liu et al., 2021). Through attention weights, we can ascertain the global proportion of each token. The formula is :

$$A_i = \sum_{j=i}^{(i+1)m} (qk^\top)_j, \quad (4)$$

where q represents the last token vector and k represents all the compressed tokens vectors. Thus,

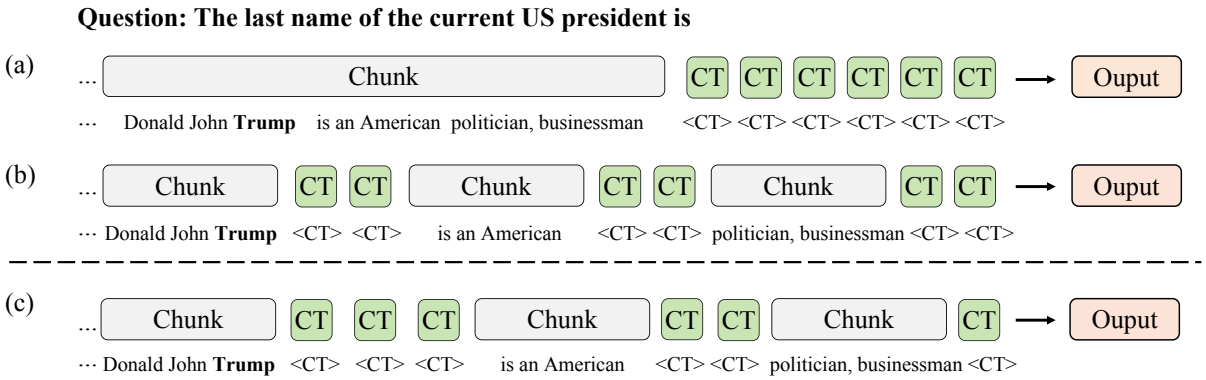

Figure 1: (a) and (b): the previous **fixed** allocation methods include **Single-chunk** and **Multi-chunks compression**; (c): our **dynamic** allocation method. Notably, our dynamic method allocated more soft tokens to the key information in the answer (highlighted in **bold**) while reducing soft tokens for less information. <CT> is compress soft token.

we can obtain the scores of i-th chunk that focus on both global and local information:

$$S_i = A_i \cdot \alpha - \frac{P_i}{\sum_{k=1}^{N}(P_k)} \cdot (1-\alpha), \quad (5)$$

where α is a parameter that balances the importance of global A_i and local P_i information. Then Softmax is used for normalization. Noted that since A_i and P_i are derived from different distributions, P_i is scaled by the number of all chunk N. Given the total number of soft tokens of context is M, we can calculate the actual number of soft tokens in the i-th chunk $d_i = M \times S_i$.

Reallocation: In order to make the tokens after dynamic allocation consistent with the training, we design a reallocation algorithm to make them divisible by L. The details of this algorithm are shown in Appendix A. It is worth mentioning that the reallocation is optional, depending on the method mechanism used.

3 Experiments

The experiment's detail are presented in Appendix B. Next, we want to answer two questions: (1) How effective is DAST? (2) How does the performance of DAST improve?

3.1 Main Results

To evaluate the dynamic distribution capability of DAST in a long context with inconsistent ground truth granularity distribution, we employ three benchmarks from LongBench (Bai et al., 2024): Single-Document, Multi-Document, and Example tasks (Few-Shot). As demonstrated in Table 1, our approach demonstrates consistent superiority over baseline methods across all evaluated tasks. This improvement can be attributed to the model's adaptive capacity to discern textual saliency within redundant, extended textual inputs. Our method strategically allocates a higher proportion of soft tokens to semantically critical chunks, thereby enhancing computational attention to pivotal content. This differential allocation mechanism ultimately optimizes task-specific performance through context-aware resource distribution.

Having established the effectiveness of our dynamic compression method, we further examine its performance-enhancing mechanisms through systematic experiments on the NaturalQuestions dataset (Kwiatkowski et al., 2019). This benchmark is particularly well-suited for analysis due to the presence of correct answers at varying contextual positions. As shown in Figure 2, our method consistently surpasses the uniform compression approach of Beacon across all positional configurations. Notably, in context chunks containing answer-relevant, our method adaptively allocates more tokens to these semantically critical regions, thereby improving performance.

3.2 Comparison of Different Constraints

We compare our method to baselines on the Long-Term Memory MSC dataset (Packer et al., 2023) to study compression intensity vs. memory retention. As presented in Table 2, our approach consistently outperforms conventional methods across all compression levels, especially in resisting degradation at higher compression constraints.

Methods	Document and Example Compression							
	Single Doc	Multi Doc	Few Shot	AVG	Single Doc	Multi Doc	Few Shot	AVG
	LLama-2-7B				Qwen-2-7B			
Original Prompt	24.9	22.5	60.0	35.8	22.0	29.3	62.3	37.9
Zero-Shot	8.1	6.1	32.2	15.5	7.1	6.6	26.8	13.5
AutoComp.† (Chevalier et al., 2023)	12.9	16.4	23.8	17.7		-		
ICAE† (Ge et al., 2024)	19.5	19.2	24.8	21.2		-		
LongLLM.† (Jiang et al., 2024)	21.5	18.8	49.5	29.9	24.7	20.3	55.9	33.6
SnapKV† (Li et al., 2024e)	24.2	22.6	60.1	35.6	38.7	37.6	67.1	47.8
Beacon† (Zhang et al., 2025a)	34.9	27.5	61.4	41.3	40.5	40.3	68.4	49.7
DAST (Ours)	**38.1**	**37.4**	**63.6**	**46.4**	**40.6**	**45.6**	**68.6**	**51.6**

Table 1: Evaluation of various Document and Example Compression tasks (top performances marked in **bold**). †: the results cited from Zhang et al. (2025a).

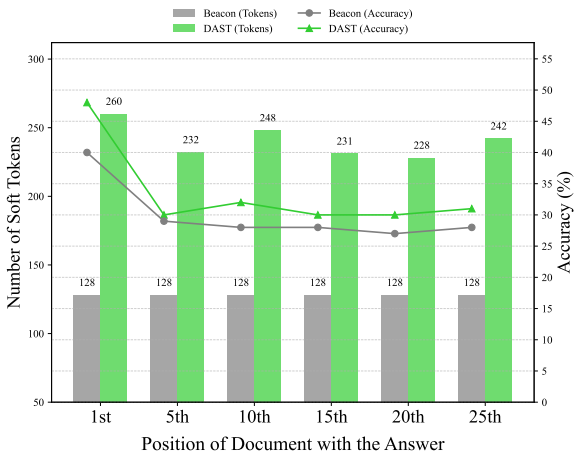

Figure 2: Performance and Number of soft tokens v.s. Key Information Position.

Method	Compression Constraint			
	∼4 x	∼8 x	∼16 x	∼24 x
AutoCom.	28.8	$27.3_{\downarrow 5.2\%}$	$25.0_{\downarrow 13.2\%}$	$24.0_{\downarrow 17.1\%}$
LongLLM.	22.4	$19.5_{\downarrow 13.0\%}$	$17.8_{\downarrow 20.5\%}$	$15.9_{\downarrow 29.0\%}$
ICAE	18.1	$16.6_{\downarrow 8.3\%}$	$15.3_{\downarrow 15.5\%}$	$14.5_{\downarrow 19.9\%}$
Beacon	39.0	$36.5_{\downarrow 6.4\%}$	$33.6_{\downarrow 13.9\%}$	$32.3_{\downarrow 17.2\%}$
DAST	**55.9**	$\mathbf{55.5}_{\downarrow 0.7\%}$	$\mathbf{52.6}_{\downarrow 5.9\%}$	$\mathbf{51.9}_{\downarrow 7.2\%}$

Table 2: Evaluation of Long-Term Memory on MSC. ↓: percentages showing relative performance drop compared to the ∼4x compression baseline.

Method	Single-Doc
Random Allocation	34.52
Uniform Allocation	34.90
Dynamic Allocation (ours)	**38.14**
w/o PPL	37.60
w/o Attn	37.24

Table 3: Ablation study of DAST.

3.3 Ablation Study

In this section, we analyze our method's performance and the impact of each module (see Table 3). Random and uniform tokens allocations performed poorly, showing insufficient focus on critical context segments. Removing either the global attention (Attn) or local perplexity (PPL) module individually caused performance drops, highlighting their importance in prioritizing important chunks.

3.4 Parameter Sensitivity Analysis

It is important to analyze the sensitivity of the parameters α. As shown in Figure 3. The results indicate that the performance remains stable across different values of α. Consequently, we selected a default value of $\alpha = 0.5$ in main experiments, which simplifies the application of our method to other models, as it eliminates the need for specialized parameter tuning.

4 Conclusion

In this paper, we propose DAST, a simple yet effective method that dynamically allocates soft tokens by leveraging the LLM's intrinsic perception of information density. By integrating perplexity-based local information and attention-driven global relevance, DAST adaptively focuses compression capacity on high-information regions without re-

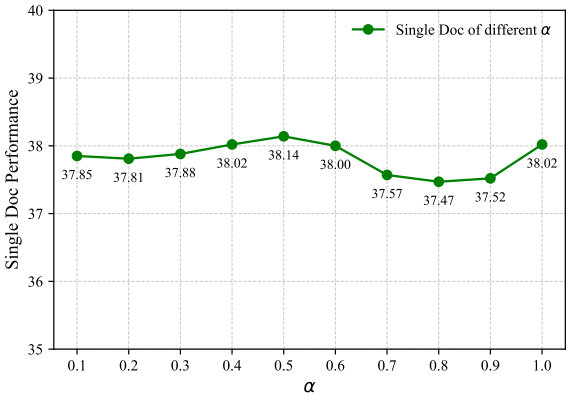

Figure 3: Parameter Sensitivity Analysis of α.

lying on external models. Our experiments show that DAST outperforms prior methods in both compression quality and downstream task performance, underscoring the value of model-guided dynamic allocation.

Limitations

Current model compression research remains primarily limited to the approximately 7B parameter scale due to computational resource constraints. While our study has demonstrated that our method outperforms other compression approaches, we have not been able to systematically investigate whether existing compression techniques, including our approach, can maintain their effectiveness when applied to larger architectures. Given that the practical value of compression techniques becomes more pronounced with increasing model sizes, this represents a critical direction for future research. Furthermore, it is also imperative to examine whether the compression process induces more severe : (1) hallucination phenomena and (2) catastrophic forgetting in compressed models, which constitutes another essential aspect requiring thorough investigation.

Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No. 62276154), Research Center for Computer Network (Shenzhen) Ministry of Education, the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012914 and 44030024103310080177O), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033, JCYJ20240813112009013 and GJHZ20240218113603006), the Major Key Project of PCL (NO. PCL2024A08).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. LongloRA: Efficient fine-tuning of long-context large language models. In *The Twelfth International Conference on Learning Representations*.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xrag: Extreme context compression for retrieval-augmented generation with one token. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3829–3846.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2023. A survey of natural language generation. *ACM Comput. Surv.*, 55(8):173:1–173:38.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 2024.

Llms assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5081–5099. Association for Computational Linguistics.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2024. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations*.

Haojing Huang, Jingheng Ye, Qingyu Zhou, Yinghui Li, Yangning Li, Feng Zhou, and Hai-Tao Zheng. 2023. A frustratingly easy plug-and-play detection-and-reasoning module for chinese spelling check. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11514–11525. Association for Computational Linguistics.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Haitao Zheng. 2024. Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10186–10197. ELRA and ICCL.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13358–13376. Association for Computational Linguistics.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. Booksum: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558.

Jiayi Kuang, Jingyou Xie, Haohao Luo, Ronghao Li, Zhe Xu, Xianfeng Cheng, Yinghui Li, Xika Lin, and Ying Shen. 2024. Natural language understanding and inference with MLLM in visual question answering: A survey. *CoRR*, abs/2411.17558.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and Philip S. Yu. 2025a. Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yangning Li, Tingwei Lu, Hai-Tao Zheng, Yinghui Li, Shulin Huang, Tianyu Yu, Jun Yuan, and Rui Zhang. 2024a. MESED: A multi-modal entity set expansion dataset with fine-grained semantic classes and hard negative entities. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 8697–8706. AAAI Press.

Yinghui Li, Haojing Huang, Jiayi Kuang, Yangning Li, Shu-Yu Guo, Chao Qu, Xiaoyu Tan, Hai-Tao Zheng, Ying Shen, and Philip S. Yu. 2025b. Refine knowledge of large language models via adaptive contrastive learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023. On the (in)effectiveness of large language models for chinese text correction. *CoRR*, abs/2307.09007.

Yinghui Li, Jiayi Kuang, Haojing Huang, Zhikun Xu, Xinnian Liang, Yi Yu, Wenlian Lu, Yangning Li, Xiaoyu Tan, Chao Qu, Ying Shen, Hai-Tao Zheng, and Philip S. Yu. 2025c. One example shown, many concepts known! counterexample-driven conceptual reasoning in mathematical llms. *CoRR*, abs/2502.10454.

Yinghui Li, Yangning Li, Yuxin He, Tianyu Yu, Ying Shen, and Hai-Tao Zheng. 2022a. Contrastive learning with hard negative entities for entity set expansion. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1077–1086. ACM.

Yinghui Li, Shirong Ma, Qingyu Zhou, Zhongli Li, Yangning Li, Shulin Huang, Ruiyang Liu, Chao

Li, Yunbo Cao, and Haitao Zheng. 2022b. Learning from the dictionary: Heterogeneous knowledge guided fine-tuning for chinese spell checking. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 238–249. Association for Computational Linguistics.

Yinghui Li, Shang Qin, Jingheng Ye, Shirong Ma, Yangning Li, Libo Qin, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, and Philip S. Yu. 2024b. Rethinking the roles of large language models in chinese grammatical error correction. *CoRR*, abs/2402.11420.

Yinghui Li, Zishan Xu, Shaoshen Chen, Haojing Huang, Yangning Li, Shirong Ma, Yong Jiang, Zhongli Li, Qingyu Zhou, Hai-Tao Zheng, and Ying Shen. 2024c. Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8656–8668. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022c. The past mistake is the future wisdom: Error-driven contrastive probability optimization for chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3202–3213. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S. Yu. 2024d. When llms meet cunning texts: A fallacy understanding benchmark for large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024e. Snapkv: Llm knows what you are looking for before generation. *arXiv preprint arXiv:2404.14469*.

Ruiyang Liu, Yinghui Li, Linmi Tao, Dun Liang, and Hai-Tao Zheng. 2022. Are we ready for a new paradigm shift? A survey on visual deep MLP. *Patterns*, 3(7):100520.

Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2024. Can llms learn uncertainty on their own? expressing uncertainty effectively in A self-training manner. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 21635–21645. Association for Computational Linguistics.

Yichao Liu, Zongru Shao, and Nico Hoffmann. 2021. Global attention mechanism: Retain information to enhance channel-spatial interactions. *CoRR*, abs/2112.05561.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Yangning Li, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen. 2022. Linguistic rules-based corpus generation for native chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 576–589. Association for Computational Linguistics.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. Memgpt: Towards llms as operating systems. *CoRR*, abs/2310.08560.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. 2024. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 963–981. Association for Computational Linguistics.

Jiwei Tang, Jin Xu, Tingwei Lu, Zhicheng Zhang, Yiming Zhao, Lin Hai, and Hai-Tao Zheng. 2025a. Perception compressor: A training-free prompt compression framework in long context scenarios. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4093–4108. Association for Computational Linguistics.

Jiwei Tang, Zhicheng Zhang, Shunlong Wu, Jingheng Ye, Lichen Bai, Zitai Wang, Tingwei Lu, Jiaqi Chen, Lin Hai, Hai-Tao Zheng, and Hong-Gee Kim. 2025b. Gmsa: Enhancing context compression via group merging and layer semantic alignment. *Preprint*, arXiv:2505.12215.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan,

Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Zhikun Xu, Yinghui Li, Ruixue Ding, Xinyu Wang, Boli Chen, Yong Jiang, Haitao Zheng, Wenlian Lu, Pengjun Xie, and Fei Huang. 2025. Let llms take on the latest challenges! A chinese dynamic question answering benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10435–10448. Association for Computational Linguistics.

Jingheng Ye, Yinghui Li, Yangning Li, and Hai-Tao Zheng. 2023a. Mixedit: Revisiting data augmentation and beyond for grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10161–10175. Association for Computational Linguistics.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023b. CLEME: debiasing multi-reference evaluation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6174–6189. Association for Computational Linguistics.

Tianyu Yu, Chengyue Jiang, Chao Lou, Shen Huang, Xiaobin Wang, Wei Liu, Jiong Cai, Yangning Li, Yinghui Li, Kewei Tu, Hai-Tao Zheng, Ningyu Zhang, Pengjun Xie, Fei Huang, and Yong Jiang. 2024. Seqgpt: An out-of-the-box large language model for open domain sequence understanding. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19458–19467. AAAI Press.

Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. 2025a. Long context compression with activation beacon. In *The Thirteenth International Conference on Learning Representations*.

Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou, Aiwei Liu, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, Feiran Huang, Sheng Zhou, Jiajun Bu, Allen Lin, James Caverlee, Fakhri Karray, Irwin King, and Philip S. Yu. 2025b. Cold-start recommendation towards the era of large language models (llms): A comprehensive survey and roadmap. *CoRR*, abs/2501.01945.

A Reallocation Algorithm

For the current soft tokens set $\boldsymbol{T}$ of all chunks given, the total soft tokens $\boldsymbol{S}$ can be assigned, and the optional compression rate set $\boldsymbol{R}$, our goal is to reassign $\boldsymbol{T}$ to get $\tilde{\boldsymbol{T}}$, the algorithm is as follows Algorithm 1.

Algorithm 1 Reallocation

Require: $\boldsymbol{T}$, $\boldsymbol{S}$,$\boldsymbol{R}$,
Ensure: Reallocated $\tilde{\boldsymbol{T}}$,

$\tilde{\boldsymbol{T}} \leftarrow \emptyset$
$\boldsymbol{T}$ are allocated to each chunk based on the closest compression constraint $\boldsymbol{R}$ to get the $\tilde{\boldsymbol{T}}$.
repeat
    Get the disposable $\boldsymbol{M}$ from $\boldsymbol{S}$ - $\boldsymbol{sum}(\tilde{\boldsymbol{T}})$
    In the remaining tokens, double the tokens of the chunk that meets the conditions with the highest score.
    Update $\tilde{\boldsymbol{T}}$
    Remove the highest score temporarily,
until $\boldsymbol{M} = 0$
return $\tilde{\boldsymbol{T}}$

B Settings

B.1 Implementation

To ensure strict methodological consistency in model comparisons, we employ the Llama-2-7B (chat) (Touvron et al., 2023) and Qwen-2-7B (Bai et al., 2023) architectures. For training data selection, we adopt the same approach as Beacon(Zhang et al., 2025a), utilizing 1B tokens sampled from Repajama (Weber et al., 2024) during pre-training, supplemented by LongAlpaca (Chen et al., 2024), BookSum (Kryściński et al., 2022), and synthetic data generated by GPT-3.5 for fine-tuning (see Beacon (Zhang et al., 2025a) for detailed data curation protocols). Our implementation uses a standard α

value of 0.5, with sensitivity analyses for alternative parameter configurations provided in §3.4. We use the HuggingFace framework (Wolf et al., 2019) and all experiments were conducted on a computational cluster equipped with 8 × A800 GPUs (80GB).

B.2 Baselines

We compare our method to a baseline (represented by the Original Prompt) with the same constraints and an uncompressed baseline with no long context data (represented by Zero-Shot). In addition, We also compared with the current mainstream including text pruning or summarization long context compression method and semantic vector-based long context compression methods. These include AutoCompressors(Chevalier et al., 2023), ICAE(Ge et al., 2024), LongLLMLingua(Jiang et al., 2024), SnapKV(Li et al., 2024e), and Beacon(Zhang et al., 2025a).

C Latency Analysis

LLama-2-7b		**Qwen-2-7b**	
Model	**Latency**	**Model**	**Latency**
AutoComp.	1.8	AutoComp.	-
ICAE	1.2	ICAE	-
LongLLM.	2.4	LongLLM.	4.5
SnapKV	**0.9**	SnapKV	**2.8**
Beacon	1.2	Beacon	3.3
DAST	1.6	DAST	3.8

Table 4: Latency Analysis.

As shown in Table 4, we analyzed the latency of DAST and each baseline method. The lantency of DAST is higher than some baseline, such as ICAE and Beacon, because they compress directly. The latency is lower than LongLLMlingua because our approach relies on the model itself for dynamic tokens allocation, whereas these baselines rely on external models.