FocusLLM: Scaling LLM's Context by Parallel Decoding

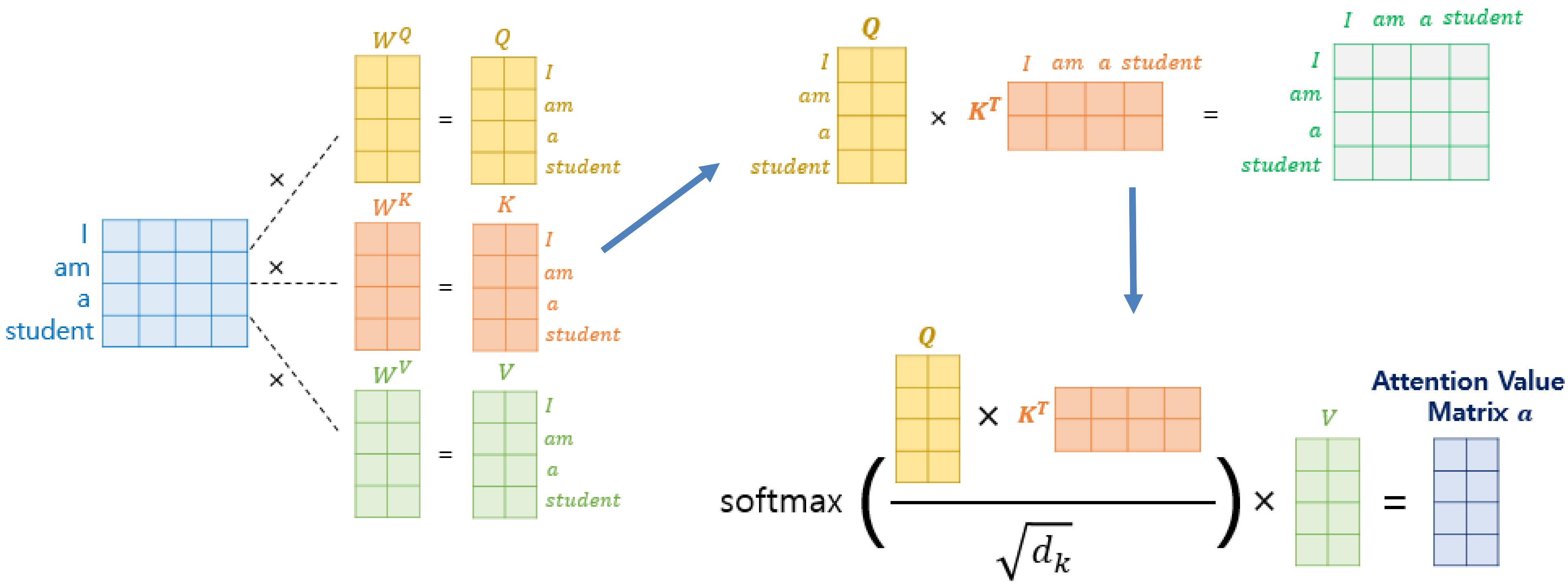
CIDA Lab.

전자전기컴퓨터공학부
2022440119 전형준

2024.08.29

Complexity of Attention Mechanism

Due to the attention mechanism, the computational Complexity is $O(L^2)$



https://wikidocs.net/31379

FocusLLM

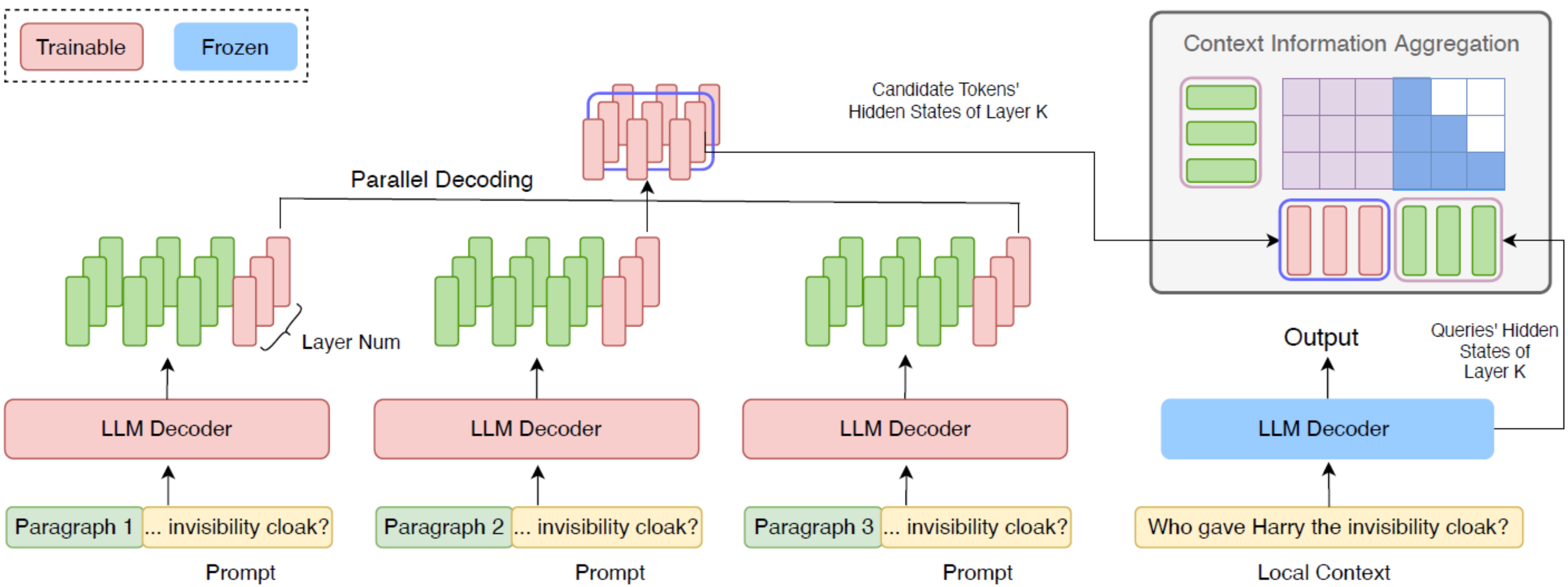
Framework designed to extend the context length of any decoder-only LLM, enabling the model to focus on relevant information from very long sequences.

FocusLLM – Notations

Long sequence with S tokens {x1, ..., xS}

Memory tokens {x1, ..., xm} and **Local tokens** {xm+1, ..., xS}

Concurrently, we divide the memory into **chunks**, labeled as C1,C2, ...,Ck

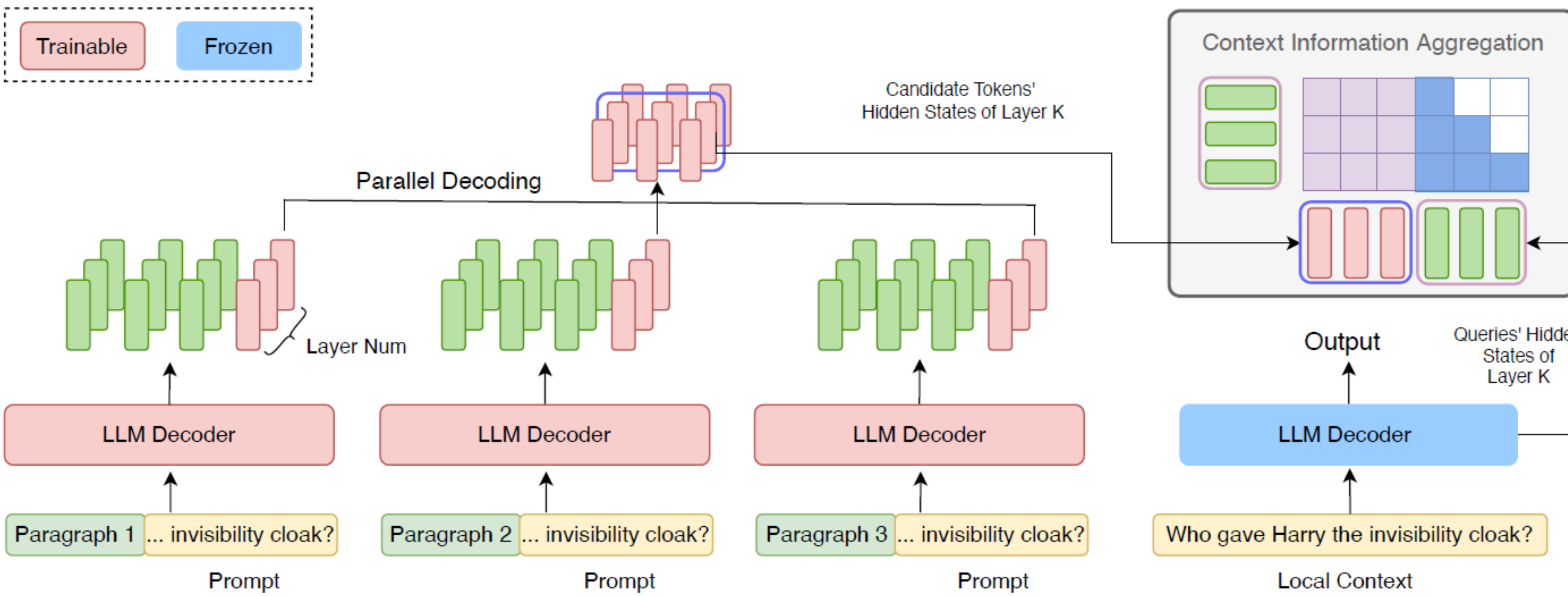


FocusLLM – Notations

Original decoder model as $\boldsymbol{F_{dec}}$, and its hidden dimension $\boldsymbol{d_{dec}}$

To generate candidate tokens, we introduce a small

set of new parameters resulting in the modified model $\boldsymbol{F'_{dec}}$



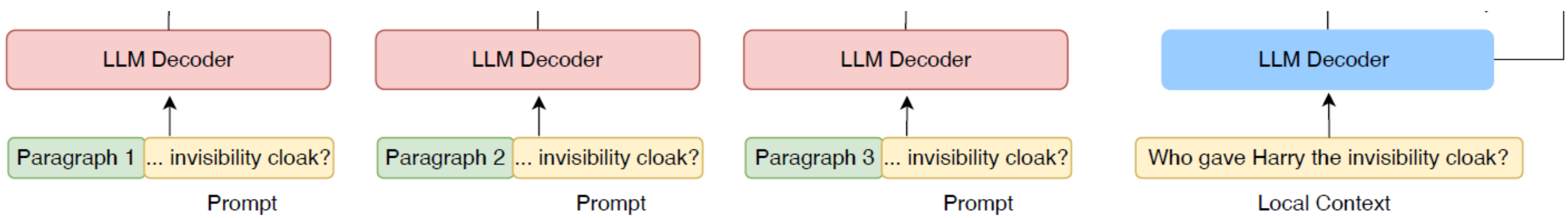
FocusLLM – Notations

Local Context Injection

We append a small fragment of local tokens

Behind each chunk and perform **parallel decoding.**

$$\hat{C}_i \leftarrow \{C_i; x_{m+j}, \ldots, x_S\} \; i = 1, \ldots, k; 1 \leq j \leq S-m$$

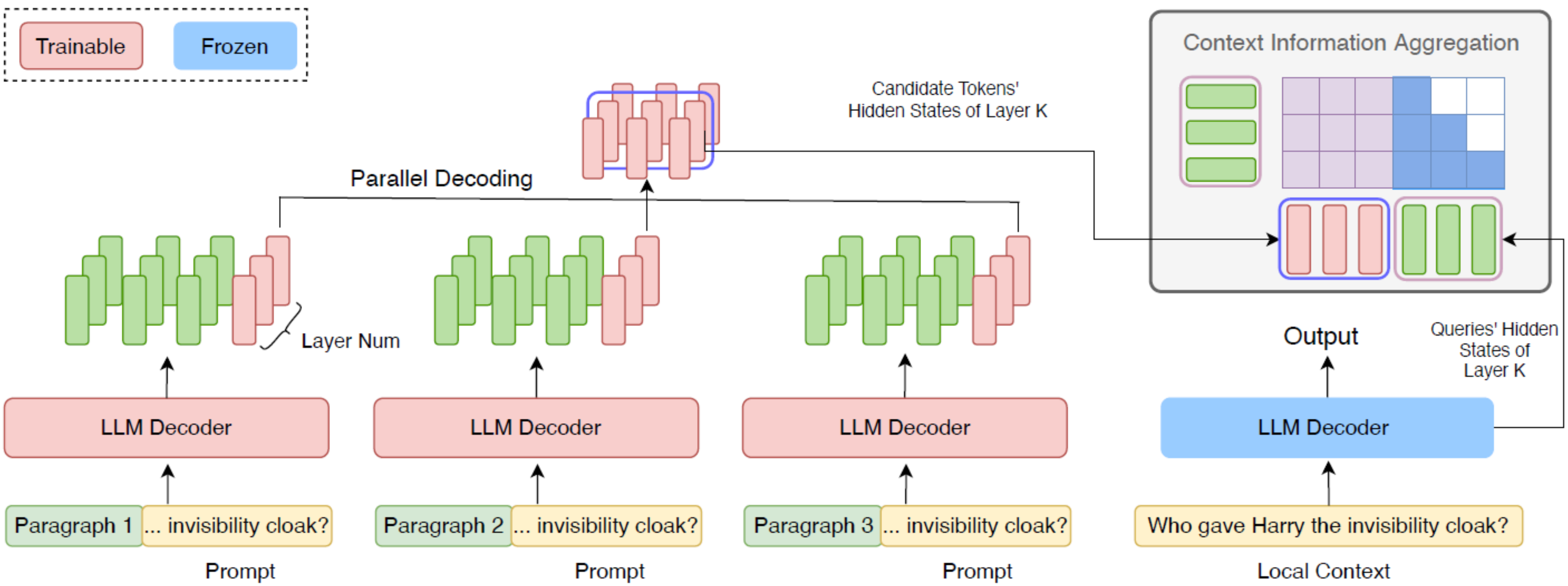


FocusLLM – Notations

The **candidate token** is the trainable hidden states

corresponding to the last local token xS in each chunk.

Whether this chunk contains information relevant to the local context.



FocusLLM – Notations

We only add a new set of trainable parameters to the

Linear projection matrices of each layer.

$$\{W'_Q, W'_K, W'_V, W'_O\}_l$$

$$e_i = F'_{dec}(\hat{C}_i)$$

Where e_i consists of key-value hidden states K_e and V_e of the last token in each layer.

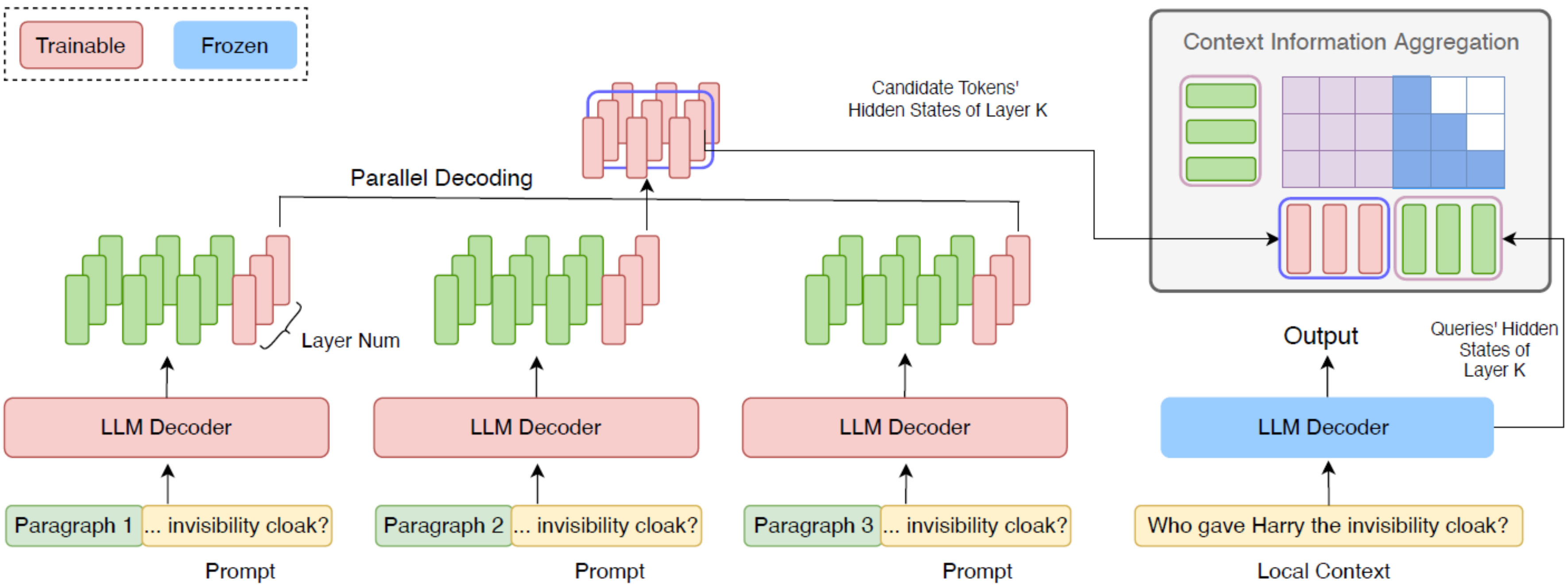
$$Q_e \leftarrow HW'_Q \quad K_e \leftarrow HW'_K \quad V_e \leftarrow HW'_V$$

$$A \leftarrow softmax\left(Q_e\left(K \oplus K_e\right)^T\right)$$

$$O_e \leftarrow V_e {W'_O}^T \quad V_e \leftarrow A\left(V \oplus V_e^T\right)$$

FocusLLM – Notations

Finally, the generated candidate tokens are concatenated with the local tokens and are subsequently processed by a frozen decoder.



FocusLLM – Notations

FocusLLM is trained using a natural auto-regressive method. Specifically, we train the model to predict the next token, which encourages the candidate token to aggregate useful information from each chunk.

$$\min_{F'_{dec}} - \sum_{i=2}^{L} \log(p(x_i \mid e_1, \ldots, e_k, x_1, \ldots, x_{i-1}; F'_{dec}))$$

FocusLLM – Notations

Specifically, based on the different selection methods for local tokens, we design two types of loss functions for joint training.

Continuation Loss: Last L tokens from a long document are selected as local tokens

Repetition loss: Take the entire long document as memory and then randomly select L continuous tokens from it as local tokens

Experiments - Long-context Language Modeling

We perform the evaluation on three datasets:

PG19, Proof-Pile, and CodeParrot.

These three datasets encompass 100 long test cases related to

books, arXiv papers, and code repositories, respectively.

Experiments - Long-context Language Modeling

Table 2: Language Modeling Assessment: perplexity analysis of various context scaling methods on the PG19, Proof-Pile, and CodeParrot. FocusLLM successfully extends context of the original Llama model and maintains low perplexity on extremely long sequences.

	PG19				Proof-Pile				CodeParrot			
Method	4K	16K	32K	100K	4K	16K	32K	100K	4K	16K	32K	100K
Llama-2-7B	9.21	$>10^3$	$>10^3$	OOM	3.47	$>10^3$	$>10^3$	OOM	2.55	$>10^3$	$>10^3$	OOM
PI	9.21	19.5	$>10^2$	OOM	3.47	5.94	33.7	OOM	2.55	4.57	29.33	OOM
NTK	9.21	11.5	37.8	OOM	3.47	3.65	7.67	OOM	2.55	2.86	7.68	OOM
StreamingLLM	9.21	9.25	9.24	9.32	3.47	3.51	3.50	3.55	2.55	2.60	2.54	2.56
AutoCompre.-6K	11.8	$>10^2$	$>10^3$	OOM	4.55	$>10^2$	$>10^3$	OOM	5.43	$>10^2$	$>10^3$	OOM
YaRN-128K	6.68	6.44	6.38	OOM	2.70	2.47	2.41	OOM	2.17	2.04	2.00	OOM
LongChat-32K	9.47	8.85	8.81	OOM	3.07	2.70	2.65	OOM	2.36	2.16	2.13	OOM
LongAlpaca-16K	9.96	9.83	$>10^2$	OOM	3.82	3.37	$>10^3$	OOM	2.81	2.54	$>10^3$	OOM
LongLlama	9.06	8.83	OOM	OOM	2.61	2.41	OOM	OOM	1.95	1.90	OOM	OOM
Activation Beacon	9.21	8.54	8.56	8.68	3.47	3.42	3.39	3.35	2.55	2.54	2.53	2.55
FocusLLM	9.21	9.19	9.17	10.59	3.47	3.17	3.43	2.57	2.55	2.01	2.27	3.02

Shortcomings

1. The dataset used in the experiment is formatted in a way that makes it easy to parse natural language into SAT format..
2. While the approach quickly and accurately derives solutions when a solution exists, it does not address cases where no solution exists.