

Parse Trees Guided LLM Prompt Compression

Wenhai Mao, Chengbin Hou, Tianyu Zhang, Xinyu Lin, Ke Tang, *Fellow IEEE*, Hairong Lv

Abstract—Offering rich contexts to Large Language Models (LLMs) has shown to boost the performance in various tasks, but the resulting longer prompt would increase the computational cost and might exceed the input limit of LLMs. Recently, some prompt compression methods have been suggested to shorten the length of prompts by using language models to generate shorter prompts or by developing computational models to select important parts of original prompt. The generative compression methods would suffer from issues like hallucination, while the selective compression methods have not involved linguistic rules and overlook the global structure of prompt. To this end, we propose a novel selective compression method called PartPrompt. It first obtains a parse tree for each sentence based on linguistic rules, and calculates local information entropy for each node in a parse tree. These local parse trees are then organized into a global tree according to the hierarchical structure such as the dependency of sentences, paragraphs, and sections. After that, the root-ward propagation and leaf-ward propagation are proposed to adjust node values over the global tree. Finally, a recursive algorithm is developed to prune the global tree based on the adjusted node values. The experiments show that PartPrompt receives the state-of-the-art performance across various datasets, metrics, compression ratios, and target LLMs for inference. The in-depth ablation studies confirm the effectiveness of designs in PartPrompt, and other additional experiments also demonstrate its superiority in terms of the coherence of compressed prompts and in the extreme long prompt scenario.

Index Terms—Large Language Models, Prompt Compression, Parse Trees, Prompt Structure, Text Pattern Analysis.

I. INTRODUCTION

LARGE Language Models (LLMs) have achieved remarkable performance on various tasks such as question answering, summarization, multimodal generation, and information extraction [1], [2]. Prompting LLMs with adequate task-related contexts can enhance their performance. The prompting techniques, like in-context learning [3], chain-of-thought [4], and retrieval augmented generation [5], have shown noticeable performance gains for answering questions that require long-tail knowledge [6] and reasoning ability [7].

Most LLMs adopt the transformer architecture, which results in the computational cost of LLMs being proportional to the square of input context length [8]. In addition, there is an upper token limit that can be processed by LLMs [8]. The use of long prompt for providing rich contexts to LLMs would

Wenhai Mao, Tianyu Zhang, and Hairong Lv are with the Ministry of Education Key Laboratory of Bioinformatics, BNRIST Bioinformatics Division, Department of Automation, Tsinghua University, Beijing, China.

Chengbin Hou and Xinyu Lin are with the School of Computer Science and Engineering, Fuyao University of Science and Technology, Fuzhou, China.

Ke Tang is with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China.

Corresponding authors: Chengbin Hou; Hailong Lv

E-mail: chengbin.hou10@foxmail.com; lvhairong@tsinghua.edu.cn

Submitted to IEEE TPAMI

Manuscript received xxxx xx, 2024; revised xxxx xx, 2024

also significantly increase the computational cost and might exceed the input token limit. To this end, prompt compression, by shortening the length of prompt, is suggested to reduce the computational cost during LLM inference and make LLMs possible for handling prompts beyond the input token limit (or indirectly increasing the limit).

Recently, some prompt compression methods [9], [10] feed the original prompt into a language model for generating the compressed prompt. These generative compression methods exploit the language understanding and generation ability of language models, but would encounter issues like hallucination [11] due to the limitations of generative language models. Another typical type of prompt compression methods [12]–[16] obtain the compressed prompt by selecting a portion of the original prompt without generating new contents, and thus alleviate the hallucination issue. These selective compression methods employ a measure to evaluate the importance of tokens in the original prompt, and preserve the important parts while outputting the compressed prompt.

Regarding existing selective prompt compression methods, the parts of the original prompt to remain or remove are determined mainly by a computational model, e.g., using a small language model to calculate information entropy as the measure. The computational models have not yet involved the linguistic rules, which have been previously shown effective in various learning tasks [17]–[20]. Besides, the computational cost of prompt compression methods is also deserved to consider, since one fundamental motivation of prompt compression is to reduce the computational cost for LLM inference [12], [13]. Moreover, the streaming processing of original prompt by compression methods may overlook the connecting patterns among sentences and the hierarchical structure of the whole prompt, which we refer as the human writing logic especially while writing long prompts.

To address these challenges, this work proposes to leverage Parse trees to guide the Prompt compressing process (namely PartPrompt) incorporating with the local information entropy and the global prompt patterns. Specifically, PartPrompt first analyzes a parse tree for each sentence and obtains the local information entropy of tokens within each sentence. Second, a global tree is constructed to reflect connecting patterns among sentences and hierarchical structure of the whole prompt. Third, the node value based on the entropy is adjusted over the global tree by the newly proposed root-ward propagation and leaf-ward propagation. And finally, a recursive algorithm is developed to prune the global tree based on the adjusted node value. It is noted that, during prompt compression, the linguistic rules are introduced by the parse trees of sentences; the computational cost is reduced by the local approximated entropy; the human writing logic is considered in the construction of global tree and the adjustment of node value.

Extensive experiments are conducted on various datasets to present the performance for understanding, summarization, in-context learning, and math question answering prompts. The results demonstrate that PartPrompt considerably outperforms the state-of-the-art prompt compression methods over several compression ratios and metrics. And the superior performance of PartPrompt can be also observed when feeding the compressed prompt to different LLMs as the target model for inference. Besides, in-depth ablation experiments confirm the effectiveness of the key components in PartPrompt such as introducing parse trees, constructing the global tree, and adjusting the node value. Moreover, we further explore the performance of PartPrompt under the extreme long prompt case, and investigate the coherence of the compressed prompt via quantitative metrics and intuitive examples.

Apart from extensive experiments and in-depth analysis, the main technical contributions are as follows:

- This work introduces parse trees, a kind of linguistic rules and text patterns, to guide the prompt compression process for the first time in literature. For this purpose, we transform the prompt compression problem of selecting tokens into the tree constructing and pruning problem.
- To capture the global structure, we propose to construct a global tree for the original prompt based on the dependency of sentences, paragraphs, and sections. And we further propose the novel root-ward propagation and leaf-ward propagation to adjust the value of nodes in the tree. Both techniques are motivated by the human writing logic.
- Unlike sequentially evaluating and removing unimportant parts as previous methods done, a new recursive algorithm is developed to remove the unimportant parts of original prompt by pruning over the global tree. And with the help of the global tree, it becomes reasonable to adopt local approximated information entropy, thereby saving computational cost compared to using global information entropy.
- To benefit future research, the code is freely available at <https://github.com/LegendaryHippopotamus/PartPrompt>

II. RELATED WORK

A. Large Language Models and Prompt Engineering

Early language models were applied for tasks such as text classification and machine translation [21], [22]. Subsequently, some works [8], [23], [24] discovered that the performance of language models is highly correlated to their parameter scales, leading to development of LLMs in pursuit of more powerful capabilities [25], [26]. After that, many LLMs with billions of parameters [1], [2] were developed and have achieved remarkable success in a wide variety of tasks.

The recent LLMs have also emerged many new abilities with the help of prompts [27] such as in-context learning [3], chain-of-thought [4], and retrieve augmented generative [5]. A multitude of works then attempt to enhance the performance of LLMs via prompt engineering. Some of them manually design prompts [4], whereas other works focus on designing prompts in a more automatic way [28].

Prompting LLMs with adequate task-related contexts can enhance the ability of LLMs. However, the prompt techniques

such as [3]–[5] often require a relatively longer prompt, which would considerably increase the computational cost during inference due to the transformer architecture. To tackle this issue, recent works have begun to explore prompt compression, i.e., using the compressed prompt to replace the given prompt while preserving performance. And this work also intends to investigate the prompt compression problem.

B. Prompt Compression

The prompt compression methods can be divided into two categories: generative compression and selective compression. Generative compression utilizes a language model to take the original given prompt as the input, which is then asked to generate the compressed prompt typically from the same language model.

Specifically, [10] directly employs LLMs to compress the given prompt, and this work then analyzes the ability of LLMs in retaining semantics and understanding contents after compression; [9] trains an encapsulation model with a semantic loss and a reward function, and the trained model is then adopted to generate the compressed prompt.

Selective compression, on the other hand, selects a portion of the original prompt as the compressed prompt, which can better preserve the original contents and avoid hallucinations.

Some selective compression methods directly employ an existing pretrained language model to evaluate the importance of tokens. Selective-Context [12] obtains the compressed prompt by retaining tokens with higher information entropy, which is computed by a pretrained language model. Then, LLMLingua [13] further introduces the budget controller, iterative token-level compression, and distribution alignment for a better information entropy estimation and a higher compression rate. Apart from them, other selective compression methods try to train a new model to evaluate the importance of tokens. [14] trains a model to learn the token compression given by the GPT4. [15] exploits reinforcement learning to train a model for deciding which tokens to remove, and [16] also uses reinforcement learning to train a model for pruning the prompt with the chain-of-thought style.

Distinguished from previous selective compression works, this work tries to transform the prompt compression problem of selecting tokens as the tree pruning problem. During such transformation, we consider linguistic rules, hierarchical structure of prompts, and human writing logic, while building the tree and updating the values of nodes for tree pruning.

C. Text Pattern Analysis and Compression

Text data contains the latent patterns that can be used to assist test analysis and processing. For example, [29] employs backbone information to help the transformer model to learn the text encoding and representation. Some explicit text patterns, e.g., the patterns given by linguistic parse trees, can be easily understood by humans and have been applied to many machine learning tasks. For instances, [17] employs the parse tree to aid natural language processing tasks like machine translation. Both [18] and [19] employ parse trees to enhance the reasoning ability of models. Note that, the most

related work to our work, [20] also utilizes the explicit text patterns, i.e., parse trees, to compress sentences.

Unlike the sentence compression, the prompt compression for LLMs has its own challenges. First, the prompt often consists of multiple sentences or even multiple paragraphs and sections, which involves much more abundant patterns. Second, the compressed output is not only aiming for human understanding but also for assisting LLMs in various tasks. Also note that, the patterns in LLM prompt, such as linguistic parse trees and the hierarchical structure of sentences, have not yet been explored in the prompt compression problem.

III. PROBLEM FORMULATION

Definition 1. Prompt $R \triangleq r_1r_2\cdots r_n$ is a sequence of tokens, which is the natural language input to LLMs for performing a specific task. For each token r_i in R , function $C(\cdot)$ calculates the length of a token, and function $E(\cdot)$ calculates the importance or value of a token. Given a prompt R containing n tokens, the length of the prompt can be obtained by $C(R) \triangleq \sum_{i=1}^n C(r_i)$, and the value of the prompt can be obtained by $E(R) \triangleq \sum_{i=1}^n E(r_i)$.

Definition 2. Selective Compression refers to the compressed prompt $R_{cp} \triangleq r_{w_1}r_{w_2}\cdots r_{w_l}$ being selected from a part of the original prompt $R = r_1r_2\cdots r_n$, where $\{w_1, w_2, \dots, w_l\}$ is a subset of $\{1, 2, \dots, n\}$ satisfying $w_1 < w_2 < \dots < w_l$, which indicates that the selective compression does not generate new tokens and preserves the order. Accordingly, the compression ratio between the compressed prompt and original prompt can be defined by

$$\tau \triangleq \frac{C(R_{cp})}{C(R)}. \quad (1)$$

Definition 3. Selective Prompt Compression Problem can be formulated as: for a given prompt R and compression ratio τ where $0 < \tau < 1$, find the compressed prompt R_{cp} , selecting from the original R , such that the overall important score or value of R_{cp} is maximized without exceeding the upper token limit set by τ of R during the compression process, i.e.,

$$\max E(R_{cp}) \text{ s.t. } C(R_{cp}) \leq \tau C(R). \quad (2)$$

Definition 4. Parse Tree $T \triangleq (R, V, v_{root}, f)$ consists of four terms, where R is the given prompt; $V = \{v_1, v_2, \dots, v_n\}$ is a node set for the parse tree wherein v_{root} is the root node without parent node; for each node $v_i \in V$, function $f : V \setminus \{v_{root}\} \mapsto V$ maps each node to its parent node except for the root node. Note that, each node v_i corresponds one-to-one

to a token r_i , and hence to obey the order of tokens in the prompt, the nodes in a parse tree also retains the same order. We define $v_{w_1} < v_{w_2}$ (here $<$ indicates that it is ordered earlier in the sequence among sibling nodes) for subscript $w_1 < w_2$. The node with the smallest index among child nodes is called the first child node. The length and value of a node can be defined by the corresponding token, i.e., $C(v_i) \triangleq C(r_i)$ and $E(v_i) \triangleq E(r_i)$ respectively.

Remarks: We expand the definition of the parse tree based on the foundation of traditional linguistics. If the given prompt contains only one sentence, the parse tree is exactly the traditional linguistic parse tree of the sentence. If the given prompt includes multiple sentences, the parse tree becomes the global parse tree, which integrates multiple linguistic (also referred as local) parse trees following some logic.

Definition 5. Subtree $T_j = (R_j, V_j, v_{root,j}, f_j)$ is a portion of a given parse tree $T = (R, V, v_{root}, f)$. The node $v_{root,j}$, a particular node in the node set V of given parse tree T , is the root node of the subtree. And it, along with all its child nodes, constitutes the set V_j .

Definition 6. Compressed Tree $T_{cp} \triangleq (R_{cp}, V_{cp}, v_{root,cp}, f_{cp})$ is the remaining part of the given parse tree $T = (R, V, v_{root}, f)$ after pruning some subtrees T_1, T_2, \dots, T_l . Let k -th pruned subtree be $T_k = (R_k, V_k, v_{root,k}, f_k)$ for $k \in \{1, 2, \dots, l\}$, then the deleted node set is $\bigcup_{k=1}^l V_k$, the retained node set is $V_{cp} = V \setminus (\bigcup_{k=1}^l V_k)$. Since all nodes in the retained node set are also the nodes in the given tree T , the compressed prompt R_{cp} is a selective compression of R .

IV. THE PROPOSED METHOD

This section elaborates the proposed method in detail, and the overall framework of proposed PartPrompt is illustrated in Figure 1. To be more specific, for a given prompt R , we first slice it into sentences $[R_1, R_2, \dots, R_m]$ and calculate the information entropy of each token within each sentence. A parse tree is then built based on linguistic rules for each sentence so that we obtain corresponding m local parse trees $[T_1, T_2, \dots, T_m]$. Second, a global parse tree T is constructed by using virtual nodes to merge these local parse trees based on the structure of prompt. Third, a node value adjustment module, including the root-ward propagation and leaf-ward propagation, is proposed to adjust the original value of each node based on human writing logic. Finally, a recursive algorithm is developed to prune the global tree, which yields the compressed tree T_{cp} and accordingly the compressed prompt R_{cp} for a given compress ratio τ .

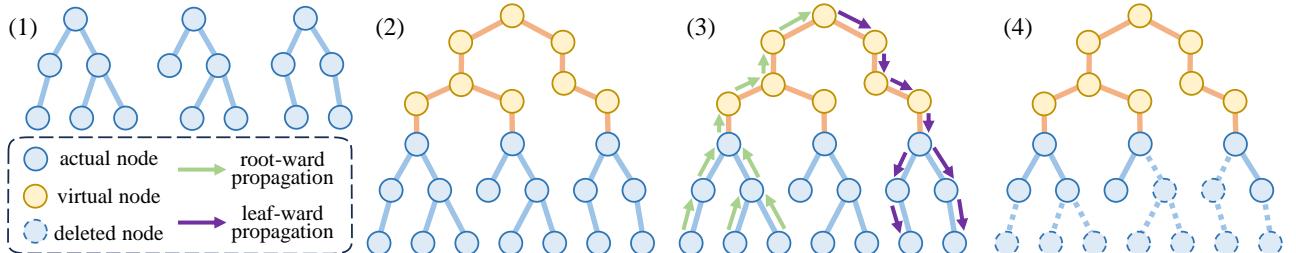


Fig. 1. Overall framework of PartPrompt: (1) local parse trees for sentences, (2) global parse tree construction, (3) node value adjustment, (4) tree compression.

A. Information Entropy Approximation

Given a prompt R that contains multiple sentences, we slice it into a list of sentences $[R_1, R_2, \dots, R_j, \dots, R_m]$. For each sentence R_j , it can further be divided into a sequence of tokens $r_{j,1} r_{j,2} \dots r_{j,i} \dots r_{j,n_j}$. The conditional probability of a token $r_{j,i}$ given its preceding token sequence can be computed by

$$p(r_{j,i} | r_{j,<i}, r_{<j}), \quad (3)$$

where the subscript j denotes sentence index and $< j$ indicates all the tokens prior to sentence j , while the subscript i denotes token index and $< i$ indicates all the tokens prior to the token i within the sentence j . The corresponding information entropy [30] of a token $r_{j,i}$ then becomes

$$E(r_{j,i}) = -\log p(r_{j,i} | r_{j,<i}, r_{<j}). \quad (4)$$

As shown in Equation (4), to obtain the information entropy of a token, we can compute its conditional probability, which is exactly what the language model does. Accordingly, the information entropy of a token $r_{j,i}$ estimated by language model can be calculated by

$$-\log p_{LM}(r_{j,i} | r_{j,<i}, r_{<j}). \quad (5)$$

In order to reduce the computational cost of calculating information entropy, we make the following approximation

$$p(r_{j,i} | r_{j,<i}, r_{<j}) \approx p(r_{j,i} | r_{j,<i}), \quad (6)$$

where the term $r_{<j}$ is removed, which means the information entropy of a token is only considered within its own corresponding sentence. Consequently, the information entropy calculated by this approximation finally becomes

$$E_{LM}(r_{j,i}) = -\log p_{LM}(r_{j,i} | r_{j,<i}). \quad (7)$$

Regarding the information entropy approximation in Equation (7), it is worth noting that, the computational cost is saved especially for a long prompt containing multiple sentences. Besides, to tackle the potential performance drop caused by the information entropy approximation and make the approximation more promising, we suggest to consider the global structure of a given prompt using a well-organized tree and adjust the approximated information entropy based on the tree, which are presented in the following sections.

B. Global Parse Tree Construction

A local parse tree, based on linguistic rules, is a tree that analyzes word dependencies for a given sentence. Concretely, in a local parse tree T_j corresponding to a sentence R_j , the word r_i , attached to the child node v_i , is considered dependent on the word of its parent node via mapping f_j on v_i . The word of global verb in this sentence is located at the root node $v_{root,j}$ without its parent dependency. We adopt Stanford NLP toolkit [31] to build all local parse trees $[T_1, T_2, \dots, T_m]$ for all sentence $[R_1, R_2, \dots, R_m]$.

For the series of local parse trees $T_j = (R_j, V_j, v_{root,j}, f_j)$ for $j \in \{1, 2, \dots, m\}$, we introduce a new virtual node \tilde{v} and make $v_{root,1}, v_{root,2}, \dots, v_{root,m}$ as its child nodes. In this way, these trees are aggregated into a single tree, which is referred

as the global parse tree, denoted by T . For the virtual node \tilde{v} introduced to aggregate trees, there is no actual token attached to the virtual node, and the initial value or information entropy is zero. In contrast, the nodes in local parse trees contain actual token(s) and are referred as actual nodes.

The aforementioned global tree connects all local parse trees into a global tree, but cannot reflect the hierarchical structure of the entire given prompt. Considering a common sentence-paragraph-section-document writing style, we suggest to carry out the aforementioned aggregation process in several steps. First, the virtual sentence nodes are created for each local parse tree corresponding to each sentence. Second, the virtual paragraph nodes are created to aggregate their subtree(s) belonging to the same paragraph. Third, the virtual section nodes are created to aggregate their subtree(s) belonging to the same section. Fourth, a virtual document node is created to aggregate all subtree(s) for the entire prompt.

The global parse tree T has dependencies between child nodes and parent nodes: actual nodes (from local parse trees) depend on the corresponding virtual sentence nodes; virtual sentence nodes depend on virtual paragraph nodes; virtual paragraph nodes depend on virtual section nodes; virtual section nodes depend on the virtual document node (i.e., the root node of global tree T) for the entire given prompt. This hierarchical dependency relationship clearly reflects the hierarchical structure of the entire prompt.

C. Token Alignment

For the actual nodes in the local parse trees analyzed by linguistic rules, there is a token attached to each actual node. However, the token in an actual node might not be one-to-one correspondence to the token analyzed by an LLM tokenizer, which is utilized in obtaining the value of an actual node to reflect its importance by a small well-trained LLM. Consequently, a token alignment module is needed to compute the value and length of the token in each actual node.

The aim of token alignment is to ensure the smallest token retaining sufficient semantic integrity. To this end, we choose the tokenizer of establishing local parse trees (denoted as parse tree tokenizer) as the base, and utilize a small well-trained LLM with LLM tokenizer to compute the information entropy of the token(s) attached to actual nodes in local parse trees. The alignment between the two tokenizers achieved through a rule-based matching algorithm, yielding the aligned information entropy or value $E_{Aligned}(v_i)$ and its length $C(v_i)$ for each node v_i in local parse trees.

Figure 2 illustrates a toy example. The token *Almaty* is recognized as a single unit by the parse tree tokenizer, whereas the LLM tokenizer divides it into three distinct tokens: *Al*, *mat*, and *y*. The information entropy E_{LM} computed by LLM for the three tokens are 6.69, 7.15, and 0.02. To maintain the semantic integrity, we adhere to make the parse tree tokenizer as the base, treating *Almaty* as a unified token. Accordingly, the aligned information entropy $E_{aligned}$ of this unified token is the sum of the information entropy of the three tokens by LLM tokenizer, i.e., $(6.69 + 7.15 + 0.02)$. And the length of this unified token $C(\text{Almaty})$ attached to the node is 3.

Sentence	<i>Almaty is the capital of Kazakhstan</i>
Parse tree tokenizer	<i>Almaty---is---the---capital---of---Kazakhstan</i>
LLM tokenizer	<i>Al---mat---y---is---the---capital---of---Kaz---akh---stan</i>
Token value E_{LM}	6.69 7.15 0.02 3.00 0.73 2.56 0.70 0.22 0.003 0.002
Token alignment	<i>Almaty---is---the---capital---of---Kazakhstan</i>
Token value $E_{Aligned}$	(6.69+7.15+0.02) 3.00 0.73 2.56 0.70 (0.22+0.003+0.002)
Token length C	3 1 1 1 1 3

Fig. 2. An example of token alignment.

D. Node Value Adjustment

Section IV-B have introduced a novel global tree composing of actual nodes and virtual nodes. The hierarchical structure of virtual nodes simulates the typical writing style of sentence-paragraph-section-document. The aim of virtual nodes is to harmonize the compression requests for the segments at the global scale, where each segment refers to a subtree rooted by a virtual node. For this purpose, two pivotal criteria have to consider. First, each virtual node should reflect the value of its corresponding segment. Second, the compression request for each segment should be able to propagate to its corresponding actual nodes as well as virtual nodes.

Given the aim and two criteria, a node value adjustment algorithm is developed to adjust the original values attached to the nodes of the global parse tree. It consists of two components: root-ward propagation and leaf-ward propagation. The pseudocode is presented in Algorithm 1, where Lines 1-11 corresponds to the root-ward propagation and Lines 12-20 corresponds to the leaf-ward propagation.

Root-ward propagation tries to update the value of each virtual node (see Lines 5 and 7) such that the value of a virtual node reflects the value of its corresponding segment. Inspired from the forward and backward propagation in neural networks, we employ a momentum-based approach to aggregate node values from leaf nodes to the root node of a segment, thereby assigning the averaged value to the corresponding virtual node. Note that, the value of actual nodes are not updated and we append the aggregated value to a list or vector \vec{M} , whereas the value of virtual nodes are updated recursively and we also append the aggregated value to \vec{M} .

Leaf-ward propagation attempts to update the value of each actual node (see Lines 17 and 18) based on its original value and the compression request. The compression request is propagated recursively from the root node to leaf nodes. For a virtual node, we employ a scalar M (at very beginning $M = 1$) to cache its value, and we adjust M with a hyper-parameter a_2 if it is the first virtual node at a hierarchy of the global tree. For an actual node, its adjusted value is obtained by adding its original value (i.e., the aligned information entropy) and the cached M with an experiential adjustment hyper-parameter a_1 . Note that, a_1 ensures the adjusted value retaining a notable distinction. And a_2 targets at emphasizing the value of the first part at each hierarchy of the global tree, such as the first sentence in a paragraph and the first paragraph in a section.

Algorithm 1: Node Value Adjustment

```

Input: Global Parse Tree  $T = (R, V, v_{root}, f)$ , Aligned
Information Entropy  $E_{Aligned}$ , Hyper-parameters
for Node Value Adjustment  $a_1, a_2$ 
Output: Adjusted Node Value  $E_{Adjusted}$ 

Function RootwardPropagation( $v_i$ ):
1    $\vec{M} \leftarrow$  Empty list []
2   foreach  $v_k \in \{v \in V \mid f(v) = v_i\}$  do
3      $\vec{M}$  appends RootwardPropagation( $v_k$ )
4
5   if  $v_i$  is virtual node then
6      $N \leftarrow$  Average( $\vec{M}$ )
7      $E_{Adjusted}(v_i) \leftarrow N$ 
8   else
9      $\vec{M}$  appends  $E_{Aligned}(v_i)$ 
10     $N \leftarrow$  Average( $\vec{M}$ )
11
12 return  $N$ 

Function LeafwardPropagation( $v_i, M$ ):
13  if  $v_i$  is virtual node then
14     $M \leftarrow M \cdot E_{Adjusted}(v_i)$ 
15    if  $v_i$  is first child node then
16       $M \leftarrow M \cdot a_2$ 
17
18  else
19     $E_{Adjusted}(v_i) \leftarrow E_{Aligned}(v_i) + M^{a_1}$ 
20    foreach  $v_k \in \{v \in V \mid f(v) = v_i\}$  do
21      LeafwardPropagation( $v_k, M$ )
22
23 RootwardPropagation( $v_{root}$ )
24 LeafwardPropagation( $v_{root}, 1$ )

```

E. Tree Compression Based on Node Value

After obtaining the global parse tree as well as the length and the value of each node, the selective prompt compression problem, as defined in Definition 3, can be transformed into a tree pruning problem that considers the length and the value of the nodes. Formally, it can be formulated as follows.

Definition 7. Parse Tree Pruning Problem Given a tree $T = (R, V, v_{root}, f)$ for a specified prompt R and a specified compression ratio τ where $0 < \tau < 1$, the task is to derive a compressed tree $T_{cp} = (R_{cp}, V_{cp}, v_{root, cp}, f_{cp})$ for a selective compression R_{cp} . The compressed tree is the optimal solution of the following optimization problem

$$\max E(V_{cp}) \quad (8)$$

$$\text{s.t. } C(V_{cp}) \leq \tau C(V), \quad (9)$$

$$f_{cp} : V_{cp} \setminus \{v_{root, cp}\} \mapsto V_{cp}. \quad (10)$$

The objective function (8) and constraint (9) are derived from the general selective prompt compression problem. The constraint (9), which corresponds to the constraint $C(R_{cp}) \leq \tau C(R)$ in the general problem, ensures that the actual compression ratio $C(V_{cp}) / C(V)$ does not exceed the given compression ratio τ , thereby allowing for precise control over the length of the compressed prompt. The objective function (8), which corresponds to objective function $\max E(R_{cp})$ in

the general problem, aims to preserve the value of the nodes as much as possible.

The constraint (10) guarantees that the retained nodes maintain the tree structure, a requirement derived from the property of the parse tree. This reflects the inherent dependencies within the parse tree, whereby the tokens on the child nodes are considered dependent on the tokens of the parent nodes. Consequently, if the compressed prompt includes tokens from the child nodes, the corresponding parent node tokens should also be preserved.

A recursive algorithm is then developed to derive the optimal solution. The pseudocode is provided in Algorithm 2. The output of algorithm is a list, denoted by Q , which is defined as below: $Q = [Q_0, Q_1, Q_2, \dots]$, where each Q_l represents the optimal solution limited to length l . To be more specific, $Q_l = (Q_{l,1}, Q_{l,2}) = (E(V_{cp,l}), V_{cp,l})$, with $V_{cp,l}$ being the node set of the optimal compressed tree limited to length l . Thus, the tree constructed by node set $Q_{\lfloor \tau C(V) \rfloor,2}$ is the optimal solution of the parse tree pruning problem in Definition 7. After that, the compressed prompt is obtained by concatenating the tokens corresponding to the aforementioned nodes in sequence. In Algorithm 2, each call of function `CalculateSolution` recursively derives the optimal solution for the subtree rooted at v_i from the optimal solutions of the subtrees rooted at all child nodes of v_i . This process is divided into two steps: First merging the solutions from the child nodes using the function `MergeSolution` (Lines 11-13), and second incorporating v_i into the merged solution (Lines 14-18). As both of these steps maintain the optimal state, the final result is also optimal.

F. Algorithm and Complexity Analysis of PartPrompt

The complete procedure of the PartPrompt method is presented in Algorithm 3. For clarity, Section IV-A describes Line 3, where function `SmallModel` calculates the entropy of token by Equation (7); Section IV-B describes Line 4 and Line 6, where function `SentenceParser` builds the local parse tree and function `BuildGlobalTree` builds the hierarchical global tree; Section IV-C describes Line 5; Section IV-D describes Line 7; and Section IV-E describes Line 8. Moreover, function `PromptStructureParser` slices the original prompt into sentences, and function `TokenConcatenation` concatenates tokens into compressed prompts. The innovative recursive tree compression technique allows for the concurrent management of multiple compression ratios without additional computational costs. Users are thus able to compare the results of several compression ratios before selecting one.

To analyze the computational complexity of each component within the process described in Algorithm 3, consider a prompt composed of m sentences, with each sentence containing n tokens. The computational complexity of Algorithm 3 is broken down as follows: Line 1, Line 6 requires $\mathcal{O}(m)$; Line 3 requires $\mathcal{O}(kmn^2)$, where k is related to the size of language model; Line 4 requires $\mathcal{O}(mn^3)$; Line 5, Line 7, and Line 9 requires $\mathcal{O}(mn)$; Line 8 requires $\mathcal{O}(m^2n^2)$. In comparison, the computational complexity of Selective-Context [12] is $\mathcal{O}(km^2n^2)$, while LLMLingua [13] requires

Algorithm 2: Recursive Tree Compression

```

Input: Parse Tree  $T = (R, V, v_{root}, f)$ , Length Function  $C$ , Adjusted Value  $E$ 
Output: List  $Q = [Q_0, \dots, Q_l, \dots]$  where  $Q_l$  is the optimal solution limited to length  $l$ 

1 Definition For a list  $Q$ ,  $C(Q)$  is defined as the length of  $Q$ , e.g.  $C([Q_0, Q_1, \dots, Q_n]) = n + 1$ 
2 Function MergeSolution( $Q, P$ ):
3   foreach  $0 \leq l \leq (C(Q) + C(P) - 2)$  do
4      $k \leftarrow \underset{j}{\operatorname{argmax}}(Q_{j,1} + P_{l-j,1})$ 
5      $S_l \leftarrow ((Q_{k,1} + P_{l-k,1}), (Q_{k,2} \cup P_{l-k,2}))$ 
6      $S \leftarrow [S_0, \dots, S_{C(Q)+C(P)-2}]$ 
7   return  $S$ 
8 Function CalculateSolution( $v_i$ ):
9    $Q_0 \leftarrow (0, \emptyset)$ 
10   $Q \leftarrow [Q_0]$ 
11  foreach  $v_k \in \{v \in V \mid f(v) = v_i\}$  do
12     $P \leftarrow \text{CalculateSolution}(v_k)$ 
13     $Q \leftarrow \text{MergeSolution}(Q, P)$ 
14  foreach  $0 \leq l \leq (C(Q) + C(v_i) - 1)$  do
15    if  $l < C(v_i)$  then
16       $S_l = (0, \emptyset)$ 
17    else
18       $S_l = (E(v_j) + Q_{l-C(v_i),1}, \{v_j\} \cup Q_{l-C(v_i),2})$ 
19     $S \leftarrow [S_0, \dots, S_{C(Q)+C(P)-1}]$ 
20  return  $S$ 
21 CalculateSolution( $v_{root}$ )

```

$\mathcal{O}(kmn^2(c^2m+1))$, with c representing the compression ratio. Under typical conditions, m and n range around $20 \sim 50$, c lies between $0.2 \sim 0.5$, and $k \gg m, n$. Therefore, the overall computational complexity of the PartPrompt method is $\mathcal{O}(kmn^2)$, resulting in the following leaf-ward propagation of computational demands among these three methods as $\mathcal{O}(\text{PartPrompt}) < \mathcal{O}(\text{LLMLingua}) < \mathcal{O}(\text{Selective-Context})$.

G. Theoretical Analysis

This section elaborates the theoretical analysis of information entropy approximation and the theoretical connections of PartPrompt to the two prominent selective prompt compression methods: Selective-Context [12] and LLMLingua [13].

In the information entropy approximation module, the term $r_{<j}$ is excluded when calculating $E(r_{j,i})$. This removed term is same for all tokens $r_{j,1}, r_{j,2}, \dots, r_{j,n_j}$ within the same sentence, indicating a roughly consistent approximation error across these tokens. However, for tokens across different sentences, those with a larger j have more $r_{<j}$ omitted, leading to a greater neglect of information and a potential overestimation of $E(r_{j,i})$. Considering that the objective Function (8) of tree compression aims to preserve nodes with larger value, the approximation error leads to a bias towards retaining tokens with a larger j . Therefore, this inter-sentence error need to be managed efficiently, or may results in performance loss.

Algorithm 3: PartPrompt

Input: Original Prompt R , Compression Ratio List $[c_1, c_2, \dots, c_u]$, Node Value Adjustment Parameters a_1, a_2
Output: Compressed Prompts List of each Compression Ratio $[R_{cp,1}, R_{cp,2}, \dots, R_{cp,u}]$

- 1 Sentence List $[R_1, R_2, \dots, R_m]$, Structure Label List $[g_1, g_2, \dots, g_m] \leftarrow \text{PromptStructureParser}(R)$
- 2 **foreach** Sentence R_j in Sentence List **do**
- 3 Token List $[s_1, s_2, \dots, s_{l_j}]_j$, Entropy List $[e_1, e_2, \dots, e_{l_j}]_j \leftarrow \text{SmallModel}(R_j)$
- 4 Node List $[r_1, r_2, \dots, r_{n_j}]_j$, Edge List $[(r_{w_1}, r_{w_2}), (r_{w_2}, r_{w_3}), \dots, (r_{w_{h_j}}, r_{w_{n_j}})]_j \leftarrow \text{SentenceParser}(R_j)$
- 5 Aligned Entropy List $[\tilde{e}_1, \tilde{e}_2, \dots, \tilde{e}_{n_j}]_j$, Length List $[C_1, C_2, \dots, C_{n_j}]_j \leftarrow \text{TokenAlignment}(\text{Token List}, \text{Node List}, \text{Entropy List})$
- 6 Global tree $T \leftarrow \text{BuildGlobalTree}(\text{Edge Lists}, \text{Structure Label List}, \text{Node Lists})$
- 7 Adjusted Node Value Lists $[\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n] \leftarrow \text{nodeValueAdjustment}(T, \text{Aligned Entropy Lists}, a_1, a_2)$
- 8 Solution $Q \leftarrow \text{RecursiveTreeCompression}(T, \text{Length Lists}, \text{Adjusted Node Value Lists})$
- 9 Compressed Prompts List $[R_{cp,1}, R_{cp,2}, \dots, R_{cp,u}] \leftarrow \text{TokenConcatenation}(\text{Solution}, \text{Compression Ratio List})$

Next, we present the theoretical connections of PartPrompt to the prominent methods. Selective-Context retains tokens with higher information entropy and employs a parse-based tokenizer to ensure token completeness [12]. Accordingly, Selective-Context can be reformulated within our parse tree compression framework as follows. For a given prompt R , a tree $T = (R, V, v_{\text{root}}, f)$ is constructed, where v_{root} is the only virtual node, and all actual nodes v_i are its child nodes with $f(v_i) = v_{\text{root}}$. The tree pruning problem in Definition 7 is then solved with $E(v_i)$ calculated by Equation (12). Therefore, Selective-Context can be considered as a simplified version of PartPrompt that flattens the global parse tree and omits the approximation of information entropy.

LLMLingua, another prominent selective prompt compression method, consists of the budget controller, iterative token-level prompt compression, and distribution alignment [13]. The budget controller divides the prompt into paragraphs, sorting them by information entropy and preserving those with higher entropy. In PartPrompt, the node value adjustment algorithm for actual nodes is defined as $E_{\text{Adjusted}}(v_i) \leftarrow E_{\text{Aligned}}(v_i) + M^{a_1}$ (Line 18 of Algorithm 1). Here M is derived from the root-ward propagation and convergence of the information entropy of the corresponding segments. In light of the fact that $M > 1$ is typically the case, setting $a_1 \gg 1$ gives rise to the conclusion that $M^{a_1} \gg E_{\text{Aligned}}(v_i)$, which leads to the approximation that $E_{\text{Adjusted}}(v_i) \approx M^{a_1}$. In this case, parse tree compression is equivalent to retaining the part with the larger value of M . Therefore, the budget controller of LLMLingua can be considered as a variant of parse tree compression which involves flattening the virtual part of the global parse tree and setting parameter $a_1 \gg 1$ in the node value adjustment module of PartPrompt.

Considering $r_{j,i}$ as the i -th token of the j -th sentence, the iterative token-level prompt compression of LLMLingua employs Equation (11) to calculate the information entropy. This approach may be regarded as an intermediate stage between the initial version employed by Selective-Context, i.e., Equation (12), and the local approximation version of PartPrompt, i.e., Equation (13).

$$E(r_{j,i}) = -\log p_{\text{LM}}(r_{j,i} | r_{j,<i}, r_{<j,\text{retained}}). \quad (11)$$

$$E(r_{j,i}) = -\log p_{\text{LM}}(r_{j,i} | r_{j,<i}, r_{<j}). \quad (12)$$

$$E(r_{j,i}) = -\log p_{\text{LM}}(r_{j,i} | r_{j,<i}). \quad (13)$$

It might be worth mentioning that the approximation from Equation (12) to Equation (13) does not lead to a significant performance loss, when the approximation error is effectively managed as demonstrated by the empirical study in Section VI-D. On the other hand, the approximation significantly reduces the computational cost, since the entropy is computed within each sentence rather than the whole prompt.

V. EXPERIMENTAL SETTINGS

A. Datasets

Four representative datasets are employed to evaluate the proposed method. Concretely, BBCnews dataset is collected from articles on the BBC News website [32], and arXiv dataset is collected from papers on the arXiv preprint platform [33]. Following Selective-Context [12] and LLMLingua [13], these two datasets are used for testing contextual understanding prompts with the task of summarizing articles. To ensure that the model has not seen these data during training, we re-crawl new data for BBC News and arXiv, with all the release dates of these data occurring after Jan 1, 2024.

Due to the length of arXiv dataset, which exceeds the input length that Selective-Context and LLMLingua can handle [12], [13], we truncate the first 3000 tokens of the main text, referring as truncate arXiv. The original arXiv dataset is also used to verify the applicability of our method for texts exceeding the input limit of compared methods.

Regarding another two datasets, HotpotQA [34] is a multi-hop question answering dataset, which is designed to test the performance for question answering and multi-hop reasoning prompts. We use the same setting as in [35] and select 500 questions for testing. GSM8K [36] is a classic mathematical reasoning dataset. Following [13], we adopt the chain-of-thought prompt provided by [7] to evaluate the performance for in-context learning and chain-of-thought prompts.

B. Compared Methods for Prompt Compression

As our method belongs to selective prompt compression, we compare it to the state-of-the-art selective prompt compression methods. Selective-context [12] is the first to propose selective

prompt compression method. It uses a smaller LLM to calculate the information entropy of tokens and removes words with lower information entropy. The question text is fully preserved for better performance. LLMLingua [13] further proposes the model distribution alignment, budget controller and token-level iterative compression algorithm based on Selecting-context. LLMLingua receives significant improvements in context-understanding prompts, and makes selective compression method viable for chain-of-thought prompts.

In addition to selective compression methods, our method is also compared with the generative compression method. Qwen2-72B [37], the up-to-date generative LLM trained by Alibaba, is exploited to directly compress the original prompt. To be more specific, we feed the original prompt into Qwen2-72B together with the prompt of asking LLM for compression¹ to generate the compressed prompt. To further highlight the advantage of our method, we compare it with the generative compression method over the four normal datasets, despite using LLMs like Qwen2-72B for compression requires huge computational resources. Besides, we additionally conduct experiments for a very long prompt scenario that most selective compression methods are unable to manage.

C. Target Models for Inference

The prompt is employed to assist the inference and generation of target LLMs. In fact, the same prompt for different target LLMs would have different performances. To verify the effectiveness of our method across various target LLMs, we conduct experiments for the following LLMs: Mixtral-8x7B [38], Llama3-70B [2], and Qwen2-72B [37]. Among them, Mixtral-8x7B is a famous mix-of-expert LLM performing well on various tasks, and is set as the default model unless specified otherwise. Llama3-70B is a recent high-performance LLM trained by Meta; Qwen2-72B is the up-to-date LLM trained by Alibaba and receives the higher performance than Llama3-70B according to the HELM Leaderboard².

D. Evaluation

Following [12], [13], we take BLEU [39], Rouge (including Rouge1, Rouge2, RougeL) [40], and BERTScore (specifically BS-F1) [41] as the metric for evaluation on BBCnews and (truncate) arXiv datasets. BLEU is a composite metric of four metrics: 1-gram, 2-gram, 3-gram, and 4-gram, and is often used in the natural language processing. Rouge is classic linguistic metrics based on rule matching, while BERTScore calculates the similarity of the embeddings from BERT. Regarding above two datasets, the ground-truth answer is set to the output of the uncompressed prompt. For hotpotQA dataset, we follow [35] to compute precision, recall, and F1, since the answers is quite brief. For GSM8K dataset, as the questions are math-related, we adopt exact matching to calculate the EM score obeying the same setting as in [7], [13].

¹The prompt of asking LLM for compression is as follows: "Eliminate repetitive elements and present the text concisely, ensuring that key details and logical processes are retained.", which comes from [13].

²HELM Leaderboard: <https://crfm.stanford.edu/helm/mmlu/v1.6.0>

E. Other Settings

Apart from above settings, other experimental settings are clarified as follows. The information entropy is calculated using Llama2-7B [42], and the local parse tree is analyzed using the Stanford CoreNLP [31]. When adjusting the value in parse trees, the related hyper-parameters are set as follows: for BBCnews, $a_1 = 4$, $a_2 = 100$; for truncate arXiv, $a_1 = 3$, $a_2 = 100$; for hotpotQA, $a_1 = 3$, $a_2 = 20$; for GSM8K, $a_1 = 5$, $a_2 = 25$. And the typical range of them are $0 \leq a_1 \leq 5$, $1 \leq a_2 \leq 1000$. The maximum token length for target LLMs is set to 300 during inference for all experiments. For the pretrained language models used in this work, Mixtral-8x7B and Llama3-70B are accessed through HuggingFace Pro API³; Qwen2-72B is accessed via Aliyun or can be downloaded from HuggingFace⁴; Llama2-7B is downloaded from HuggingFace⁵ and run on an NVIDIA GeForce RTX 3090 GPU with 24G memory.

VI. EXPERIMENTS

Extensive experiments are conducted to demonstrate the effectiveness and superiority of PartPrompt (the proposed method) regarding various prompt compression methods, datasets, metrics, compression ratios, and scenarios. Section VI-A presents a thorough comparison between PartPrompt and other compression methods to confirm the superiority of the proposed method. Considering the effect of compression ratios and target LLMs for inference, PartPrompt is further compared with other methods for various compression ratios in Section VI-B and different target LLMs in Section VI-C. Section VI-D provides a comprehensive ablation study to verify the positive role of each component in PartPrompt. Section VI-E investigates the potential of PartPrompt under the extreme long prompt scenario. In Section VI-F and VI-G, the compressed prompt is directly compared with uncompressed one; accordingly more abundant metrics and an intuitive example are employed to show the advantages of PartPrompt in terms of text similarity and coherence maintenance.

A. Comparative Study: Main Experiments

Thorough experiments are conducted in this section to fully compare the proposed method with other selective compression methods. Four diverse datasets and multiple metrics are employed to evaluate the performance. For clarity, we choose three compression ratios for analysis: 20%, 30%, and 50%. As the actual compression ratio is not under the control of some methods, the average length of the compressed prompt (named as "Tokens") and inverse compression ratio ($1/\tau$) are also listed to show the actual compression result. The LLM Generation is presented with a single compression ratio, since there is no compression ratio to set.

The experimental results are shown in Table I. Overall, it is evident that PartPrompt achieves remarkable improvements in almost all datasets, compression ratios, and metrics, which

³Mixtral-8x7B and Llama3-70B: <https://huggingface.co/blog/inference-pro>

⁴Qwen2-72B: <https://huggingface.co/Qwen/Qwen2-72B-Instruct>

⁵Llama2-7B: <https://huggingface.co/meta-llama/Llama-2-7b>

TABLE I

THE PERFORMANCE OF PARTPROMPT IN COMPARISON WITH BASELINE METHODS ON MULTIPLE DATASETS, COMPRESSION RATIOS, AND METRICS. THE BEST PERFORMANCE AMONG THREE SELECTIVE METHODS IS IN BOLD. PARTPROMPT ACHIEVES THE BEST PERFORMANCE IN ALMOST ALL SCENARIOS. THE BASELINE OF LLM GENERATION IS ONLY AS A REFERENCE, AS USING GENERATIVE LLMs FOR PROMPT COMPRESSION IS NOT PRACTICAL.

methods	BBCnews							truncate arXiv						
	BLEU	Rouge1	Rouge2	RougeL	BS-F1	Tokens	$1/\tau$	BLEU	Rouge1	Rouge2	RougeL	BS-F1	Tokens	$1/\tau$
LLM Generation	11.31	42.32	18.60	39.22	69.17	261.3	3.29	8.77	39.18	16.00	36.22	68.21	643.8	4.65
20% token constraint														
Selective-Context	3.58	31.63	9.22	28.73	60.64	188.6	4.56	5.28	34.86	12.25	31.48	65.10	656.6	4.56
LLMLingua	9.56	35.09	15.28	32.89	61.97	206.0	4.18	7.35	36.10	14.15	33.18	65.03	576.7	5.19
PartPrompt	13.69	43.54	22.87	41.05	67.58	173.7	4.96	11.16	39.17	17.96	36.36	66.99	556.6	5.38
30% token constraint														
Selective-Context	6.68	37.49	13.54	34.15	64.32	294.2	2.93	7.65	38.14	14.92	34.65	66.64	1048.1	2.86
LLMLingua	10.22	38.55	16.66	35.88	64.64	278.2	3.09	10.39	39.23	17.47	36.26	66.95	867.9	3.45
PartPrompt	18.38	48.57	28.39	46.21	70.20	257.7	3.34	15.35	43.12	22.50	40.26	68.95	839.4	3.57
50% token constraint														
Selective-Context	14.26	46.74	22.63	43.41	69.53	481.0	1.79	13.63	44.56	21.78	41.32	69.95	1763.7	1.70
LLMLingua	15.71	45.98	23.21	43.17	69.52	435.8	1.97	17.26	45.85	25.00	43.08	70.46	1445.9	2.07
PartPrompt	28.28	56.48	38.30	54.29	74.57	424.9	2.03	19.50	46.78	26.87	44.26	70.93	1405.8	2.13
hotpotQA														
methods	BLEU	Rouge1	Rouge2	RougeL	BS-F1	precision	recall	F1	Tokens	$1/\tau$	EM	Tokens	$1/\tau$	
LLM Generation	1.54	31.02	16.58	30.89	51.29	16.51	57.89	22.12	610.5	2.55	38.51	477.5	1.89	
20% token constraint														
Selective-Context	0.43	20.74	10.47	20.62	45.63	9.92	37.51	13.79	309.7	5.02	0.00	261.4	3.45	
LLMLingua	0.97	21.12	11.44	21.06	47.28	10.42	37.00	14.26	309.5	5.02	3.03	155.2	5.81	
PartPrompt	1.15	24.62	13.04	24.57	48.68	12.61	42.31	17.13	303.9	5.11	23.88	179.4	5.01	
30% token constraint														
Selective-Context	0.73	21.15	11.26	21.05	46.48	10.70	39.77	14.63	486.9	3.19	0.23	345.4	2.61	
LLMLingua	0.65	21.24	10.57	21.15	46.21	10.56	37.63	14.31	453.2	3.43	67.85	298.2	3.02	
PartPrompt	1.96	28.18	15.95	28.16	50.17	14.88	48.04	19.91	456.8	3.40	53.30	270.4	3.33	
50% token constraint														
Selective-Context	1.25	25.63	12.75	25.60	48.76	13.25	47.39	17.78	834.3	1.86	0.08	524.4	1.72	
LLMLingua	0.87	21.99	10.97	21.86	46.88	10.70	40.83	14.74	764.4	2.03	12.74	425.5	2.12	
PartPrompt	1.51	27.89	14.43	27.85	49.90	14.05	47.95	19.10	770.4	2.02	55.27	446.4	2.01	

demonstrates the advantages of PartPrompt. In particular, the proposed PartPrompt receives the best performance compared to Selective-Context and LLMLingua, i.e., the two state-of-the-art selective prompt compression methods, on BBCnews, truncated arXiv, and hotpotQA datasets across all metrics and compression ratios. For the GSM8K dataset, PartPrompt is more robust than LLMLingua under different compression ratios and achieves the best averaged performance regarding the three compression ratios.

Apart from comparing to selective compression methods, PartPrompt is further compared to the generative compression method. Qwen2-72B is employed for generating compressed prompts, which is denoted as LLM Generation in Table I. Note that, using an LLM for compression needs a significant amount of computational resources, which is contrary to the original goal of prompt compression and therefore impractical. Despite that, LLM Generation may still serve as a reference method. The results show that PartPrompt outperforms LLM Generation on the BBCnews, truncated arXiv, and GSM8K datasets. This not only proves the advantage of PartPrompt but also validates the rationality of selective compression methods. For hotpotQA, LLM Generation obtains better performance than all selective compression methods, which may indicate the limitation of current selective compression methods in handling this kind of datasets with incohesive texts.

B. Comparative Study: Under Various Compression Ratios

The compression ratio is a vital aspect in the prompt compression problem. To further investigate the effectiveness of these methods across a range of compression ratios, we carry out the experiments with nine different compression ratios, i.e., $\tau = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$. The results for BBCnews, truncated arXiv, and hotpotQA are plotted in Figure 3. The horizontal axis of each subplot depicts the compression ratio τ and the vertical axis corresponds to the metric score.

It can be observed that PartPrompt achieves considerable improvements for almost all compression ratios on these datasets, which again proves the superiority of the proposed method. Notably, the performance curve of PartPrompt displays a more pronounced convexity towards the upper left. This indicates that as the compression ratio decreases (i.e., the shorter compressed prompt), PartPrompt can better prioritize the deletion of less important parts, thereby slowing down the performance decline. It is worth mentioning the low to medium compression ratios are more challenging cases than the high compression ratios. And comparing to other methods, the proposed PartPrompt tends to receive more performance gains under the low and medium compression ratios. Also note that, the fluctuations on hotpotQA can be attributed to the brevity of the standard answers.

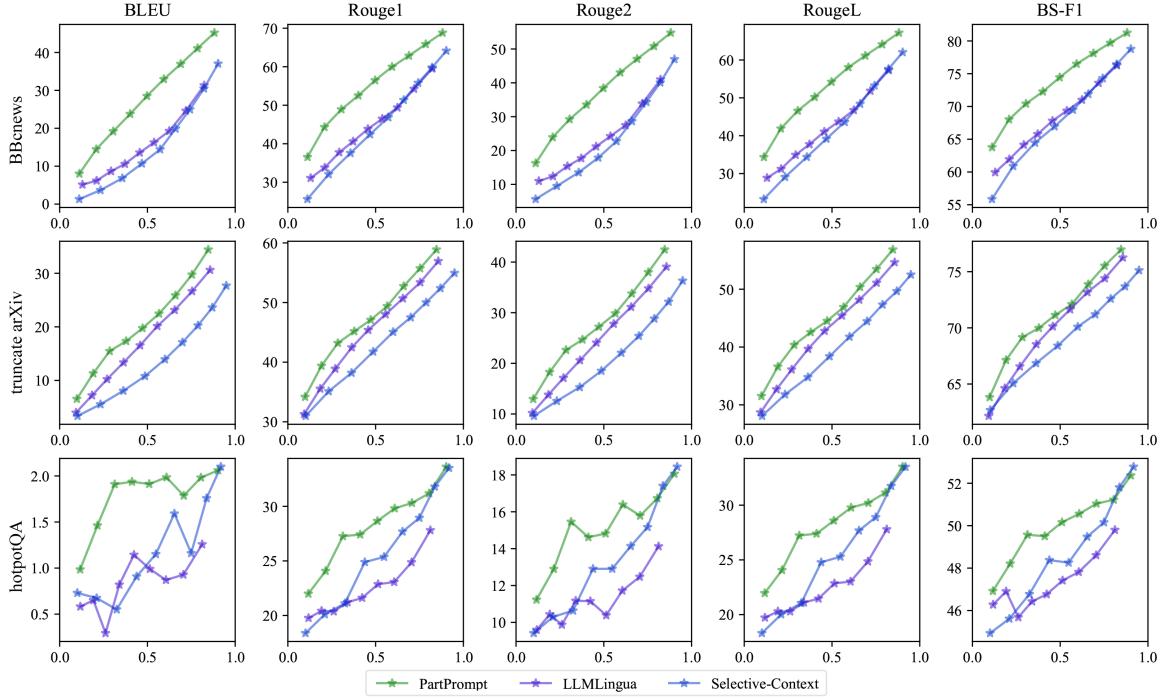


Fig. 3. The performance of selective prompt compression methods over various compression ratios. The horizontal axis of each subplot depicts the ratios.

It is interesting to observe the similarity in the curves between BLEU and Rouge2, and between Rouge1 and RougeL. The similarity can be owing to the underlying principles of the metrics. BLEU is derived from the average of its 1-gram to 4-gram, which could be considered as 2.5-gram. This is similar to Rouge2, which also employs 2-gram, and both count only continuous matching words. Conversely, Rouge1 and RougeL account for non-continuous matching words in their computations and exhibit greater similarity to each other. The consistency of the underlying principles and results in turn renders the reliability of our experimental results.

C. Comparative Study: Using Different LLMs for Inference

The Prompt is designed to be used with an LLM, and the same prompt with different LLMs would have different performances. Consequently, it is necessary to evaluate the performance for different LLMs. In addition to Mixtral-8x7B as the default LLM, two additional LLMs are also employed in this section: Llama3-70B and Qwen2-72B. The experiments are conducted on BBCnews, and the ground truth is set as the output of the corresponding LLM given the uncompressed prompt. The results are presented in Table II.

First, we observe that PartPrompt not only consistently outperforms baselines on Mixtral-8x7B but also consistently achieves better performances on Llama3-70B and Qwen2-72B. The observation substantiates the superiority and robustness of PartPrompt across different LLMs during inference. Second, it is worth mentioning that Selective-Context and LLMLingua yield quite low performances on Llama3-70B, which is caused by a large number of empty responses from Llama3-70B, a phenomenon not observed while feeding the compressed prompts given by PartPrompt to Llama3-70B.

D. Ablation Study of PartPrompt

To elucidate the effect of each PartPrompt component, a comprehensive ablation study is carried out. A total of nine variants of PartPrompt are designed and evaluated, and their organization is illustrated by a tree in Figure 4.

The complete PartPrompt is used as a point of departure for the sequential removal of components, so as to observe the effect of each component. Initially, the token value adjustment module is omitted, resulting in the hierarchical global tree degenerating into a global tree with a single virtual node. This modification is referred variant ①. Subsequently, the global parse tree is removed, keeping only the local parse trees, which is denoted as variant ②. In this instance, each local parse tree is pruned independently and then merged. Finally, the local parse trees are removed, with only information entropy used for prompt compression, which is denoted as variant ③. Note that, the complete PartPrompt and its variants ①, ②, and ③ all employ local approximated information entropy. When the information entropy approximation module is further removed, the complete PartPrompt and the corresponding variants ①, ②, ③ become ④, ⑤, ⑥, ⑦ respectively. These four variants can also be considered as an ablation chain of the global tree, and the performance difference among them can also reflect the contributions of each component of global tree.

On the other hand, the variants ⑧ and ⑨ are developed to investigate the impact of entirely removing the information entropy module and solely utilizing the parse tree for compression. Regarding a compression ratio of 0.5 on BBCnews, the performance of the complete PartPrompt, all nine variants, as well as two baseline methods is shown in Table III.

According to Table III, the complete PartPrompt achieves the best performance among all variants, demonstrating the

TABLE II

THE PERFORMANCE OF SELECTIVE PROMPT COMPRESSION METHODS BY FEEDING THE COMPRESSED PROMPT TO DIFFERENT LLMs.

methods	BLEU	Rouge1	Rouge2	RougeL	BS-F1
Mixtral-8x7B					
20% token constraint					
Selective-Context	3.58	31.63	9.22	28.73	60.64
LLMLingua	9.56	35.09	15.28	32.89	61.97
PartPrompt	13.69	43.54	22.87	41.05	67.58
30% token constraint					
Selective-Context	6.68	37.49	13.54	34.15	64.32
LLMLingua	10.22	38.55	16.66	35.88	64.64
PartPrompt	18.38	48.57	28.39	46.21	70.20
50% token constraint					
Selective-Context	14.26	46.74	22.63	43.41	69.53
LLMLingua	15.71	45.98	23.21	43.17	69.52
PartPrompt	28.28	56.48	38.30	54.29	74.57
Llama3-70B					
20% token constraint					
Selective-Context	1.74	12.04	3.85	11.01	22.16
LLMLingua	2.64	6.75	3.50	6.31	11.45
PartPrompt	13.64	41.71	22.07	39.35	65.88
30% token constraint					
Selective-Context	1.88	9.11	3.48	8.26	15.09
LLMLingua	4.06	13.79	6.07	12.82	23.77
PartPrompt	18.26	45.52	26.58	43.09	67.04
50% token constraint					
Selective-Context	1.28	4.06	1.97	3.79	6.16
LLMLingua	8.19	21.25	11.21	19.99	31.92
PartPrompt	25.87	51.80	33.67	49.55	70.87
Qwen2-72B					
20% token constraint					
Selective-Context	2.23	27.34	6.54	24.16	57.95
LLMLingua	6.40	28.99	10.57	26.47	58.43
PartPrompt	9.27	37.29	15.60	34.13	65.56
30% token constraint					
Selective-Context	4.97	33.14	10.21	29.66	62.68
LLMLingua	6.91	32.00	11.78	29.12	61.10
PartPrompt	13.51	41.84	20.03	38.55	68.70
50% token constraint					
Selective-Context	11.72	42.88	18.55	39.08	69.01
LLMLingua	11.29	40.06	17.35	36.83	67.92
PartPrompt	19.80	48.34	26.60	45.21	72.59

effectiveness of each component. Notably, variants ⑧ and ⑨, which rely solely on the parse tree, also achieve commendable results. The variant ⑧ even outperforms two state-of-the-art selective methods: Selective-Context and LLMLingua. These observations imply that the parse tree, analyzed by linguistic rules, holds valuable information for prompt compression. Furthermore, the performance of variants ④, ⑤, ⑥, and ⑦ decreases as the corresponding component removes, which also reflects the positive contributions of each component.

Figure 4 annotates the rise and drop of BS-F1 scores in the ablation study. First, it is evident that the token value adjustment yields a significant performance improvement (3.78%), which verifies the significance of considering human writing logic in prompt compression. Second, variants ⑧ and ⑨ consider the parse tree, while variants ① and ② consider both the parse tree and information entropy and thereby obtain better

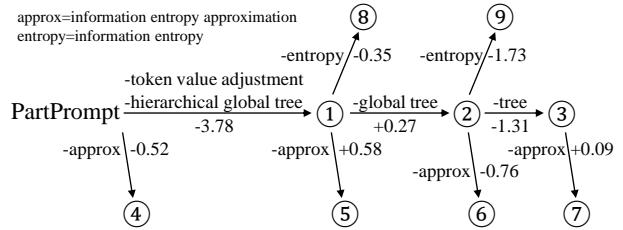


Fig. 4. The organization of variants of PartPrompt in the ablation study. Each arrow comes with the removed component(s) and the corresponding performance drop (if “-” or gain if “+”). Each circled number ①-⑨ at the end of an arrow indicates the variant of PartPrompt after removing the component of that arrow based on the variant at the beginning of that arrow. The detailed performance of these variants can be found in Table III.

TABLE III
THE PERFORMANCE OF VARIANTS OF PARTPROMPT IN ABLATION STUDY.
THE CIRCLED NUMBERS ①-⑨ DENOTE THE VARIANTS FOLLOWING THEIR ORGANIZATION AS SHOWN IN FIGURE 4.

methods	BLEU	Rouge1	Rouge2	RougeL	BS-F1
PartPrompt	28.28	56.48	38.30	54.29	74.57
①	17.08	49.68	25.50	46.72	70.79
②	18.10	50.21	26.29	47.30	71.06
③	14.61	47.24	22.55	43.94	69.75
④	28.20	55.95	37.86	53.77	74.04
⑤	18.83	50.88	27.38	47.85	71.37
⑥	16.63	48.95	24.86	45.91	70.29
⑦	16.12	47.74	23.93	44.66	69.83
⑧	18.01	49.04	26.56	46.30	70.44
⑨	15.44	46.94	23.82	44.18	69.33
Selective-Context	14.26	46.74	22.63	43.41	69.53
LLMLingua	15.71	45.98	23.21	43.17	69.52

performance (0.35% and 1.73%) respectively. This observation indicates that the parse tree and the information entropy are complementary to each other in prompt compression.

Next, we provide an analysis for variants ④, ⑤, ⑥, ⑦ without information entropy approximation. In Section IV-G, we have concluded that the inter-sentence error of information entropy approximation should be managed or may result in performance loss. For example, the performance of variant ③ with approximation is inferior to that of variant 6 without approximation. Moreover, variant ② conducts the compression process within each sentence, which avoids the comparison of token values among inter-sentences and the inter-sentence error, thereby resulting in good performance. Unlike variant ②, variant ① employs the global tree that compares token values globally. In this case, the inter-sentence error would lead to the performance loss. Overall, variant ① can benefit from either not using the global tree (i.e., ②) or from not using the approximation (i.e., ⑤). Nonetheless, the complete PartPrompt do use the global tree as in variant ① and further exploits a node value adjustment mechanism to correct the approximation error, and receives the best performance.

E. The Extreme Long Prompt Scenario

PartPrompt as well as Selective-Context and LLMLingua, employs a small language model to calculate the information

entropy of each token. When the original prompt is extreme long and exceeds the input limit of the small language model, both Selective-Context and LLMLingua become impractical as they computes the information entropy over the whole prompt. In contrast, PartPrompt calculates the information entropy within each sentence, thereby enabling it to manage the extreme long prompt with many sentences.

To demonstrate the capacity of PartPrompt in processing the extreme long prompt, we select the papers, exceeding 6000 tokens, from the arXiv dataset without truncation. Since the length far exceeds the input limit of the small language model for calculating entropy, Selective-Context and LLMLingua (two state-of-the-art selective prompt compression methods) are no longer applicable. To this end, we employ Qwen2-72B to directly generate compressed prompt for comparison. It is important to note that using an LLM for compression requires a significant amount of computational resources, which may not be feasible in practical scenarios. The experimental results are shown in Table IV, and the abstract written by humans is employed as the ground truth.

TABLE IV

THE PERFORMANCE OF PARTPROMPT FOR THE EXTREME LONG PROMPT SCENARIO. OTHER SELECTIVE COMPRESSION METHODS CANNOT HANDLE THIS SCENARIO, SO A GENERATIVE LLM IS USED FOR COMPARISON.

methods	BLEU	Rouge1	Rouge2	RougeL	BS-F1
LLM Generation	0.00	13.33	0.76	12.29	52.67
PartPrompt	0.00	14.37	0.77	13.41	51.07

It can be observed that PartPrompt outperforms LLM Generation (by Qwen2-72B) on more metrics, despite LLM Generation receives better performance on the BS-F1 metric. Besides, the BLEU metric for both methods is zero due to the extreme large compression request. Considering the first time to test the extreme long prompt scenario, achieving such results is particularly encouraging for the selective prompt compression methods in the literature.

F. Non-discourse and Discourse Metrics

To analyze whether the compressed prompt can effectively preserve the original prompt while maintaining coherence, a direct comparison is made between the compressed prompt and the original prompt, with both non-discourse and discourse metrics employed for evaluation. Non-discourse metrics is employed to measure the similarity of texts at the lexical level, including BLEU, Rouge, and BERTScore. Rouge (including Rouge1, Rouge2, and RougeL) measures the overlap between the hypothesis and reference texts. BERTScore (BS-F1 is used) offers a more comprehensive analysis of lexical similarity. BLEU consists of 1-gram, 2-gram, 3-gram, and 4-gram, where the n-gram calculates the proportion of n consecutive identical words. BLEU is employed to evaluate the continuity of the compressed text at the lexical level while measuring the lexical similarity. All the above non-discourse metrics are converted into percentages.

In contrast to non-discourse metrics, discourse metrics concern more on the global structure of texts, which can further

evaluate the coherence of the compressed prompt [43]–[45]. Concretely, RC and LC [46] measure the number and proportion of words that serve the connecting role. EntityGraph [47] evaluates text coherence through graphs. LexicalChain [48] measures the overlap of lexical chains between the hypothesis and reference texts. DiscoScore (including DS-Focus and DS-SENT) measures the difference between the hypothesis and reference texts through their focus. We adopt the same settings as in [45] for these non-discourse metrics.

Both BERTScore and DiscoScore have a maximum text length of 512 tokens, the BBCnews dataset is thus truncated to 500 tokens in the experiments. The compression ratio is set to 0.5. The results are shown in Table V, and note that the lower DS-Focus indicates the better performance.

TABLE V

THE PERFORMANCE OF A DIRECT COMPARISON BETWEEN THE COMPRESSED PROMPT AND THE ORIGINAL PROMPT. THE LOWER SCORE IS BETTER WITH ↓ SYMBOL, OTHERWISE THE HIGHER SCORE IS BETTER.

metrics	Selective-Context	LLMLingua	PartPrompt
Rouge1	66.20	63.58	74.80
Rouge2	33.50	45.10	59.96
RougeL	65.37	62.80	74.68
BS-F1	61.76	67.90	78.29
BLEU 1-gram	31.26	37.04	37.15
BLEU 2-gram	16.40	28.99	31.73
BLEU 3-gram	8.85	25.92	28.28
BLEU 4-gram	4.99	24.63	25.77
RC	0.378	0.388	0.607
LC	0.769	0.891	2.426
EntityGraph	0.402	0.310	0.480
LexicalChain	0.159	0.128	0.190
DS-Focus ↓	1.262	1.050	1.049
DS-SENT	0.762	0.882	0.869

PartPrompt exhibits the superior performance in almost all metrics for both non-discourse and discourse metrics. In particular, PartPrompt considerably outperforms other methods for most discourse metrics, i.e., better at maintaining the global structure and coherence of the original prompt. This would not only enhance the effectiveness of the compressed prompt while feeding it to LLMs, but also facilitate humans to understand how and why prompt compression may work.

G. Visualization of Contents Before and After Compression

This section intends to provide intuitive examples of the prompts before and after compression. We extract a segment from BBCnews and compare four selective prompt compression methods: Selective-Context, LLMLingua, PartPrompt, and PartPrompt without adjustment. The results are visualized in Figure 5, and the retained tokens are marked in green.

LLMLingua tends to preserve incomplete words, which may cause by the direct use of LLM tokenizer. For instance, it preserves '8' in '1988', 'o' in 'duo', and 'elling' in 'selling'. The incompleteness of these words would lead to a fragmented compressed prompt, which makes it less coherent and comprehensible. While Selective-Context employs a LLM tokenizer along with some semantic rules, which results in the better completeness of words comparing to LLMLingua. Note that,

Fig. 5. The compression results of PartPrompt (without adjustment) and the comparative methods on a brief intuitive example. All retained words are highlighted in green. LLMLingua retains tokens with incomplete semantics. Selective-Context retains too many function words. PartPrompt without adjustment can effectively identify key information and retain it. PartPrompt's compressed prompt is more coherent.

LLMLingua
Colorado-based New Belgium Brewing can trace its roots to 1988 and a cycle trip through Belgium. The experience inspired co-founders Kim Jordan and Jeff Lebesch to bring Belgian brewing techniques back to their home town. Three years later and the duo were selling Fat Tire, one of their first beers at a local festival, and they now have over a dozen beers in production.
Selective-Context
Colorado-based New Belgium Brewing can trace its roots to 1988 and a cycle trip through Belgium. The experience inspired co-founders Kim Jordan and Jeff Lebesch to bring Belgian brewing techniques back to their home town. Three years later and the duo were selling Fat Tire, one of their first beers at a local festival, and they now have over a dozen beers in production.
PartPrompt without adjustment
Colorado-based New Belgium Brewing can trace its roots to 1988 and a cycle trip through Belgium. The experience inspired co-founders Kim Jordan and Jeff Lebesch to bring Belgian brewing techniques back to their home town. Three years later and the duo were selling Fat Tire, one of their first beers at a local festival, and they now have over a dozen beers in production.
PartPrompt
Colorado-based New Belgium Brewing can trace its roots to 1988 and a cycle trip through Belgium. The experience inspired co-founders Kim Jordan and Jeff Lebesch to bring Belgian brewing techniques back to their home town. Three years later and the duo were selling Fat Tire, one of their first beers at a local festival, and they now have over a dozen beers in production.

retained token deleted token

both methods employ information entropy to determine the value of words, so that some less informative words might be retained. For example, Selective-Context retains '*and*' in '*and a cycle trip through Belgium*', and '*now over*' in '*they now have over a dozen beers*'. Besides, many function words alone, such as '*at*' and '*over*', cannot offer meaningful information for LLM inference and human comprehension.

The variant of PartPrompt without node value adjustment is better able to preserve the main contents of sentences and remove function words with the help of the parse tree. This allows the proposed method to preserve more useful information and specific contents given the same compression ratio. After adding the node value adjustment module, PartPrompt can be regulated to preserve the first part and keep the completeness of the compressed prompt by adjusting its hyper-parameters. The resulting example by PartPrompt clearly obeys the desired goal, and looks more coherent and easier to understand than other resulting examples by other methods.

VII. CONCLUSION

In this work, we introduced a novel prompt compression method called PartPrompt, which leverages parse trees to guide the prompt compressing process. To this end, the prompt compression problem was transformed into a tree pruning problem. And during the prompt compression process, the linguistic rules and human writing logic were achieved via constructing the global parse tree and adjusting the node value in the tree. The comparative experiments demonstrated the state-of-the-art performance of PartPrompt across a wide range of datasets, compression ratios, metrics, and target LLMs for inference. A comprehensive ablation study verified the usefulness of each component in PartPrompt. And additional experiments further confirmed the advantages of PartPrompt in terms of more scenarios and metrics such as the extreme long prompt case and the coherence of compressed prompts.

For future work, although PartPrompt has been shown capable of handling extremely long prompts with encouraging results, further efforts are still needed to enhance the performance of current prompt compression methods under the extreme long prompt scenario. Besides, the patterns of LLM prompt analyzed by linguistic parse trees and human writing logic have been shown effective in improving the performance for the prompt compression problem, it remains unknown whether other patterns of LLM prompt are also helpful, which is another promising future direction.

REFERENCES

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Meta. (2024) Introducing meta llama 3: The most capable openly available llm to date. Web page. [Online]. Available: <https://ai.meta.com/blog/meta-llama-3/>
- [3] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.
- [4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kütller, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [6] D. Li, J. Yan, T. Zhang, C. Wang, X. He, L. Huang, H. Xue, and J. Huang, "On the role of long-tail knowledge in retrieval augmented large language models," *arXiv preprint arXiv:2406.16367*, 2024.
- [7] Y. Fu, L. Ou, M. Chen, Y. Wan, H. Peng, and T. Khot, "Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance," *arXiv preprint arXiv:2305.17306*, 2023.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [9] Y.-N. Chuang, T. Xing, C.-Y. Chang, Z. Liu, X. Chen, and X. Hu, "Learning to compress prompt in natural language formats," *arXiv preprint arXiv:2402.18700*, 2024.

- [10] H. Gilbert, M. Sandborn, D. C. Schmidt, J. Spencer-Smith, and J. White, “Semantic compression with large language models,” in *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2023, pp. 1–8.
- [11] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [12] Y. Li, B. Dong, C. Lin, and F. Guerin, “Compressing context to enhance inference efficiency of large language models,” *arXiv preprint arXiv:2310.06201*, 2023.
- [13] H. Jiang, Q. Wu, C.-Y. Lin, Y. Yang, and L. Qiu, “Llmlingua: Compressing prompts for accelerated inference of large language models,” *arXiv preprint arXiv:2310.05736*, 2023.
- [14] Z. Pan, Q. Wu, H. Jiang, M. Xia, X. Luo, J. Zhang, Q. Lin, V. Rühle, Y. Yang, C.-Y. Lin *et al.*, “Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression,” *arXiv preprint arXiv:2403.12968*, 2024.
- [15] H. Jung and K.-J. Kim, “Discrete prompt compression with reinforcement learning,” *IEEE Access*, 2024.
- [16] X. Huang, L. L. Zhang, K.-T. Cheng, and M. Yang, “Boosting llm reasoning: Push the limits of few-shot learning with reinforced in-context pruning,” *arXiv preprint arXiv:2312.08901*, 2023.
- [17] Z. Zhang, Y. Wu, J. Zhou, S. Duan, H. Zhao, and R. Wang, “Sgnet: Syntax guided transformer for language representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3285–3299, 2020.
- [18] R. Hong, D. Liu, X. Mo, X. He, and H. Zhang, “Learning to compose and reason with language tree structures for visual grounding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 684–696, 2019.
- [19] Q. Cao, X. Liang, B. Li, and L. Lin, “Interpretable visual question answering by reasoning on dependency trees,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 887–901, 2019.
- [20] Y. Unno, T. Ninomiya, Y. Miyao, and J. Tsujii, “Trimming cfg parse trees for sentence compression using machine learning approaches,” in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 2006, pp. 850–857.
- [21] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimedia tools and applications*, vol. 82, no. 3, pp. 3713–3744, 2023.
- [22] K. Chowdhary and K. Chowdhary, “Natural language processing,” *Fundamentals of artificial intelligence*, pp. 603–649, 2020.
- [23] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [24] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.
- [25] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [27] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, “Emergent abilities of large language models,” *arXiv preprint arXiv:2206.07682*, 2022.
- [28] Z. Zhang, A. Zhang, M. Li, and A. Smola, “Automatic chain of thought prompting in large language models,” *arXiv preprint arXiv:2210.03493*, 2022.
- [29] Z. Li, Z. Zhang, H. Zhao, R. Wang, K. Chen, M. Utiyama, and E. Sumita, “Text compression-aided transformer encoding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3840–3857, 2021.
- [30] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [31] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, “The stanford corenlp natural language processing toolkit,” in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.
- [32] BBC. (2024) BBC. Web page. [Online]. Available: <https://www.bbc.com/>
- [33] Cornell. (2024) arXiv. Web page. [Online]. Available: <https://www.arxiv.org/>
- [34] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, “Hotpotqa: A dataset for diverse, explainable multi-hop question answering,” *arXiv preprint arXiv:1809.09600*, 2018.
- [35] R. Li and X. Du, “Leveraging structured information for explainable multi-hop question answering and reasoning,” *arXiv preprint arXiv:2311.03734*, 2023.
- [36] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, “Training verifiers to solve math word problems,” *arXiv preprint arXiv:2110.14168*, 2021.
- [37] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang *et al.*, “Qwen2 technical report,” *arXiv preprint arXiv:2407.10671*, 2024.
- [38] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand *et al.*, “Mixtral of experts,” *arXiv preprint arXiv:2401.04088*, 2024.
- [39] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [40] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.
- [41] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” *arXiv preprint arXiv:1904.09675*, 2019.
- [42] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [43] B. Grosz, A. Joshi, and S. Weinstein, “Centering: A framework for modeling the local coherence of discourse,” *Computational linguistics*, 1995.
- [44] W. C. Mann and S. A. Thompson, “Rhetorical structure theory: Toward a functional theory of text organization,” *Text-interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 243–281, 1988.
- [45] W. Zhao, M. Strube, and S. Eger, “Discoscore: Evaluating text generation with bert and discourse coherence,” *arXiv preprint arXiv:2201.11176*, 2022.
- [46] B. T. Wong and C. Kit, “Extending machine translation evaluation metrics with lexical cohesion to document level,” in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 2012, pp. 1060–1068.
- [47] C. Guinaudeau and M. Strube, “Graph-based local coherence modeling,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 93–103.
- [48] Z. Gong, M. Zhang, and G. Zhou, “Document-level machine translation evaluation with gist consistency and text cohesion,” in *Proceedings of the Second Workshop on Discourse in Machine Translation*, 2015, pp. 33–40.