

LanguaShrink:Reducing Token Overhead with Psycholinguistics

Anonymous ACL submission

Abstract

As large language models (LLMs) improve their capabilities in handling complex tasks, the issues of computational cost and efficiency due to long prompts are becoming increasingly prominent. To accelerate model inference and reduce costs, we propose an innovative prompt compression framework called LanguaShrink. Inspired by the observation that LLM performance depends on the density and position of key information in the input prompts, LanguaShrink leverages psycholinguistic principles and the Ebbinghaus memory curve to achieve task-agnostic prompt compression. This effectively reduces prompt length while preserving essential information. We referred to the training method of OpenChat. The framework introduces part-of-speech priority compression and data distillation techniques, using smaller models to learn compression targets and employing a KL-regularized reinforcement learning strategy for training(Wang et al., 2023). Additionally, we adopt a chunk-based compression algorithm to achieve adjustable compression rates. We evaluate our method on multiple datasets, including LongBench, Zero-Scrolls, Arxiv Articles, and a newly constructed novel test set. Experimental results show that LanguaShrink maintains semantic similarity while achieving up to 26 times compression. Compared to existing prompt compression methods, LanguaShrink improves end-to-end latency by 1.43 times. Our demo are available at <https://anonymous.4open.science/r/LanguaShrink-8159/> and demo at <https://languashrink.cpolar.top>

1 Introduction

In recent years, the field of large language models (LLM) has seen the emergence of various prompting

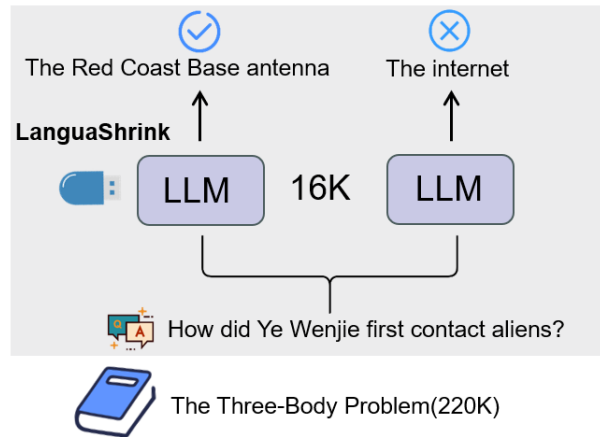


Figure 1: Illustration of the Plug-and-Play Document Module. The document encoding is decoupled from specific tasks. By inserting the document plugin into the task model, we can separate compressed text from downstream task reasoning and reduce computational costs.

techniques, such as Chain of Thought (CoT)(Wei et al.), In-context Learning (ICL)(Dong et al., 2022), and Retrieval Augmented Generation (RAG)(Lewis et al., 2020). These techniques have greatly expanded the capabilities of LLMs in handling complex and diverse tasks, by using prompts that can contain up to tens of thousands of vocabulary tokens(Manathunga and Hettigoda, 2023). However, while such lengthy prompts enhance processing capabilities, they also bring higher computational costs and financial burdens, posing challenges to the information processing and comprehension abilities of LLMs(Zhou et al., 2023).

To alleviate these issues, prompt compression techniques have emerged, aiming to reduce the length of the original prompts while preserving the core information and key instructions as much as possible, in order to optimize costs and efficiency(Mu

et al., 2023). Currently, many methods have been proposed for task-specific prompt compression, but these methods lack generality and portability. On the other hand, some other studies have explored task-agnostic prompt compression methods to pursue better generality and efficiency. These methods assume that natural language contains redundant information(Jiang et al., 2023), which may be useful for human understanding but might not be necessary for LLMs.

However, current task-agnostic methods face several challenges. Existing compression techniques mainly rely on simple token classification, which may lead to the loss of important sentence structure information(Kuvshinova and Khritankov, 2019). For complex long-text processing, effectively compressing without sacrificing the inherent logic and semantic structure of sentences remains an inadequately addressed issue(Wang and Chen, 2019). Additionally, most existing models do not effectively evaluate the importance of each sentence within a paragraph, which is crucial for maintaining the coherence and completeness of information in long texts(jian Luo et al., 2022).

To address these issues, we propose a new framework based on psycholinguistics, called LanguaShrink. LanguaShrink combines plug-and-play modules and psycholinguistic models to parse document information, using the Ebbinghaus memory curve to filter important information. This enables task-agnostic prompt compression and adapts to various open-source and proprietary large models. As shown in Figure 1, LanguaShrink can decouple compressed texts from downstream task reasoning and reduce computational costs (Hu et al., 2013; Murre and Dros, 2015).

Specifically, we use plug-and-play modules for compression, segment the text into chunks, and evaluate the semantic and structural importance of each chunk to avoid losing critical information. By using a comprehensive weighting method, we assess the relevance and perplexity of the chunks, selecting those with high relevance and low perplexity to improve the coherence and completeness of the compressed text. Additionally, we propose a data distillation method that uses small models to learn the compression target, thereby reducing latency(Ma et al., 2020). We incorporate a reinforcement learn-

ing framework based on KL regularization, refining the training process with different reward weights.

We validate the effectiveness of our method on three datasets from different domains, namely Longbench(Bai et al., 2023), ZeroScrolls(Shaham et al., 2023), and Arxiv Articles(Clement et al., 2019), and we also construct a new long-text novel test set. Experimental results show that our method achieves better semantic similarity compared to existing prompt compression methods at the same compression rate, while reducing end-to-end latency by 1.43 times and achieving a compression ratio of 2x to 8x.

The main contributions of our work are as follows:

- We propose a plug-and-play prompt compression system based on psycholinguistics and the Ebbinghaus memory curve to filter important information.
- We propose a data distillation method that uses smaller models to learn the compression target, optimizing training through a reinforcement learning framework based on KL regularization.
- We conduct extensive experiments on various datasets, and the results demonstrate that our method achieves up to 26x compression without compromising performance.

2 Related work

2.1 Psycholinguistics

Psycholinguistic research has two main areas: sentence processing and text processing(McKoon and Ratcliff, 1998). Sentence processing focuses on how the syntactic structure of sentences is computed(Alyahya et al., 2018). Text processing involves understanding the meaning of larger units of text. Conversely, function words and key nouns play crucial roles in sentences (Kalyuga, 2012).

Existing research indicates that removing redundant information can effectively improve the efficiency of foreign language vocabulary learning(Ellis and Beaton, 1993). Additionally, this helps optimize storage space(Schmidhuber, 2000). Based on these research findings, we propose a psycholinguistics-based Part-of-Speech Priority Compression (PPC) algorithm that uses lexical classification and priority

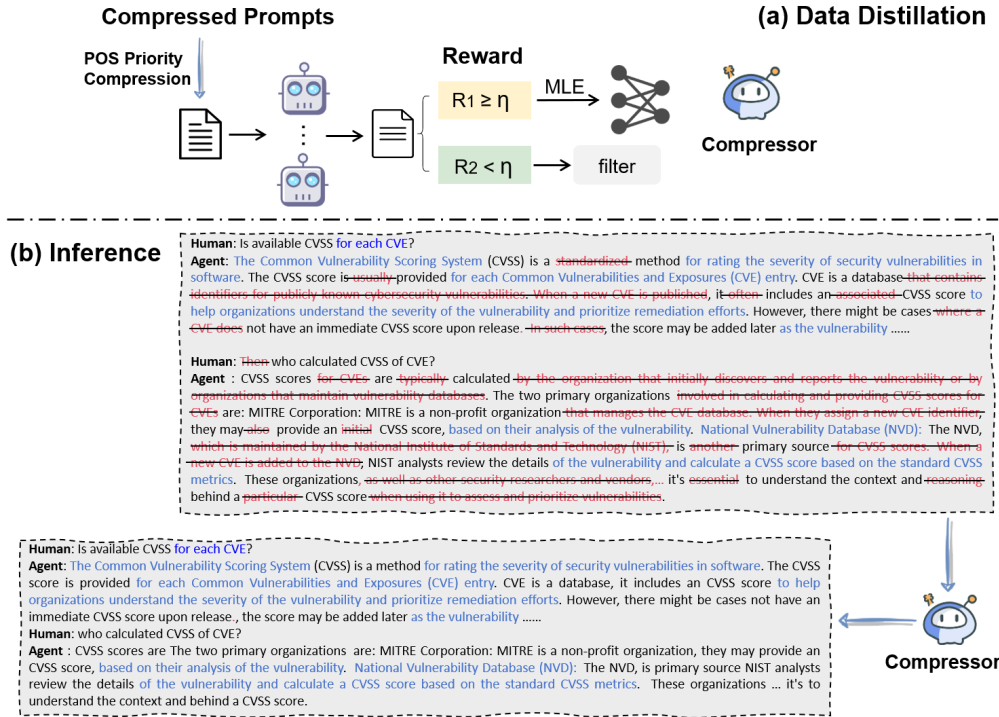


Figure 2: (a) Data distillation. Initial text compression is first performed using POS priority compression. Next, the compressed prompts are evaluated based on the similarity and compression ratio between the compressed prompt and the original prompt. If the similarity is above the threshold, the model receives a reward; otherwise, the reward is zero and it is filtered out. Then, the model is fine-tuned using Maximum Likelihood Estimation (MLE), and finally, the compressor generates the compressed prompts. (b) Inference. The application of the compressor in actual question-answering tasks is demonstrated. The effect of LinguaShrink compression processing on the original dialogue is shown. Red indicates the parts that are most likely to be compressed, blue indicates the parts that are next most likely to be compressed.

149 assignment to more efficiently retain core informa- 167
 150 tion and eliminate redundant content(Graça et al., 168
 151 2011). 169

152 2.2 Prompt Compression 170

153 LLMs face significant challenges in handling long 171
 154 contexts. Due to the quadratic growth in memory 172
 155 and computational demands of the attention mech- 173
 156 anism, the computational cost of processing long 174
 157 texts is extremely high(Han et al., 2023; Zhuang 175
 158 et al., 2022; Chen et al., 2023). Existing LLMs typi- 176
 159 cally use a fixed context window during pre-training, 177
 160 which further limits their ability to handle longer 178
 161 contexts. To address this issue, researchers have pro- 179
 162 posed methods such as sparse attention and local 180
 163 dense attention to reduce computational and memory 181
 164 costs. Additionally, soft prompt tuning and reinforc- 182
 165 ement learning-based compression methods have been 183
 166 applied to save context costs during inference(Shen 184
 185

et al., 2018; Liu et al., 2023).

Prompt compression is one of the direct meth- 168
 169 ods to address the problem of LLMs handling long 170
 171 contexts, aiming to shorten prompt length while re- 172
 173 taining important information. Typical prompt com- 174
 175 pression methods are divided into task-aware and 176
 177 task-agnostic categories. Task-aware compression 178
 179 adjusts the context based on downstream tasks or cur- 180
 181 rent queries, such as LongLLMLingua(Jiang et al., 182
 183 2023), which adjusts the compression ratio by esti- 184
 185 mating token information entropy. Task-agnostic 185
 compression, on the other hand, applies to a wide 186
 range of applications and typically uses information 187
 entropy measures to remove redundant information 188
 from the prompt. Although these methods have im- 189
 proved computational efficiency and model perfor- 190
 mance to some extent, further exploration and opti- 191
 mization are needed to effectively handle long texts 192
 and complex tasks in a wide range of real-world 193
 194
 195

186 applications(Hsieh et al., 2023).

187 The OpenChat project proposed an effective mech- 220
188 anism for prompt compression fine-tuning. This 221
189 method dynamically adjusts prompts through rein- 222
190 forcement learning strategies to retain the most im- 223
191 portant information for tasks, thereby reducing un- 224
192 necessary computational overhead. Additionally, the 225
193 Kullback-Leibler (KL) divergence is used to measure 226
194 changes in prompt information before and after com- 227
195 pression. By minimizing KL divergence, it ensures 228
196 that critical information from the original prompt is 229
197 preserved during compression. 230

198 In this paper, we propose the Prompt Compression 231
199 Fine-Tuning (PC-RLFT) method, which achieves 232
200 efficient prompt compression and core information 233
201 retention by incorporating chunk compression tech- 234
202 niques. 235

Algorithm 1 Compression Algorithm

Require: T : input text
1: $C \leftarrow \text{SplitToChunks}(T)$
2: $n \leftarrow \text{ChunkCount}(C)$
3: **for** $i \leftarrow 1$ to n **do**
4: $c_i \leftarrow C[i]$
5: $\text{TokenCompression}(c_i)$
6: **end for**
7: $T' \leftarrow \text{JoinChunks}(C)$
8: **return** T'

203 3 Method

204 3.1 POS Priority Compression

205 PPC is achieved by inputting carefully designed 236
206 prompts into LLMs. To implement PPC, we need 237
207 to design a series of specific prompts that achieve 238
208 part-of-speech priority compression through the CoT 239
209 approach. Below is the design idea for CoT: 240

210 **Relation Word Extraction** We design prompts 241
211 to guide the model in identifying relation words in 242
212 the text. By using dependency syntax analysis, the 243
213 model can understand the relationships between sen- 244
214 tences. The model assigns different priorities to each 245
215 relation word based on the context of the entire sen- 246
216 tence. 247

217 **Part-of-Speech Classification** The prompts guide 248
218 the model to classify words in the text by their parts 249
219 of speech, such as adjectives, adverbs, nouns, and 250

220 prepositions. This is based on psycholinguistic part-
221 of-speech analysis, assigning a priority to each part
222 of speech. For example, the priority standard is:
223 nouns > verbs > adjectives > adverbs. 224

225 **Priority Filtering** After completing the extraction
226 of relation words and part-of-speech classification,
227 the model uses this information to filter out the words
228 and sentences that contribute most to the core mean-
229 ing of the text, while deleting lower-priority words
230 and sentences that have a minimal impact on the
overall understanding.

Algorithm 2 Chunk-Based Compression Algorithm

Require: T : input text, Q : query, α : relevance
weight, β : importance weight, R_t : target com-
pression rate
 $T' \leftarrow \text{Compression}(T)$
 $C \leftarrow \text{SplitToChunks}(T')$
 $n \leftarrow \text{ChunkCount}(C)$
 $R_0 \leftarrow \text{CalcCompRate}(T, T')$
for $i \leftarrow 1$ to n **do**
 $c_i \leftarrow C[i]$
 $rel_i \leftarrow \text{CosineSim}(c_i, Q)$
 $imp_i \leftarrow \text{CalcImportance}(c_i)$
 $ppl_i \leftarrow \text{CalcPerplexity}(c_i)$
 $w_i \leftarrow \alpha \times rel_i + \beta \times imp_i$
 $C[i] \leftarrow (c_i, w_i, ppl_i)$
end for
 $\text{SortByWeightAndPerp}(C)$
 $k \leftarrow n$
 $R \leftarrow R_0$
while $R > R_t$ **do**
 $T'' \leftarrow \text{JoinTopKChunks}(C, k)$
 $R \leftarrow \text{CalcCompRate}(T, T'')$
 if $R \leq R_t$ **then**
 break
 end if
 $k \leftarrow k - 1$
end while
return T'', R

231 3.2 Dataset Distillation

232 We propose a data distillation method that extracts
233 knowledge from large language models (LLMs) to
234 generate compressed prompts that retain key informa-
235 tion while reducing latency by using smaller models
236 to learn the compression targets. Additionally, we en- 237

sure the compressed prompts remain highly faithful to the original content.

Dataset: We source our data from reading materials in the Chinese Gaokao and the postgraduate entrance exam English sections. These reading materials provide a rich variety of texts suitable for compression training. The dataset contains 20,000 samples, each processed by segmenting the reading materials into blocks of three consecutive sentences. This structure allows the model to learn effective compression while avoiding the loss of critical contextual information.

To generate compressed data, we use various LLMs, including both open-source and proprietary models such as GPT-4¹, Yi², GLM³, and Qwen⁴. These LLMs are used to distill knowledge according to psycholinguistic principles, creating compressed prompts that retain subject-verb-object structures and maximize the preservation of key information. We also ensure the compressed content remains semantically similar to the original content. The selection of multiple LLMs is based on their complementary strengths in language understanding and generation. By integrating knowledge from multiple models, we achieve a more comprehensive compression perspective, enhancing the quality of the compressed data and the model’s generalization capabilities.

3.3 Prompt Compress-RLFT

3.3.1 Reward Design

The reward consists of two components: one is based on the cosine similarity score, which measures the similarity between the output sequences generated from the original and compressed prompts; the other is the compression ratio τ , reflecting the reduction in prompt length. If the cosine similarity score exceeds a certain threshold τ , the model receives the compression ratio as a reward; if it does not, the reward is zero.

3.3.2 Tuning

We selected the pre-trained Qwen as the smaller language model (SLM). The distilled dataset is then

¹<https://openai.com/>

²<https://www.lingyiwanwu.com/>

³<https://chatglm.cn/>

⁴<https://tongyi.aliyun.com/qianwen/>

used for Prompt Compress Reinforcement Learning Fine Tuning (PC-RLFT). During fine-tuning, we combine the PC-RLFT method based on a KL regularization reinforcement learning framework. We assign different reward weights to the data to refine the training process.

KL-regularized RL objective is defined as follow:

$$J_{PC-RLFT}(\theta) = \mathbb{E}_{y \sim \pi_\theta} [r_c(x, y)] - \beta D_{KL}(\pi_\theta, \pi_c) \quad (1)$$

where π_θ is the policy parameterized by θ , $r_c(x, y)$ is the class-conditioned reward function, π_c is the higher-quality class-conditioned behavior policy, β is a scaling factor for the KL divergence term, and D_{KL} represents the KL divergence (Wang et al., 2023).

Previous work has demonstrated that the optimal solution to the KL-regularized reward maximization objective is as follows:

$$\pi^*(y|x, c) \propto \pi_c(y|x, c) \exp\left(\frac{1}{\beta} r_c(x, y)\right) \quad (2)$$

where π^* signifies the optimal policy for a given class c and input x .

The method to extract the optimized policy π_θ by minimizing the KL divergence:

$$\begin{aligned} \pi_\theta &= \arg \min_{\theta} \mathbb{E}_{(x,c) \sim D_c} [D_{KL}(\pi^*(\cdot|x, c) || \pi_\theta(\cdot|x, c))] \\ &= \max_{\theta} \mathbb{E}_{(x,y,c) \sim D_c} \left[\exp\left(\frac{1}{\beta} r_c(x, y)\right) \log \pi_\theta(y|x, c) \right] \end{aligned} \quad (3)$$

Equation 3 outlines the process for minimizing the KL divergence between π^* and π_θ over the class-conditioned dataset D_c . The final expression represents the reward-weighted regression objective for the optimized policy π_θ .

In this study, we propose a chunk-based compression algorithm. First, the input text T is preprocessed through a standard token compression process, segmenting it into chunks $C[i]$ consisting of three consecutive sentences. This step achieves a high initial compression rate R_0 .

3.4 Chunk-Based Compression

As shown in the alg 1 and alg 2.A For each chunk $C[i]$, the algorithm evaluates its relevance to the query Q by calculating the cosine similarity $CS(C[i], Q)$. The relevance is denoted as rel_i . The

semantic and structural importance of the chunk is calculated using the function $CI(C[i])$, producing an importance score imp_i . Additionally, the perplexity ppl_i of the chunk, as a measure of information content, is calculated using the function $CP(C[i])$.

The combined weight w_i of each chunk is calculated using the following formula:

$$w_i = \alpha \times rel_i + \beta \times imp_i$$

where α and β are coefficients that adjust the influence of relevance and importance. This weight determines the retention priority of the chunk in the final compressed text.

The chunks are then sorted based on their weight w_i and perplexity ppl_i , with higher weights and lower perplexities being prioritized for retention to optimize information preservation and compression effectiveness. The number of retained chunks k is adjusted in a decremental manner until the compression rate R reaches the target compression rate R_t . Finally, the selected chunks are recombined to form the compressed text.

4 Experiment

4.1 Settings

Implementation Details We use the PPC method to analyze the text, extracting a large compressed dataset. By training the model using the PC-RLFT approach, we obtain a smaller model. All reported metrics use GPT-3.5 as the target LLM for downstream tasks. To improve the stability of the outputs generated by the LLM, we apply greedy decoding with a temperature of 0 in all experiments. In our experiments, we utilize Qwen-1.8B as the smaller pre-trained language model for compression.

Dataset and Metrics To comprehensively evaluate the effectiveness of compressed prompts in retaining LLM capabilities, we assess their performance across multiple datasets. For long-context scenarios, we use LongBench and ZeroSCROLLSE.

Baselines We adopt two state-of-the-art prompt compression methods as the primary baselines for comparison: Selective-Context and the llmlingua(Jiang et al., 2023; Pan et al., 2020) series. Additionally, we compare our method with several task-aware prompt compression methods, such as retrieval-based methods and longllmlingua.

4.2 Main Results

Table 1 presents the performance of various methods under different compression constraints. Despite our compression model being much smaller than LLAMa-2-7B⁵ or other models used as baselines, our approach achieves better performance in both QA and synthesis tasks. Compared to the original prompts, our compressed prompts achieve comparable performance at a lower cost. Our model exhibits superior performance compared to other task-agnostic baselines, demonstrating the effectiveness of our constructed dataset and highlighting the importance and benefits of optimizing compression models using prompt compression knowledge.

Compression-based methods, such as selective context and LLMLingua, perform poorly on most tasks. This is due to their purely information entropy-based compression mechanism, which includes too much noise in the compressed results. Retrieval-based methods rely on finding the most relevant fragments to the query from a large number of documents. However, in practical applications, these fragments may contain a lot of redundant information, leading to lower overall information density.

We find that compressing tokens leads to a decline in mathematical capabilities, possibly because psycholinguistics is less sensitive to mathematical content. In the Longbench test, LLMLingua2 has a slight advantage in few-shot and code tasks, but LanguaShrink performs better in text or Q&A compression tasks. This difference indicates that while psycholinguistic techniques have significant advantages in text compression, they are still inadequate in handling mathematical content.

In the ArXiv tests, as shown in Table 2, our method also performs outstandingly in Compression Semantic Efficiency (CSE), a new metric that combines compression ratio and semantic similarity. Our results are 2.46 times better than llmlingua2. In terms of BLEU scores, our method shows significant improvement compared to Select context and llmlingua2. Regarding ROUGE scores, since our method alters sentence structures to maximize semantic retention, our performance in this metric is comparable to other methods, without significant improvement.

⁵<https://github.com/Meta-Llama/llama>

Table 1: Performance of different methods under different compression ratios on LongBench and ZeroSCROLLS using GPT-3.5-Turbo.

Methods	LongBench									ZeroSCROLLS		
	SingleDoc	MultiDoc	Summ.	FewShot	Synth.	Code	AVG	Tokens	1/ τ	AVG	Tokens	1/ τ
2,000 tokens constraint												
Retrieval-based Methods												
BM25	30.1	29.3	21.3	12.5	19.5	29.1	23.63	1802	5x	20.1	1,799	5x
SBERT	33.8	36.0	25.8	23.5	12.5	29.0	23.6	1947	5x	20.5	1,773	5x
OpenAI	34.3	36.4	24.6	26.3	32.4	24.8	30.47	1991	5x	20.6	1,784	5x
LongLLMLingua	<u>37.8</u>	<u>41.7</u>	26.9	64.3	53.0	52.4	<u>46.0</u>	1960	5x	24.9	1,771	5x
Compression-based Methods												
Selective-Context(Li, 2023)	16.2	34.8	24.4	8.4	15.7	49.2	24.8	1925	5x	19.4	1,865	5x
LLMLingua	22.4	32.1	24.5	61.2	10.4	56.8	34.6	1,950	5x	27.2	1,862	5x
LLMLingua2-small	29.5	32.0	24.5	<u>64.8</u>	22.3	56.2	38.2	1,891	5x	<u>33.3</u>	1,862	5x
LLMLingua2	29.8	33.1	25.3	66.4	21.3	58.9	39.1	1,954	5x	<u>33.3</u>	1,898	5x
LanguaShrink	42.1	54.3	<u>26.3</u>	62.3	<u>33.0</u>	<u>58.4</u>	46.1	1,988	5x	39.0	1,871	5x
3,000 tokens constraint												
Retrieval-based Methods												
BM25	32.3	34.3	25.3	57.9	45.1	48.9	40.6	3,417	3x	19.8	3,379	3x
SBERT	35.3	37.4	26.7	63.4	51.0	34.5	41.4	3,399	3x	24.0	3,340	3x
OpenAI	34.5	38.6	<u>26.8</u>	63.4	49.6	37.6	41.7	3,421	3x	22.4	3,362	3x
LongLLMLingua	<u>37.6</u>	<u>42.9</u>	26.9	<u>68.2</u>	<u>49.9</u>	53.4	<u>46.5</u>	3,424	3x	33.5	3,206	3x
Compression-based Methods												
Selective-Context	23.3	39.2	25.0	23.8	27.5	53.1	32.0	3,328	3x	20.7	3,460	3x
LLMLingua	31.8	37.5	26.2	67.2	8.3	53.2	37.4	3,421	3x	30.7	3,366	3x
LLMLingua2-small	35.5	38.1	26.2	67.5	23.9	<u>60.0</u>	41.9	3,278	3x	33.4	3,089	3x
LLMLingua2	35.5	38.7	26.3	69.6	21.4	62.8	42.4	3,392	3x	<u>35.5</u>	3,206	3x
LanguaShrink	42.2	54.5	26.3	62.6	34.0	62.8	47.1	3,488	3x	39.6	3,197	3x
Original Prompt	41.7	38.7	26.5	67.0	37.8	54.2	44.9	10,295	-	34.7	9,788	-

Table 2: Statistics of Arxiv Articles CSE, BLEU, ROUGE are averaged per document.

Method	CSE	BLEU	ROUGE
Select context	3.1080	0.0010	0.2063
llmlingua2	1.4845	0.0008	0.2015
LanguaShrink	3.6555	0.0235	0.2015

Table 3: Method performance statistics. Tokens and time are averaged values.

Method	Tokens (avg)	Time (avg)
LanguaShrink	3502.75	24.29
LanguaShrink (w/o psy.)	3811.3	33.99
LanguaShrink (w/o SA)	3770.5	35.74

Table 4: F1 Scores of Different Methods

Method	F1
llmlingua2	21.7
Select context	18.3
LanguaShrink	26.0
LanguaShrink(w/o psy.)	17.3
LanguaShrink(w/o SA.)	22.1
original	27.6

4.3 Ablation Study

Our method consists of two core components: a psycholinguistic analysis module and a sentence analysis module. As shown in Table 3, when we remove the psycholinguistic core compression component, we find a nearly 10% decrease in compression capa-

406
407
408
409
410
411

Table 5: Method performance statistics. Tokens, CSE, and BLEU are averaged values.

Method	Tokens (avg)	CSE	BLEU
lmlingua2	280.10	0.9419	0.0288
our	253.15	1.0462	0.0304
Select context	270.15	0.9413	0.0273
original	343.65	-	-

bility. This is mainly due to the lack of the linguistic analysis part, which leads to an inability to quickly and accurately locate tokens. When we remove the sentence analysis component, although the ability to quickly and accurately locate tokens is regained, the compression performance and efficiency decrease due to the inability to identify the key parts of sentences to compress.

As shown in Table 4, we first use the constructed novel dataset as the original context to generate questions and answers, where these answers are considered reference answers, and then require the LLM to answer these questions. We find that with the psycholinguistic core compression component present, even without the sentence analysis module, LanguaShrink can reach the level of lmlingua2.

The CSE metric we propose measures the true effectiveness of large model compression methods. When the value is below 1, compressing the large model does not improve token performance; instead, it significantly reduces token performance. When the value exceeds 1, the compression method can increase the context length. In tests with text that is already highly refined, compression may lead to a decrease in CSE performance. As shown in Table 5, in our tests of three models, none achieve the specified compression ratio, with the compression rate being around 30%. However, our method still exceeds 1 when other methods are below 1, proving the effectiveness of our method even in extreme conditions. Please refer to the appendix F for more examples.

4.4 Latency Evaluation

We conducted tests on the A800-80GB GPU, using the same prompt as indicated in the appendix, which on average contained 10K tokens, and set the response length to 200 tokens in the API calls. In Table 6, "E2E" represents the latency of each prompt

Table 6: Latency and Speedup Factor of Different Methods

Method	Latency(s)	Speedup Factor
lmlingua	7.48	1.6x
Select context	7.56	1.6x
LanguaShrink	6.64	1.8x
original	11.84	-

compression system and the black-box API. The results show that our prompt compression system indeed accelerates the overall inference. This acceleration effect becomes more pronounced with the increase in compression rates. It is worth noting that in scenarios where the API’s cost time is longer, the actual absolute time saved by LanguaShrink may be more significant (Cao et al., 2023; Stone et al., 2008; Yazdanbakhsh et al., 2015).

5 Conclusion

In this paper, we propose LanguaShrink, an innovative prompt compression framework aimed at improving the efficiency and performance of LLMs by reducing the length of prompts while preserving core information. LanguaShrink leverages psycholinguistic models and the Ebbinghaus memory curve to achieve task-agnostic compression compatible with various LLMs. We introduce a method based on part-of-speech priority compression and data distillation techniques, using smaller models to learn compression targets and employing a KL-regularized reinforcement learning strategy for training. Additionally, we adopt a chunk-based compression algorithm, evaluating each chunk’s relevance, importance, and perplexity to adjust the retention priority and achieve adjustable compression rates. Extensive experimental results show that LanguaShrink significantly outperforms existing techniques in semantic similarity and compression efficiency across multiple datasets, achieving up to 26 times compression while maintaining performance comparable to the original prompts.

Limitations

Currently, our token compression technology mainly incorporates psycholinguistic techniques and has not

484 yet integrated RAG (Retrieval-Augmented Gener- 530
485 ation) technology. In early experiments, we tried 531
486 various psycholinguistic knowledge and ultimately 532
487 selected the two most effective methods for exper- 533
488 imentation, but we did not fully apply all the psy- 534
489 cholinguiistic knowledge. In these early experiments, 535
490 we partially used the Oxford Dictionary for train- 536
491 ing. Although this yielded good results, we were
492 unable to conduct comprehensive testing due to not
493 having collected the complete content of the Oxford
494 Dictionary.

495 LanguaShrink becomes unstable when the com-
496 pression rate exceeds 90%. Although we introduce
497 an extreme mode to address this issue, it is not an
498 ideal long-term solution. While the extreme mode
499 can temporarily mitigate the performance degrada-
500 tion caused by excessive compression, it may intro-
501 duce other complexities and resource consumption
502 in practical applications.

503 In the future, our research focuses on further op-
504 timizing the application of psycholinguistic tech-
505 niques, exploring more diverse integration methods,
506 and addressing the decline in mathematical capabili-
507 ties to achieve breakthroughs in a broader range of
508 application scenarios.

509 Ethics Statement

510 The development and application of LanguaShrink
511 also raise several ethical considerations: Bias and
512 Fairness: The datasets used for training and evaluat-
513 ing LanguaShrink must be carefully curated to ensure
514 they are representative and do not perpetuate biases.
515 Any inherent biases in the data could be amplified
516 through the compression process, leading to unfair
517 or biased outputs from the LLMs. Privacy and Confi-
518 dentiality: When applying LanguaShrink to sensitive
519 or confidential information, it is crucial to ensure
520 that the compression process does not inadvertently
521 expose or compromise any personal or sensitive data.
522 Robust data handling and privacy-preserving tech-
523 niques must be implemented. Transparency and Ac-
524 countability: The use of LanguaShrink should be
525 transparent, with clear documentation on how the
526 compression is performed and its potential impacts
527 on the data. Users should be informed about the
528 limitations and potential risks associated with the
529 compressed prompts to make informed decisions

about their use. Impact on Employment: The effi-
ciency gains from using LanguaShrink could lead to
reduced demand for certain roles involved in manual
data processing and prompt generation. It is essential
to consider the socio-economic impacts and provide
support for individuals who might be affected by
such technological advancements.

References

- 537
538 Reem S. W. Alyahya, A. Halai, Paul Conroy, and
539 M. L. Lambon Ralph. 2018. [Noun and verb processing
540 in aphasia: Behavioural profiles and neural correlates.](#)
541 *NeuroImage : Clinical*, 18:215 – 230.
- 542 Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai
543 Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan
544 Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi
545 Li. 2023. Longbench: A bilingual, multitask bench-
546 mark for long context understanding. *arXiv preprint*
547 *arXiv:2308.14508*.
- 548 Kai Cao, Qizhong Wu, Ling-Yu Wang, Nan Wang,
549 Huaqiong Cheng, Xiao Tang, Dongqing Li, and
550 Lanning Wang. 2023. [Gpu-hadvppm v1.0: a
551 high-efficiency parallel gpu design of the piecewise
552 parabolic method \(ppm\) for horizontal advection in an
553 air quality model \(camx v6.10\).](#) *Geoscientific Model
554 Development*.
- 555 Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhi-
556 jian Liu, Song Han, and Jiaya Jia. 2023. [Longlora:
557 Efficient fine-tuning of long-context large language
558 models.](#) *ArXiv*, abs/2309.12307.
- 559 Colin B. Clement, Matthew Bierbaum, Kevin P. O’Keeffe,
560 and Alexander A. Alemi. 2019. [On the use of arxiv as
561 a dataset.](#) *Preprint*, arXiv:1905.00075.
- 562 Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong
563 Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and
564 Zhifang Sui. 2022. [A survey for in-context learning.](#)
565 *Preprint*, arXiv:2301.00234.
- 566 N. Ellis and A. Beaton. 1993. [Psycholinguistic determi-
567 nants of foreign language vocabulary learning.](#) *Lan-
568 guage Learning*, 43:559–617.
- 569 João Graça, Kuzman Ganchev, Luísa Coheur, Fernando C
570 Pereira, and B. Taskar. 2011. [Controlling complexity in
571 part-of-speech induction.](#) *J. Artif. Intell. Res.*, 41:527–
572 551.
- 573 Insu Han, Rajesh Jayaram, Amin Karbasi, V. Mirrokni,
574 David P. Woodruff, and A. Zandieh. 2023. [Hyperatten-
575 tion: Long-context attention in near-linear time.](#) *ArXiv*,
576 abs/2310.05869.

577	Cho-Jui Hsieh, Si Si, Felix X. Yu, and I. Dhillon. 2023. Automatic engineering of long prompts . <i>ArXiv</i> , abs/2311.10117.	626
578		627
579		628
580	S. Hu, Y. Liu, Tupei Chen, Zengcai Liu, Q. Yu, L. Deng, Y. Yin, and S. Hosaka. 2013. Emulating the ebbinghaus forgetting curve of the human brain with a nio-based memristor . <i>Applied Physics Letters</i> , 103:133701.	629
581		630
582		631
583		
584	Tian jian Luo, Yuqi Liu, and Tianning Li. 2022. A multi-feature fusion method with attention mechanism for long text classification . <i>Proceedings of the 2022 6th International Conference on Compute and Data Analysis</i> .	632
585		633
586		
587		
588		
589	Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models .	634
590		635
591		636
592		637
593		638
594		639
595		640
596		641
597		642
598		643
599		644
600		645
601		646
602		
603		
604		
605		
606		
607		
608		
609		
610		
611		
612		
613		
614		
615		
616		
617		
618		
619		
620		
621		
622		
623		
624		
625		
	G. McKoon and R. Ratcliff. 1998. Memory-based language processing: psycholinguistic research in the 1990s . <i>Annual review of psychology</i> , 49:25–42.	626
		627
		628
	Jesse Mu, Xiang Lisa Li, and Noah D. Goodman. 2023. Learning to compress prompts with gist tokens . <i>ArXiv</i> , abs/2304.08467.	629
		630
		631
	J. Murre and Joeri Dros. 2015. Replication and analysis of ebbinghaus’ forgetting curve . <i>PLoS ONE</i> , 10.	632
		633
	Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, HVicky Zhao, Lili Qiu, Dongmei Zhang, Alexis Chevalier, Alexander Wetig, Anirudh Ajith, Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Moyer, Veselin Zettle, and Unsupervised Stoyanov. 2020. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression .	634
		635
		636
		637
		638
		639
		640
		641
		642
		643
		644
		645
		646
	J. Schmidhuber. 2000. Neural predictors for detecting and removing redundant information . pages 1152–1167.	647
		648
	Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 7977–7989, Singapore. Association for Computational Linguistics.	649
		650
		651
		652
		653
		654
	Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. 2018. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling . pages 4345–4352.	655
		656
		657
		658
	Sam S. Stone, Justin P. Haldar, Stephanie C. Tsao, Wen mei W. Hwu, Bradley P. Sutton, and Zhi-Pei Liang. 2008. Accelerating advanced mri reconstructions on gpus . <i>J. Parallel Distributed Comput.</i>	659
		660
		661
		662
	Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data . <i>arXiv preprint arXiv:2309.11235</i> .	663
		664
		665
		666
	Yifan Wang and Guang Chen. 2019. Improving a syntactic graph convolution network for sentence compression . pages 131–142.	667
		668
		669
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models .	670
		671
		672
		673

674 A. Yazdanbakhsh, Jongse Park, Hardik Sharma, P. Lotfi-
675 Kamran, and H. Esmailzadeh. 2015. *Neural accelera-
676 tion for gpu throughput processors*. *2015 48th Annual
677 IEEE/ACM International Symposium on Microarchi-
678 tecture (MICRO)*, pages 482–493.

679 Jianpeng Zhou, Wanjun Zhong, Yanlin Wang, and Jiahai
680 Wang. 2023. *Adaptive-solver framework for dynamic
681 strategy selection in large language model reasoning*.
682 *ArXiv*, abs/2310.01446.

683 Yimeng Zhuang, Jing Zhang, and Mei Tu. 2022. *Long-
684 range sequence modeling with predictable sparse atten-
685 tion*. pages 234–243.

686 A Dataset Details

687 Arxiv-March23 A dataset consisting of latest aca-
688 demic papers created in March 2023 from the arXiv
689 preprint repository. We use 500 data items collected
690 by Li as the test set. Due to the excessive length of
691 some articles, we take the first five sections of each
692 article and truncate each section to 10,000 charac-
693 ters. Then, we concatenate these sections to form the
694 original prompt and use GPT-3.5-Turbo to generate
695 the summary as the reference.

696 LongBench A multi-task long context benchmark
697 consists of 3,750 problems in English and includes
698 six categories with a total of 16 tasks. These tasks
699 encompass key long-text application scenarios, such
700 as single-document QA, multi-document QA, sum-
701 marization, few-shot learning, synthetic tasks, and
702 code completion. The average prompt token length
703 in this benchmark is 10,289.

704 ZeroSCROLLS The multi-task long context
705 benchmark consists of 4,378 problems, including
706 four categories with a total of 10 tasks. These tasks
707 cover summarization, question answering, aggre-
708 gated sentiment classification, and information re-
709 ordering. The average prompt token length in this
710 benchmark is 9,788.

711 B Other implementation details

- 712 i. **LongBench:(Bai et al., 2023)** This benchmark
713 consists of six task types: single-document QA,
714 multi-document QA, summarization, few-shot
715 learning, code completion, and synthetic tasks.
716 We evaluate using the English portion, covering
717 16 datasets. We use the provided metrics and
718 scripts for evaluation.

ii. **ZeroSCROLLS:(Shaham et al., 2023)** This
719 benchmark comprises four task types: summa-
720 rization, QA, sentiment classification, and re-
721 ordering, covering 10 datasets. We evaluate us-
722 ing the validation set and employ the provided
723 metrics and scripts for evaluation. 724

iii. **Arxiv-March:(Clement et al., 2019)** A dataset
725 composed of the latest academic papers cover-
726 ing various scientific disciplines. Due to the
727 potential length of arXiv articles, in our experi-
728 ments, we only process the first two sections of
729 each paper (usually the introduction and back-
730 ground). We compress the content of all the
731 papers and compare the effects before and after
732 compression. 733

iv. **Novel Test:** We select a novel with nearly 250K
734 context. We test the novel on Summarisation
735 and Question Answering (QA). The Summarisa-
736 tion task aims to evaluate whether selective con-
737 text affects the model’s overall understanding of
738 the input context. The Question Answering task
739 aims to assess the model’s understanding of spe-
740 cific queries. We compare compression time,
741 compression quality, similarity to the original
742 text, and end-to-end time on these tasks. Addi-
743 tionally, we propose the Compression Semantic
744 Efficiency (CSE) metric, calculated through the
745 compression ratio and similarity. 746

747 C Experiment details

748 The experiment used two types of machines, but each
749 experiment was tested on the same machine, using
750 the A800-80GB and 3090ti. We used tiktoken11 and
751 the GPT-3.5-Turbo model to calculate all tokens. We
752 open-sourced an early version of the system preset
753 instructions, which are part of the core intermediate
754 process: Here is a set of rules for simplifying text,
755 aimed at helping users choose the appropriate level
756 of simplification according to different reading and
757 comprehension needs.

758 D Rules

759 D.1 Basic Rules

- 760 1. **Remove Non-Essential Information:**

761
762
763
764
765
766

767

768
769
770
771
772

773
774

775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791

792

793
794
795

796
797
798

799
800

- If a sentence contains two commas or dashes, consider removing the part between them unless it contains essential information.
- Remove all non-essential adjectives and adverbs.

2. Simplify Clauses and Modifiers:

- If there is a restrictive clause following a single comma, consider removing that clause.
- Remove all non-essential attributive, adverbial, and appositive clauses.

D.2 Advanced Rules

1. Handle Complex Relationship Sentences:

- **Contrasting Relationships:** Retain the main information after the contrast.
- **Concessive Relationships:** Retain the crucial part according to contextual importance.
- **Causal Relationships:** Retain the reason explanation.
- **Result Relationships:** Highlight the factors leading to the result.
- **Conditional Relationships:** Retain the condition explanation.
- **Progressive Relationships:** Emphasize the information in the progressive part.
- **Comparative Relationships:** Highlight the main content of the comparison.
- **Coordinate Relationships:** Maintain equal treatment of content.

2. Optional Retention:

- Pay special attention to retaining important information such as names, places, and proper nouns.

D.3 Simplification Levels

1. **Very Light Simplification:** Only remove redundant modifiers.
2. **Light Simplification:** Apply basic comma and clause removal rules.

3. **Moderate Simplification:** Apply all basic rules. 801 802
4. **Deep Simplification:** Apply both basic and advanced rules, retaining key sentence meaning. 803 804
5. **Very Deep Simplification:** Extremely reduce details, retaining only the main parts of the sentence (subject, verb, object). 805 806 807

E Different compression modes

Original Sentence: "The economy, despite facing numerous challenges from external factors such as global market fluctuations and geopolitical tensions, continues to grow."

Very Deep Simplification: "The economy grow."

Deep Simplification: "The economy grows despite challenges."

Moderate Simplification: "The economy grows despite external challenges."

Light Simplification: "The economy, despite challenges, continues to grow."

Very Light Simplification: "The economy, despite facing numerous challenges, continues to grow."

F Cases Study

F.1 The compression ratio of 10X.

Original Prompt :

The author is a Reuters Breakingviews columnist. The opinions expressed are his own. NEWLINE CHAR
NEWLINE CHAR BP faces opposition from some shareholders for handing Chief Executive Bob Dudley a 20 percent increase in his total remuneration package for 2015 to 19.6 million. It may seem hard to square that amount with BP's 5.2 billion loss last year, and the fact that it is slashing thousands of jobs in response to falling oil prices. But that's actually the point. Managing an oil company when crude is trading at 100 per barrel is easy compared to the current environment. Instead, Dudley has to work harder than his predecessors. NEWLINE CHAR NEWLINE CHAR Dudley, whose pay was going to a non-binding shareholder vote on April 14, has done what was needed of him. His two big challenges were to clean up the financial spill from the 2010 Gulf of Mexico disaster and change the culture at BP, which was tainted by safety concerns and excessive risk taking. Last year the company saw the number of recorded oil spills and employee injuries both at five-year lows. NEWLINE CHAR NEWLINE CHAR He has also delivered decent returns when compared to peers. BP ranks third among the big six oil majors, which include Exxon Mobil and Royal Dutch Shell, in total shareholder returns over the last three years, according to Eikon data – even despite 2010's rig blowout. Drawing a line under the environmental catastrophe last year by agreeing to pay up to 18.7 billion in penalties cleared the decks for the company to start rebuilding its balance sheet. NEWLINE CHAR NEWLINE CHAR Compared to counterparts, Dudley's remuneration appears generous. Although Shell Chief Executive Ben van Beurden pocketed 24.2 million euros (27.2 million) in 2014, this figure fell to 5.6 million euros last year, according to the company. Over the same period Dudley's base salary has remained flat, with the biggest boost to his overall financial reward coming through his pension and deferred bonus shares. NEWLINE CHAR NEWLINE CHAR The mild-mannered American has had possibly the toughest job in the oil industry. His rewards look in line with that task.

Compressed Prompt :

BP faces opposition from some shareholders for handing Chief Executive Bob Dudley a 20 percent increase in his total remuneration package for 2015 to 19.6 million. Simplified: BP faces opposition from some shareholders for handing Chief Executive Bob Dudley a 20 percent increase in his total remuneration package for 2015.

Original Prompt :

LONDON – A leading shareholder advisory group has criticized BP PLC’s decision to award its top directors their maximum bonuses for 2015, despite the company’s lackluster performance, and recommended shareholders vote against the payment plans. NEWLINE CHAR NEWLINE CHAR Last month, BP announced that Chief Executive Bob Dudley would receive a 2% bump in his total compensation package in 2015. Though much of this increase related to U.K. reporting requirements that inflated the rise in Mr. Dudley’s pension, the oil executive’s cash bonus increased to 1.4 million from 1 million in 2014. His total bonus for the year, including a portion paid in deferred BP shares, amounted to 4.2 million. That was the maximum amount he was eligible to receive for the year and was up from 3 million in 2014. Chief Financial Officer Brian Gilvary also received 100% of his possible bonus. NEWLINECHAR NEWLINE CHAR The awards follow a year in which the company lost 5.2 billion as oil prices plummeted. Since the start of 2016 it has announced plans to cut 7,000 jobs and has slashed spending to help manage the slump. NEWLINE CHAR NEWLINE CHAR "“We believe shareholders should question whether payouts were fully earned in respect of the past fiscal year relative to the company’s performance,”“ proxy advisory firm Glass Lewis said in a March report seen by The Wall Street Journal. NEWLINE CHAR NEWLINE CHAR BP’s compensation committee awards executive bonuses based on the company’s performance in a number of strategic areas, including its safety record and internal targets for operational cash flow and underlying profits. NEWLINE CHAR NEWLINE CHAR "“BP executives performed strongly in a difficult environment in 2015, managing the things they could control and for which they were accountable,”“ a BP spokesman said, adding that "“safety and operational risk performance was excellent and BP responded quickly and decisively to the drop in oil price.”“ NEWLINE CHAR NEWLINE CHAR This isn’t the first time Glass Lewis has raised objections to BP’s executive pay. Last year, it also recommended that shareholders reject Mr. Dudley’s pay package, noting that his compensation outpaced that received by chief executives at similar-sized firms "“despite the company’s relative underperformance.”“ The executive’s compensation was ultimately approved by around 86% of investors. NEWLINE CHAR NEWLINE CHAR BP’s shareholders will vote on the matter this year at the company’s annual general meeting in London on April 16, along with a host of other issues. Glass Lewis has also raised concerns about the company’s proposal to reduce its notice period for calling a general meeting, but supports most of the proposals, including the re-election of Mr. Dudley and his board. NEWLINE CHAR NEWLINE CHAR Write to Sarah Kent at sarah.kent@wsj.com NEWLINE CHAR NEWLINE CHAR More from MarketWatch

Compressed Prompt :

A leading shareholder advisory group has criticized BP PLC’s decision to award its top directors their maximum bonuses for 2015, and recommended shareholders vote against the payment plans. Simplified: A leading shareholder advisory group has criticized BP PLC’s decision to award its top directors their maximum bonuses for 2015.

Original Prompt :

Angry shareholders mounted an unprecedented protest against BP on Thursday, rebelling against a 20 per cent pay rise for chief executive Bob Dudley despite the oil group making its worst ever loss. NEWLINE CHAR NEWLINE CHAR Investors voted against the company's pay decisions for the first time in living memory, with 59 per cent of proxy votes cast going against BP's decision to pay Mr Dudley nearly 20m for 2015, a year in which the company ran up a 5.2bn loss. NEWLINE CHAR NEWLINE CHAR It was the first time that a top British company was defeated over executive pay since shareholders at advertising group WPP and Xstrata, the mining company, rebelled four years ago during what was dubbed the "shareholder spring". It left BP scrambling to win back support of some of the City's biggest institutions. NEWLINE CHAR NEWLINE CHAR The rebellion highlighted a growing trend of institutional investors and advisers around the world taking a more aggressive stance over pay. NEWLINE CHAR NEWLINE CHAR Smith Nephew, the FTSE 100 medical devices group, also suffered a defeat on their remuneration report on Thursday as 53 per cent of shareholders voted against the pay package of chief executive Olivier Bohuon. Although Mr Bohuon's overall pay fell to 5.5m in 2015 compared with 6.8m in 2014, shareholders protested because the company allowed long-term incentives to vest despite falling below initial targets. NEWLINE CHAR NEWLINE CHAR US banks from Citigroup to Bank of America have faced pressure to toughen bonus "clawback" regimes, which put executives on the hook for future losses. A resolution demanding more details of JPMorgan's clawback plans attracted 44 per cent support last year. NEWLINE CHAR NEWLINE CHAR Mr Dudley's pay looked particularly out of line to shareholders because other major energy company bosses took pay cuts in 2015, a year when energy companies were hit hard by the oil price crash. NEWLINE CHAR NEWLINE CHAR According to ISS Corporate Solutions in the US, the median pay of an S P 500 energy company chief executive, excluding their pension, fell by 1.8 per cent last year after four years of increases that ranged from 4.8 to 8.2 per cent..... (Omit here)

Compressed Prompt :

Angry shareholders mounted an unprecedented protest against BP on Thursday, rebelling against a 20 per cent pay rise for chief executive Bob Dudley despite the oil group making its worst ever loss. Simplified: Angry shareholders mounted an unprecedented protest against BP on Thursday, rebelling against a 20 per cent pay rise for chief executive Bob Dudley.

Original Prompt :

Image copyright PA Image caption Bob Dudley took over as BP chief executive in the aftermath of the fatal Gulf of Mexico oil rig explosion NEWLINE CHAR NEWLINE CHAR BP shareholders have rejected a pay package of almost £14m for chief executive Bob Dudley at the oil company's annual general meeting. NEWLINE CHAR NEWLINE CHAR Just over 59% of investors rejected Mr Dudley's 20% increase, one of the largest rejections to date of a corporate pay deal in the UK. NEWLINE CHAR NEWLINE CHAR The vote is non-binding on BP, but earlier, chairman Carl-Henric Svanberg promised to review future pay terms. NEWLINE CHAR NEWLINE CHAR Mr Dudley received the rise despite BP's falling profits and job cuts. NEWLINE CHAR NEWLINE CHAR Corporate governance adviser Manifest says the vote is at or above the fifth-largest in the UK against a boardroom remuneration deal. NEWLINE CHAR NEWLINE CHAR 'Last chance saloon' NEWLINE CHAR NEWLINE CHAR In his opening address to the shareholders' meeting, before the vote had been formally announced, Mr Svanberg acknowledged the strength of feeling, saying: ""Let me be clear. We hear you."" NEWLINE CHAR NEWLINE CHAR He continued: ""We will sit down with our largest shareholders to make sure we understand their concerns and return to seek your support for a renewed policy."" NEWLINE CHAR NEWLINE CHAR ""We know already from the proxies received and conversations with our institutional investors that there is real concern over the directors' pay in this challenging year for our shareholders. NEWLINE CHAR NEWLINE CHAR ""On remuneration, the shareholders' reactions are very strong. They are seeking change in the way we should approach this in the future,"" he said. NEWLINE CHAR NEWLINE CHAR The Institute of Directors said the shareholder rebellion would ""determine the future of corporate governance in the UK"". NEWLINE CHAR NEWLINE CHAR ""British boards are now in the last chance saloon, if the will of shareholders in cases like this is ignored, it will only be a matter of time before the government introduces tougher regulations on executive pay,"" said director general Simon Walker. NEWLINE CHAR NEWLINE CHAR Media playback is unsupported on your device Media caption Dudley's pay sends 'wrong message' investor says NEWLINE CHAR NEWLINE CHAR 'Out of touch' NEWLINE CHAR NEWLINE CHAR Shareholders that criticised the pay deals included Aberdeen Asset Management and Royal London Asset Management. NEWLINE CHAR NEWLINE CHAR Investor group Sharesoc branded the pay deal ""simply too high"", while Glass Lewis, ShareSoc, Pirc and Institutional Shareholder Services have also expressed their opposition. NEWLINE CHAR NEWLINE CHAR Earlier on Thursday, Ashley Hamilton Claxton, corporate governance manager at Royal London, told the BBC: ""The executives received the maximum bonuses possible in a year when [BP] made a record loss, and to us that just does not translate into very good decision-making by the board. NEWLINE CHAR NEWLINE CHAR ""We think it sends the wrong message. It shows that the board is out of touch."" NEWLINE CHAR NEWLINE CHAR She told the BBC's Today programme that if 20%-25% of shareholders vote down the pay deal, it would force BP to ""think long and hard about their decision"". NEWLINE CHAR NEWLINE CHAR The early voting figures suggest that the opposition is even bigger than she expected. (Omit here)

Compressed Prompt :

BP shareholders have rejected a pay package of almost £14m for chief executive Bob Dudley at the oil company's annual general meeting. Simplified: BP shareholders have rejected a pay package of almost £14m for chief executive Bob Dudley.

Original Prompt :

A majority of BP PLC's shareholders voted against the company's executive pay policy, a stinging — though nonbinding — rebuke to Chief Executive Bob Dudley and his board. NEWLINE CHAR
 NEWLINE CHAR At the company's annual meeting Thursday, the oil giant said preliminary results showed 59 % of investors voting by proxy rejected the company's executive compensation decisions for 2015. That included a controversial 20 % increase in Dudley's total pay for the year, at a time when the company lost 5.2 billion. NEWLINE CHAR NEWLINE CHAR Earlier in the day, the company also signaled in its clearest terms yet that the oil giant may have to reduce its dividend, as low oil prices continue to threaten the once-sacrosanct investor payouts across the industry. NEWLINE CHAR NEWLINE CHAR Both moves heap pressure on Dudley and his board, as they try to navigate low oil prices like the rest of the industry but also contend with increasing shareholder unease. NEWLINE CHAR NEWLINE CHAR BP BP., +0.88 % BP, +0.60 % Chairman Carl-Henric Svanberg, speaking to investors before the vote, defended the pay package, which he said was based on "exceptional" company performance during a difficult year. He said, before the vote, that the board would discuss possible changes to its compensation plan for next year. NEWLINE CHAR NEWLINE CHAR After the vote, Svanberg said that despite the nonbinding vote, the company wouldn't adjust Dudley's pay. NEWLINE CHAR NEWLINE CHAR An expanded version of this report appears on WSJ.com NEWLINE CHAR NEWLINE CHAR More from MarketWatch "

Compressed Prompt :

A majority of BP PLC's shareholders voted against the company's executive pay policy, a stinging — though nonbinding — rebuke to Chief Executive Bob Dudley and his board. Simplified: A majority of BP PLC's shareholders voted against the company's executive pay policy.

816

Original Prompt :

Item 15, report from City Manager Recommendation to adopt three resolutions. First, to join the Victory Pace program. Second, to join the California first program. And number three, consenting to to inclusion of certain properties within the jurisdiction in the California Hero program. It was emotion, motion, a second and public comment. CNN. Please cast your vote. Oh. Was your public comment? Yeah. Please come forward. I thank you, Mr. Mayor. Thank you. Members of the council. My name is Alex Mitchell. I represent the hero program. Just wanted to let you know that the hero program. Has been in California for the last three and a half years. We're in. Over 20. We're in 28 counties, and we've completed over 29,000 energy efficient projects to make homes. Greener and more energy efficient. And this includes anything. From solar to water. Efficiency. We've done. Almost.\$ 550 million in home improvements.

Compressed Prompt :

Item 15 report City Manager Recommendation adopt three resolutions. join Victory Pace program. Second join California first program. three inclusion properties jurisdiction California Hero program. emotion motion second public comment. CNN. cast vote. public comment?. come forwar. Members council. Alex Mitchell. represent hero program. California three and a half years. 20. 28 counties completed 29,000 energy efficient projects. includes. solar water. Efficiency Almost \$550 million home improvements.

817

F.2 Single-document QA in 5x constraint

Compressed Prompt:

banned. In Roth, the Court ruled that obscenity was not protected by the First Amendment, but did not provide a clear definition of what constituted obscenity. This led to the Memoirs test, which required that material be ""utterly without redeeming social value"" to be considered obscene. However, this test was difficult to apply and led to inconsistent rulings across different jurisdictions.

The Miller decision In Miller v. California, the Supreme Court modified its definition of obscenity, creating the three-prong Miller test. Under this test, material is considered obscene if:

The average person, applying contemporary community standards, would find that the work, taken as a whole, appeals to the prurient interest. The work depicts or describes, in a patently offensive way, sexual conduct specifically defined by the applicable state law. The work, taken as a whole, lacks serious literary, artistic, political, or scientific value. This new test allowed for more local control over obscenity laws, as it required that the material be judged based on contemporary community standards. It also provided a clearer definition of what constituted obscenity, allowing for more consistent rulings across different jurisdictions.

Impact The Miller decision had a significant impact on obscenity laws in the United States. It allowed for more local control over what was considered obscene, as it required that material be judged based on contemporary community standards. It also provided a clearer definition of obscenity, allowing for more consistent rulings across different jurisdictions. The Miller test is still used today to determine what constitutes obscenity under the First Amendment."

Question: Which case was brought to court first Miller v. California or Gates v. Collier ?

LLMs' Response:

Miller v. California

Ground Truth:

Miller v. California

F.3 QA on LongBench with 2X constraints

Question Prompt :

The Little Drummer Boy was followed by another special in 1969, Frosty the Snowman, based on the song of the same name, with narration by the American comedian Jimmy Durante and animation by Mushi Production. The success of the special led to a sequel, Frosty's Winter Wonderland, in 1976, as well as a 1979 animated adaptation of the 1957 song ""Rudolph the Red-Nosed Reindeer"" by Johnny Marks, ""Rudolph and Frosty's Christmas in July"". In 1970, Rankin/Bass produced Santa Claus Is Comin' to Town, based on the song of the same name and starring the American actor Fred Astaire as the narrator and the voice of the mailman, Mickey Rooney as the voice of Kris Kringle/Santa Claus, and Keenan Wynn as the voice of the Winter Warlock. This was followed by The Year Without a Santa Claus in 1974, based on the 1956 book of the same name by Phyllis McGinley, with narration by the American actor Shirley Booth and the voices of Mickey Rooney as Kris Kringle/Santa Claus and Dick Shawn as the voice of Snow Miser. The success of the special led to a live-action remake in 2006, written by Larry Wilson, and a sequel, A Miser Brothers' Christmas, in 2008. In 1977, Rankin/Bass produced The Easter Bunny is Comin' to Town, narrated by the American actor Fred Astaire and starring the voices of Skip Hinnant, Vincent Price, and Robert Morse. The studio's last major holiday special was Jack Frost in 1979, narrated by the American actor Buddy Hackett and starring the voices of Robert Morse, Debra Clinger, and Paul Frees. After Jack Frost, Rankin/Bass continued to produce animated television specials and series until 1987. In addition to their holiday specials, Rankin/Bass also produced several non-holiday themed animated series, including The Jackson 5ive and The Osmonds, as well as the animated versions of The King Kong Show and The Smokey Bear Show. They also produced several animated feature films, including The Last Dinosaur and The Bermuda Depths. In 1987, Rankin/Bass was acquired by Lorimar-Telepictures, which was later acquired by Warner Bros. in 1989. The studio continued to produce animated specials and series until 1994, when it was shut down. In 2001, Warner Bros. released a DVD box set of the studio's holiday specials, and in 2018, Warner Bros. announced plans to produce new films and television specials based on the studio's classic characters and properties." Question:For what type of work is the production company for The Year Without a Santa Claus best known?

LLMs' Response:

seasonal television specials, usually done in stop motion animation

Ground Truth:

seasonal television specials, particularly its work in stop motion animation

F.4 Sum. on LongBench with 2X constraints

Question: Prompt :

Project Manager: I hope you're all prepared for this feature design meeting.

Marketing Department: Of course.

Project Manager: I'll be taking notes, and the three of you will be presenting. We need to understand the requirements for the new project, including user needs, technical implementation, and design approach.

Marketing Department: I'm the marketing expert, Sammy Benjo, and I'd like to share insights on user needs and preferences for the remote control. We conducted a survey and found that users dislike the current remote's appearance and complexity of operation. They want a more aesthetically pleasing and user-friendly remote.

Project Manager: Next up.

Industrial Designer: Please press F5 for fullscreen display.

Marketing Department: I suggest we ask users about their feelings and requirements for the current remote. The survey shows users dislike the appearance and complexity of operation of the remote and hope for improvements in these aspects.

Project Manager: Alright, what's next?

Industrial Designer: The survey indicates that users' main complaints about the remote include difficulty in finding it, too many and unused buttons, among others.

Project Manager: I agree, these are all issues that need addressing.

Marketing Department: Users want the remote to have voice recognition capabilities, but there are significant differences in demand across different age groups.

Summarize the discussion about the working design of the remote control.

LLMs' Response:

The discussion about the working design of the remote control focused on several key aspects. Firstly, the team talked about the importance of understanding user needs, with Marketing highlighting users' preferences for simplicity, aesthetics, and functionality, such as easy zapping and a desire for fewer, more relevant buttons. The Industrial Designer discussed the technical components necessary for the remote, including a power source (battery), user interface (keypad/buttons), a digital signal processing chip for handling functions, and an infrared LED source for communication with devices. User Interface emphasized the need for a simple, intuitive design that could potentially include features like speech recognition for ease of use and automatic volume adjustment based on the user's distance from the TV. The team also considered making the remote universal for various devices, not just TVs, and discussed incorporating feedback mechanisms and possibly a screen for better interaction. The idea of speech recognition was debated, with concerns about complexity versus the potential appeal to younger users who might appreciate personalized, voice-activated controls.

Ground Truth:

The industrial designer described the product as a portable device with some energy source, an interface which included some pressing buttons and keys, a processor, some infrared source which could send the information to the main device, and different encryption codes for different TVs. Since the Marketing and the user interface designer mentioned speech recognition, the industrial designer also took that into account. All these components would cost 5 Euros or even less than that. As for gesture recognition and automatic power control, the group gave them up due to price issues. At the end of the meeting, the group also talked about how to enable the users to find the controller.

F.5 Sum. on LongBench with 2X constraints

Compressed Prompt :

"MOTHER, have you heard about our summer holidays yet?" said Julian, at the breakfast-table. "Can we go to Polseath as usual?" "I'm afraid not," said his mother. "They are quite full up this year." "The three children at the breakfast-table looked at one another in great disappointment. They did solove the house at Polseath. The beach was so lovely there, too, and the bathing was fine. "Cheer up," said Daddy. "I dare say we'll find somewhere else just as good for you. And anyway, Mother and I won't be able to go with you this year. Has Mother told you?" "No!" said Anne. "Oh, Mother is it true? Can't you really come with us on our holidays? You always do." "Well, this time Daddy wants me to go to Scotland with him," said Mother. "All by ourselves! And as you are really getting big enough to look after yourselves now, we thought it would be rather fun for you to have a holiday on your own too. But now that you can't go to Polseath, I don't really quite know where to send you." "What about Quentin's?" suddenly said Daddy. Quentin was his brother, the children's uncle. They had only seen him once, and had been rather frightened of him. He was a very tall, frowning man, a clever scientist who spent all his time studying. He lived by the sea but that was about all that the children knew of him! "Quentin?" said Mother, pursing up her lips. "Whatever made you think of him? I shouldn't think he'd want the children messing about in his little house." "Well," said Daddy, "I had to see Quentin's wife in town the other day, about a business matter and I don't think things are going too well for them. Fanny said that she would be quite glad if she could hear of one or two people to live with her for a while, to bring a little money in. Their house is by the sea, you know. It might be just the thing for the children. Fanny is very nice she would look after them well." "Yes and she has a child of her own too, hasn't she?" said the children's mother. "Let me see what's her name something funny yes, Georgina! How old would she be? About eleven, I should think." "2 "Same age as me," said Dick. "Fancy having a cousin we've never seen! She must be jolly lonely all by herself. I've got Julian and Anne to play with but Georgina is just one on her own. I should think she'd be glad to see us." "Well, your Aunt Fanny said that her Georgina would love a bit of company," said Daddy. "You know, I really think that would solve our difficulty, if we telephone to Fanny and arrange for the children to go there. It would help Fanny,

Compressed Prompt:

"MOTHER have you heard about our summer holidays yet?" said Julian at the breakfast-table. "I'm afraid not," said his mother. "Cheer up," said Daddy. "No!" said Anne. "Well, this time Daddy wants me to go to Scotland with him," said Mother. "What about Quentin's?" suddenly said Daddy. "Quentin?" said Mother. "Well," said Daddy, "I had to see Quentin's wife in town the other day." "Yes and she has a child of her own too, hasn't she?" said the children's mother. "Same age as me," said Dick. "Quentin?" said Mother, pursing up her lips. "Well, your Aunt Fanny said that her Georgina would love a bit of company," said Daddy. "Yes and she will love looking after you all," said Daddy. "Well, that's settled," he said. "Next week, if Mother can manage it," said Daddy. "Yes," she said. "How lovely it will be to wear shorts again," said Anne. "Well, you'll soon be doing it," said Mother. "Anne wanted to take all her fifteen dolls with her last year," said Dick. "No, I wasn't," said Anne. "Daddy, are we going by train or by car?" he asked. "By car," said Daddy. "That would suit me well," said Mother. "So Tuesday it was," said Mother. "It's a lovely day, hurrah!" cried Julian. "It's come at last!" she said. "Are we picnicking soon?" asked Anne. "Yes," said Mother. "Oh, gracious!" said Anne. "What time shall we be at Aunt Fanny's?" asked Julian. "About six o'clock with luck," said Daddy. "We must watch out for the sea," said Dick.