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. Additionally, we adopt a chunk-based compression algorithm to achieve adjustable compression rates. We evaluate our method on multiple datasets, including LongBench, ZeroScrolls, 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. Code is available at `hppts:github.com/LanguaShrink`.

1 Introduction

In recent years, the field of large language models (LLM) has seen the emergence of various prompting techniques, such as Chain of Thought (CoT) (Wei et al., 2022a), In-context Learning (ICL) (Dong et al., 2022), and Retrieval Augmented Generation (RAG) (Lewis et al., 2020). These techniques have

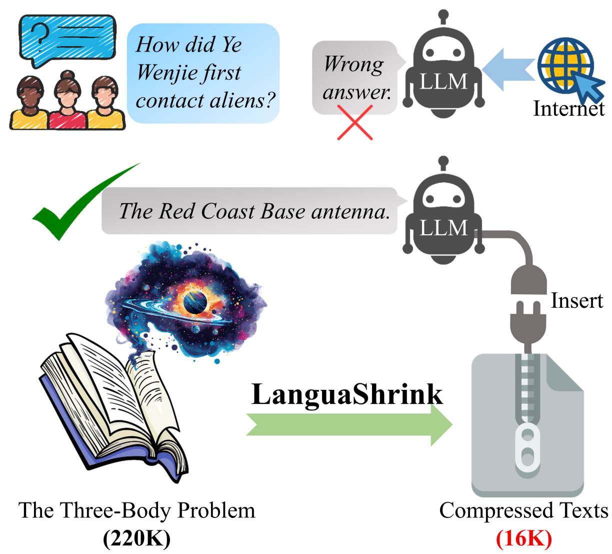

Figure 1: Illustration of the Plug-and-Play Document Module. The document encoding is decoupled from specific tasks. By inserting the document plugin into the task model, we can separate compressed text from downstream task reasoning and reduce computational costs.

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 learning 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 compression system grounded in psycholinguistics and the Ebbinghaus memory curve to highlight critical information in long texts.
- We introduce a data distillation approach that uses smaller models to learn the compression target, optimizing training through a KL-regularized reinforcement learning framework.
- We perform extensive experiments across multiple datasets, demonstrating that our method achieves up to 26x compression without significant performance loss.
- We provide the theoretical analysis and key insights underlying our method (in Section A). These results offer a rigorous foundation that supports the effectiveness and applicability of our framework design.

2 Related work

2.1 Psycholinguistics

Psycholinguistic research primarily examines two areas: sentence processing and text processing (McKoon and Ratcliff, 1998). Sentence processing focuses on how syntactic structures are computed (Alyahya et al., 2018), whereas text processing covers the understanding of larger text units. Function words and key nouns play a significant role in these processes (Kalyuga, 2012).

Previous studies show that removing redundant information can improve the efficiency of foreign-language vocabulary learning (Ellis and Beaton, 1993). Additionally, this approach also helps optimize storage space (Schmidhuber, 2000). Motivated by these findings, we propose a psycholinguistics-based Part-of-Speech Priority Compression (PPC)

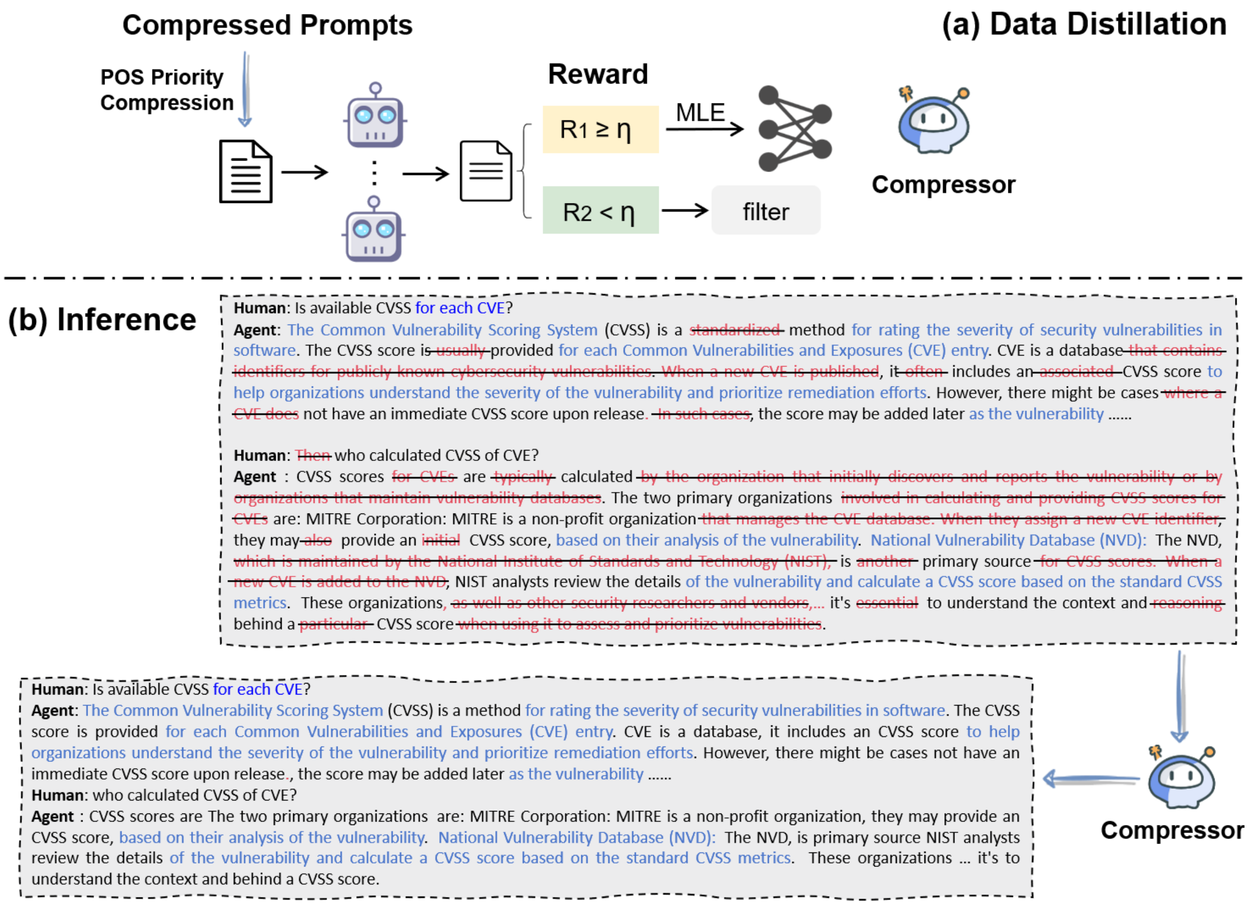

Figure 2: (a) **Data distillation.** Initial text compression is first performed using part-of-speech (POS) priority. The compressed prompts are then evaluated for similarity and compression ratio relative to the original prompt. If the similarity surpasses the threshold, the model receives a positive reward; otherwise, the reward is zero and the example is discarded. Afterward, the model is fine-tuned using Maximum Likelihood Estimation (MLE), and the compressor produces the final compressed prompts. (b) **Inference.** This figure illustrates how the compressor is employed in real-world question-answering tasks. The LanguaShrink module processes the original dialogue, where red text shows portions most likely to be compressed and blue text denotes the next most likely segments for compression.

algorithm that uses lexical classification and priority assignment to better retain core information while eliminating redundant content (Graça et al., 2011).

2.2 Prompt Compression

LLMs encounter considerable difficulties when required to manage long contexts. Due to the attention mechanism's quadratic growth in memory and computational expense, processing long text involves very high costs (Han et al., 2023; Zhuang et al., 2022; Chen et al., 2023). Many LLMs also rely on a fixed context window during pre-training, further restricting their handling of extended contexts. Researchers have explored sparse attention and local dense attention as possible ways to reduce computational and memory overhead. Additionally, soft prompt tuning and reinforcement learning-based compression methods have been investigated to save context costs at inference time (Shen et al., 2018; Liu et al., 2023).

Prompt compression offers a direct approach to address the challenge of handling long contexts by shortening prompt length while retaining crucial information. Existing methods are commonly categorized as task-aware or task-agnostic. Task-aware compression modifies context based on downstream tasks or current queries (e.g., LongLLMLingua (Jiang et al., 2023), which adjusts the compression ratio via token information entropy). Task-agnostic compression has broader applicability, typically employing information entropy to prune redundant tokens from the prompt. Although these techniques have improved model performance and efficiency, further improvements are necessary for managing long texts and complex tasks in real-world scenarios (Hsieh et al., 2023).

OpenChat (Wang et al., 2023) introduced a prompt compression fine-tuning mechanism that uses reinforcement learning to dynamically adjust prompts, retaining only the most important information for a given task and reducing unnecessary computational load. KL divergence measures changes in prompt data before and after compression; by minimizing KL divergence, crucial information from the original prompt is preserved.

In this paper, we propose the Prompt Compression Fine-Tuning (PC-RLFT) method, which merges chunk compression strategies with reinforcement learning to achieve effective prompt compression and retention of core information.

3 Method

LanguaShrink compresses input prompts T into shorter versions T', minimizing token count $|T'|$ while preserving semantic integrity $S(T, T')$. The framework prioritizes high-information tokens using part-of-speech weights w_{POS}, followed by dataset distillation where smaller models M_{small} learn to focus on the most relevant content. Reinforcement learning with policy π_θ adjusts the compression strategy, balancing compression ratio r_c and semantic similarity $S(T, T')$. Finally, prompts are divided into chunks C_i and compressed independently based on their importance $w(C_i)$, achieving efficient compression without compromising output quality.

3.1 Theoretical Foundations

The framework is grounded in two key principles: psycholinguistic information density and the Ebbinghaus memory curve. High-information tokens such as nouns and verbs are assigned larger weights w_{POS} to preserve the central meaning of T. The memory curve indicates that items at the beginning and end of a sequence are more easily recalled, guiding the selection of content by position. As a result, each chunk C_i in the compressed prompt T' sustains its essential semantic content and positional importance, allowing for effective compression.

3.2 Information Selection and Prioritization

The process of selecting and prioritizing tokens begins with Part-of-Speech (POS) analysis (Toutanova et al., 2003), where each token in the prompt T is assigned a weight w_{POS}. Tokens like nouns and verbs, which carry higher information density, receive greater importance, while lower-density tokens such as adjectives and adverbs are deprioritized. After the POS-based weighting, the framework distills the prompt through smaller models M_{small}, which learn to compress T into T' by focusing on these weighted tokens. This selection mechanism enables the preservation of key content while reducing unnecessary token overhead.

3.3 POS Priority Compression

Part-of-Speech Priority Compression (PPC) is achieved by inputting carefully designed prompts into LLMs. To implement PPC, we need to design a series of specific prompts that achieve part-of-speech priority compression through the CoT approach (Wei et al., 2022b). Below is the design idea for CoT:

Relation Word Extraction. We engineer prompts to instruct the model to detect relational terms in the text. Through dependency syntax analysis, the model captures how sentences are connected. The model then assigns different priority levels to these relational terms according to the overall sentence context.

Part-of-Speech Classification. These prompts direct the model to classify each token by its part of speech (e.g., adjective, adverb, noun, preposition), guided by psycholinguistic POS analysis. Each part of speech is given a priority value, for instance, nouns > verbs > adjectives > adverbs.

Priority Filtering. After relation word extraction and POS classification are finished, the model applies these details to filter out words and sentences that substantially contribute to the passage's main meaning, deleting lower-priority words and sentences that minimally affect overall comprehension.

3.4 Dataset Distillation

We propose a data distillation method that extracts knowledge from large language models (LLMs) to generate compressed prompts that retain key information while reducing latency by using smaller models to learn the compression targets. Additionally, we ensure the compressed prompts remain highly faithful to the original content.

Dataset: We source our data from reading materials in the Chinese Gaokao (Zhang et al., 2023;

Algorithm 1 Compression Algorithm

Require: T: input text
  $C \leftarrow SplitToChunks(T)$
  $n \leftarrow ChunkCount(C)$
  for $i \leftarrow 1$ to n **do**
    $c_i \leftarrow C[i]$
    $TokenCompression(c_i)$
  end for
  $T' \leftarrow JoinChunks(C)$
  return T'

Algorithm 2 Chunk-Based Compression Algorithm

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

Zhong et al., 2023; Sun et al., 2021) 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 formed by splitting the reading passages into blocks of three consecutive sentences. This structure allows the model to learn effective compression while retaining critical contextual information.

To generate compressed data, we use various LLMs, including both open-source and proprietary models such as GPT-4 (OpenAI, 2023), Yi (01.AI, 2023), GLM (Zeng et al., 2023), and Qwen (Cloud, 2023). These models distill knowledge under psycholinguistic principles, creating compressed prompts that preserve subject-verb-object structures and capture essential content. We ensure that the compressed versions retain high semantic similarity to the original text. By synthesizing complementary strengths from multiple LLMs, we gain a well-rounded compression strategy and better generalization.

3.5 Prompt Compress-RLFT

3.5.1 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.

3.5.2 Tuning

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

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

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:

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

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.

3.6 Chunk-Based Compression

For each chunk $C[i]$, the algorithm evaluates its relevance to the query Q by calculating the cosine similarity $\mathrm{CS}(C[i], Q)$. The relevance is denoted as rel_i. The semantic and structural importance of the chunk is calculated using the function $\mathrm{CI}(C[i])$, 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 $\mathrm{CP}(C[i])$.

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

$$w_i = \alpha \times \mathrm{rel}_i + \beta \times \mathrm{imp}_i \quad (4)$$

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 apply the PPC approach to analyze text and generate a large compressed dataset. Using the PC-RLFT method, we train a smaller model on these data. For all reported metrics, GPT-3.5 (OpenAI, 2023) is the target LLM for downstream tasks. In every experiment, we use greedy decoding with a temperature of 0 to stabilize the generated outputs. Our smaller pre-trained language model for compression is Qwen-1.8B (Cloud, 2023).

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., 2024, 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 shows the performance of several methods under different compression constraints. Although our compression model is much smaller than LLama-2-7B (Touvron et al., 2023) or other baseline models, it achieves better QA and synthesis outcomes. Relative to the original prompts, our compressed prompts maintain comparable performance at a reduced cost. Compared with other task-agnostic baselines, our method performs at a higher level, confirming the impact of our constructed dataset and underlining the benefits of optimizing compression models with prompt compression knowledge.

Compression-based methods, such as selective context (Chevalier et al., 2023) and LLMLingua (Jiang et al., 2023), often show weak results on most tasks. Their compression mechanisms rely exclusively on information entropy, which can lead to higher noise levels in the compressed text. Retrieval-based methods select fragments that most closely match a query, but in practice these fragments may still contain substantial redundant information, lowering the overall information density.

We observe that mathematical performance decreases when tokens are removed, possibly due to the lower sensitivity of psycholinguistic models to mathematical content. On the LongBench test (Bai et al., 2023), LLMLingua2 (Pan et al., 2024) has a

Methods	LongBench									ZeroSCROLLS		
	SingleDoc	MultiDoc	Summ.	FewShot	Synth.	Code	AVG	Tokens	$1/\tau$	AVG	Tokens	$1/\tau$
	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	37.8	41.7	**26.9**	64.3	**53.0**	52.4	46.0	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	64.8	22.3	56.2	38.2	1,891	5x	33.3	1,862	5x
LLMLingua2	29.8	33.1	25.3	**66.4**	21.3	**58.9**	39.1	1,954	5x	33.3	1,898	5x
LanguaShrink	**42.1**	**54.3**	26.3	62.3	33.0	58.4	**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	26.8	63.4	49.6	37.6	41.7	3,421	3x	22.4	3,362	3x
LongLLMLingua	37.6	42.9	**26.9**	68.2	49.9	53.4	46.5	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	60.0	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	35.5	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 1: Performance of different methods under different compression ratios on LongBench (Bai et al., 2023) and ZeroSCROLLS (Shaham et al., 2023) using GPT-3.5-Turbo.

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 2: Statistics on Arxiv Articles. "CSE" refers to Compression Semantic Efficiency; BLEU and ROUGE are measured against the uncompressed text.

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 3: Performance across different variants of LanguaShrink. "Tokens (avg)" and "Time (avg)" are mean values per sample. "(w/o psy.)" removes psycholinguistic weighting, and "(w/o SA)" omits syntactic analysis.

small advantage in few-shot and code tasks, whereas LanguaShrink excels in text and QA compression tasks. While psycholinguistic strategies offer significant strengths for text processing, they may not provide sufficient support for math-oriented content.

In the ArXiv tests, detailed in Table 2, our method demonstrates strong results in Compression Semantic Efficiency (CSE). Specifically, we outperform llmlingua2 by a factor of 2.46. We also see a notable rise in BLEU scores (Papineni et al., 2002) compared to Select context and llmlingua2. With ROUGE (Lin, 2004), our method remains on par with other approaches, reflecting its focus on retaining core semantics while allowing sentence structure changes.

4.3 Ablation Study

Our method has two core elements: a psycholinguistic analysis module and a sentence analysis module. From Table 3, removing the psycholinguistic module decreases compression capacity by almost 10%, largely because losing linguistic analysis reduces the ability to locate tokens efficiently. Eliminating the sentence analysis module restores some token-

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

Table 4: F1 Scores of Different Methods

Method	Tokens (avg)	CSE	BLEU
llmlingua2	280.10	0.9419	0.0288
our	253.15	1.0462	0.0304
Select context	270.15	0.9413	0.0273
original	343.65	-	-

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

location ability but lowers compression performance and efficiency because the system cannot focus on the most critical parts of sentences.

In practical test scenarios, LanguaShrink uses a semantic-retention strategy that caps the achievable compression rate at around 90% in standard mode. For even higher compression rates, we offer a performance mode that retains about 20% of the semantic content at a 96% compression rate.

Table 4 shows results from a novel dataset where questions and reference answers are generated from the original text. We then request an LLM to answer the same questions. Even without the sentence analysis module, LanguaShrink's psycholinguistic approach performs on par with llmlingua2.

We propose the CSE metric to gauge how effectively tokens are compressed without losing meaning. If CSE is below 1, the compression method reduces token performance significantly. A CSE above 1 indicates that the compression method can still boost effective context size. In evaluations with already brief text, compression may harm CSE because there is little to compress. As seen in Table 5, none of the three models achieve the specified compression ratio (around 30%). Nevertheless, our method attains a CSE above 1, while the others remain below 1, indicating that our approach remains robust even at higher compression rates.

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

Table 6: Latency and Speedup Factor of Different Methods

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 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

We propose **LanguaShrink**, a prompt compression framework that reduces prompt length for LLMs while retaining essential information. It integrates psycholinguistic priors and the Ebbinghaus memory curve to enable task-agnostic, model-compatible compression. LanguaShrink combines part-of-speech–based priority selection, data distillation with smaller models, and KL-regularized reinforcement learning. A chunk-wise compression strategy ranks segments by relevance, importance, and perplexity to support controllable compression rates. Experiments show that LanguaShrink achieves up to 26× compression with minimal performance loss, outperforming existing methods in semantic similarity and efficiency across multiple datasets.

6 Limitations

Currently, our token compression technology mainly includes psycholinguistic techniques and does not incorporate RAG (Retrieval-Augmented Generation). In initial experiments, we tested a range of psycholinguistic approaches and ultimately chose two that produced the best results. We did not fully apply all psycholinguistic methods. In these early trials, we partially used the Oxford Dictionary for training. Although this yielded promising outcomes, comprehensive testing was not possible because we did not have the complete Oxford Dictionary content.

LanguaShrink can become unstable if the compression rate exceeds 90%. Although we provide an extreme mode to address this problem, it is not a preferred long-term solution. While the extreme mode can temporarily offset performance loss from high compression rates, it may introduce other complexities and resource costs in practical scenarios.

Future work will focus on refining the application of psycholinguistic methods, exploring more diverse integration strategies, and addressing reduced mathematical performance. Our aim is to make further progress in a broader range of use cases.

Ethics Statement

The development and application of LanguaShrink also raise several ethical considerations: Bias and Fairness: The datasets used for training and evaluating LanguaShrink must be carefully curated to ensure they are representative and do not perpetuate biases. Any inherent biases in the data could be amplified through the compression process, leading to unfair or biased outputs from the LLMs. Privacy and Confidentiality: When applying LanguaShrink to sensitive or confidential information, it is crucial to ensure that the compression process does not inadvertently expose or compromise any personal or sensitive data. Robust data handling and privacy-preserving techniques must be implemented. Transparency and Accountability: The use of LanguaShrink should be transparent, with clear documentation on how the compression is performed and its potential impacts on the data. Users should be informed about the limitations and potential risks associated with the compressed prompts to make informed decisions about their use. Impact on Employment: The efficiency 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

01.AI. 2023. Yi large language model. https://www.lingyiwanwu.com/. Accessed: 2025-05-19.

Reem S. W. Alyahya, A. Halai, Paul Conroy, and M. L. Lambon Ralph. 2018. Noun and verb processing in aphasia: Behavioural profiles and neural correlates.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. Longbench: A bilingual, multitask benchmark for long context understanding.

Kai Cao, Qizhong Wu, Ling-Yu Wang, Nan Wang, Huaqiong Cheng, Xiao Tang, Dongqing Li, and Lanning Wang. 2023. Gpu-hadvppm v1.0: a high-efficiency parallel gpu design of the piecewise parabolic method (ppm) for horizontal advection in an air quality model (camx v6.10).

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models.

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

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset. *Preprint*, arXiv:1905.00075.

Alibaba Cloud. 2023. Qwen: Large language model from alibaba. https://tongyi.aliyun.com/qianwen/. Accessed: 2025-05-19.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A survey for in-context learning. *Preprint*, arXiv:2301.00234.

N. Ellis and A. Beaton. 1993. Psycholinguistic determinants of foreign language vocabulary learning.

João Graça, Kuzman Ganchev, Luísa Coheur, Fernando C Pereira, and B. Taskar. 2011. Controlling complexity in part-of-speech induction.

Insu Han, Rajesh Jayaram, Amin Karbasi, V. Mirrokni, David P. Woodruff, and A. Zandieh. 2023. Hyperattention: Long-context attention in near-linear time.

Cho-Jui Hsieh, Si Si, Felix X. Yu, and I. Dhillon. 2023. Automatic engineering of long prompts.

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.

Tian jian Luo, Yuqi Liu, and Tianning Li. 2022. A multi-feature fusion method with attention mechanism for long text classification.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. Llmlingua: Compressing prompts for accelerated inference of large language models. https://github.com/microsoft/LLMLingua. Accessed: 2025-05-19.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression.

Slava Kalyuga. 2012. Cognitive load aspects of text processing.

T. Kuvshinova and A. Khritankov. 2019. Improving a language model evaluator for sentence compression without reinforcement learning.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Filippo Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Yucheng Li. 2023. Unlocking context constraints of llms: Enhancing context efficiency of llms with self-information-based content filtering. *Preprint*, arXiv:2304.12102.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Ré, and Beidi Chen. 2023. Deja vu: Contextual sparsity for efficient llms at inference time.

Xinyin Ma, Yongliang Shen, Gongfan Fang, Chen Chen, Chenghao Jia, and Weiming Lu. 2020. Adversarial self-supervised data free distillation for text classification.

Supun Manathunga and Isuru Hettigoda. 2023. Aligning large language models for clinical tasks.

G. McKoon and R. Ratcliff. 1998. Memory-based language processing: psycholinguistic research in the 1990s.

Jesse Mu, Xiang Lisa Li, and Noah D. Goodman. 2023. Learning to compress prompts with gist tokens.

J. Murre and Joeri Dros. 2015. Replication and analysis of ebbinghaus' forgetting curve.

OpenAI. 2023. GPT-4 technical report. https://openai.com/research/gpt-4. Accessed: 2025-05-19.

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

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 Wettig, Anirudh Ajith, Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, and 17 others. 2020. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

J. Schmidhuber. 2000. Neural predictors for detecting and removing redundant information.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

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.

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.

Hanyu Sun, Yanwei Zhang, Liang Hou, Zhou Yu, Yajuan Zhang, and Jing Liu. 2021. Gcrc: A benchmark for chinese reading comprehension with explanation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8912–8925. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Timothy Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data.

Yifan Wang and Guang Chen. 2019. Improving a syntactic graph convolution network for sentence compression.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022a. Chain of thought prompting elicits reasoning in large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

A. Yazdanbakhsh, Jongse Park, Hardik Sharma, P. Lotfi-Kamran, and H. Esmaeilzadeh. 2015. Neural acceleration for gpu throughput processors.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Zhilin Xu, and 1 others. 2023. Glm-130b: An open bilingual pre-trained model. `https://chatglm.cn/`. Accessed: 2025-05-19.

Yang Zhang, Yitao Zhang, Peng Liu, Liang Wang, Jiaxian Guo, Canwen Xu, Haiyun Wang, Yanan Zheng, Xuan Zhou, and Zhiyuan Liu. 2023. Gaokao-bench: Evaluating large language models for chinese college entrance exam. `https://github.com/OpenLMLab/GAOKAO-Bench`. Accessed: 2025-05-19.

Yitong Zhong, Tianyu Tang, Jingyuan Xia, Zhi Yan, Xuezhi Zhang, Shuohang Wang, Xu Tan, Xiang Lin, Weizhu Chen, and Bowen Zhou. 2023. Agieval: A human-centric benchmark for evaluating foundation models. `https://huggingface.co/datasets/agieval`. Includes gaokao-english, gaokao-mathqa, and gaokao-mathcloze subsets.

Jianpeng Zhou, Wanjun Zhong, Yanlin Wang, and Jiahai Wang. 2023. Adaptive-solver framework for dynamic strategy selection in large language model reasoning.

Yimeng Zhuang, Jing Zhang, and Mei Tu. 2022. Long-range sequence modeling with predictable sparse attention.

A Theoretical Analysis

To formally characterize the semantic preservation behavior of our chunk-based compression strategy, we analyze the impact of individual chunk relevance, fluency, and semantic alignment on the overall similarity between the original and compressed prompts. We begin by assuming that the input text can be decomposed into semantically independent chunks and that the similarity contribution of each chunk can be decoupled into three measurable components: cosine similarity to the query, contextual importance, and perplexity. Under these assumptions, we derive a lower bound on the semantic similarity retained after compression.

A.1 Lower Bound on Semantic Similarity under Chunked Compression

Assumption 1 (Semantic Factorization)**.** *The text T is decomposed into semantically independent chunks $\{C_i\}_{i=1}^{n}$, such that the overall semantic similarity between the original and compressed text satisfies*

$$S(T, T') = \prod_{i=1}^{k} S(C_i, C_i'), \tag{5}$$

where T' is the compressed version of T, and $k \le n$ denotes the number of retained chunks.

Assumption 2 (Chunk Similarity Decomposition)**.** *For each retained chunk C_i, its semantic similarity after compression is determined by:*

$$S(C_i, C_i') = \frac{\mathit{CS}(C_i, Q) \cdot \mathit{imp}(C_i)}{\mathit{ppl}(C_i)}, \tag{6}$$

where:

- *$\mathit{CS}(C_i, Q) \in [0,1]$ denotes the cosine similarity between chunk C_i and the query Q,*
- *$\mathit{imp}(C_i) \in [0,1]$ denotes the importance weight of chunk C_i,*
- *$\mathit{ppl}(C_i) \ge 1$ denotes the perplexity of chunk C_i.*

Theorem 1 (Lower Bound on Semantic Similarity under Chunked Compression)**.** *Under Assumptions*

1 and 2, the semantic similarity between the original prompt T and the compressed prompt T' is bounded below by

$$S(T,T') \geq \left(\prod_{i=1}^{k} \frac{CS(C_i,Q) \cdot imp(C_i)}{ppl(C_i)}\right)^{1/k}, \quad (7)$$

where k is the number of retained chunks.

Proof. Based on psycholinguistic assumptions, the semantic importance of chunk C_i is quantified as:

$$S(C_i, C_i') = \mathrm{CS}(C_i, Q) \cdot \frac{\mathrm{imp}(C_i)}{\mathrm{ppl}(C_i)}. \quad (8)$$

(Note: Lower $\mathrm{ppl}(C_i)$ indicates higher information density, hence the reciprocal.)

Due to chunk independence, the global similarity is the product of chunk similarities. Applying the AM-GM inequality to $\log S(T,T')$:

$$\begin{aligned}\log S(T,T') &= \sum_{i=1}^{k} \log S(C_i, C_i') \\ &\geq k \cdot \left(\prod_{i=1}^{k} \log S(C_i, C_i')\right)^{1/k}.\end{aligned} \quad (9)$$

Taking exponentials on both sides:

$$\begin{aligned}S(T,T') &\geq \left(\prod_{i=1}^{k} S(C_i, C_i')\right)^{1/k} \\ &= \left(\prod_{i=1}^{k} \frac{\mathrm{CS}(C_i,Q) \cdot \mathrm{imp}(C_i)}{\mathrm{ppl}(C_i)}\right)^{1/k}.\end{aligned} \quad (10)$$

The Equality (10) holds when all chunks satisfy $\mathrm{CS}(C_i,Q) \cdot \mathrm{imp}(C_i)/\mathrm{ppl}(C_i) = \text{const}$. □

Theorem 1 demonstrates that our framework guarantees a nontrivial semantic similarity between the original and compressed prompts, with the bound explicitly controlled by interpretable factors: relevance to the query, informativeness, and language model confidence. Importantly, the formulation provides a concrete guideline for selecting chunks that maximize semantic retention. By jointly optimizing cosine alignment, importance weighting, and perplexity reduction, our method preserves task-critical information while significantly reducing prompt length. This theoretical guarantee underpins the effectiveness of our compression strategy in practical applications, especially in scenarios where prompt budget is constrained but semantic fidelity remains essential.

A.2 Convergence of KL-Regularized Reinforcement Learning

To theoretically support the stability and reliability of the PC-RLFT framework, we analyze its convergence behavior under KL-regularized reinforcement learning. Specifically, we consider policies parameterized by differentiable function classes, such as neural networks, and assume bounded reward functions and full support from the initial policy. Under these conditions, we show that as the regularization strength increases, the learned policy progressively aligns with the optimal solution in terms of KL divergence. The following theorem formalizes this convergence guarantee.

Assumption 3 (Policy Regularity)**.** *The policy $\pi_\theta(y \mid x, c)$ is a differentiable function of the parameter θ, where π_θ is typically instantiated as a neural network, and the action space $\mathcal{Y}$ is compact.*

Assumption 4 (Bounded Reward)**.** *There exists a constant $R_{\max} > 0$ such that the context-conditioned reward function satisfies*

$$|r_c(x,y)| \leq R_{\max}, \quad \forall (x,y,c). \quad (11)$$

Assumption 5 (Support of Initial Policy)**.** *The reference (initial) policy π_c has full support over the action space, i.e.,*

$$\pi_c(y \mid x, c) > 0, \quad \forall (x,y,c). \quad (12)$$

Theorem 2 (Convergence of KL-Regularized Reinforcement Learning)**.** *Let π_θ be the policy learned via KL-regularized reinforcement learning under the PC-RLFT objective $J_{\text{PC-RLFT}}(\theta)$. Under Assumptions 1–3, the learned policy converges almost surely to the optimal policy π^* as the regularization strength $\beta \to \infty$, in the sense that*

$$\lim_{\beta \to \infty} D_{\mathrm{KL}}(\pi^* \| \pi_\theta) = 0, \quad (13)$$

with convergence rate bounded as

$$D_{\mathrm{KL}}(\pi^* \| \pi_\theta) = O\left(\frac{R_{\max}}{\beta}\right). \quad (14)$$

Proof. From Eq. (2) in the paper, the optimal policy is:

$$\pi^*(y|x,c) = \frac{\pi_c(y|x,c)\exp\left(\frac{r_c(x,y)}{\beta}\right)}{Z(x,c)}, \tag{15}$$

where $Z(x,c) = \mathbb{E}_{y\sim\pi_c}\left[\exp\left(\frac{r_c(x,y)}{\beta}\right)\right]$ is the partition function. So we have:

$$\begin{aligned} D_{\mathrm{KL}}(\pi^*\|\pi_\theta) &= \mathbb{E}_{\pi^*}\left[\log\frac{\pi^*}{\pi_\theta}\right] \\ &= \mathbb{E}_{\pi^*}\left[\frac{r_c(x,y)}{\beta} - \log Z(x,c)\right] - \mathbb{E}_{\pi^*}\left[\log\pi_\theta\right]. \end{aligned} \tag{16}$$

Since $\log Z(x,c) \leq \frac{R_{\max}}{\beta}$ (because $\exp(r_c/\beta) \leq e^{R_{\max}/\beta}$), so we have:

$$D_{\mathrm{KL}}(\pi^*\|\pi_\theta) \leq \frac{2R_{\max}}{\beta} - \mathbb{E}_{\pi^*}\left[\log\pi_\theta\right]. \tag{17}$$

As $\beta\to\infty$, $\exp(r_c/\beta)\to 1$, so $\pi^*\to\pi_c$. Thus:

$$\lim_{\beta\to\infty} D_{\mathrm{KL}}(\pi^*\|\pi_\theta) = D_{\mathrm{KL}}(\pi_c\|\pi_\theta) = 0, \tag{18}$$

because π_θ approximates π_c via gradient descent.

Using the Taylor expansion $\exp(r_c/\beta) \sim 1 + \frac{r_c}{\beta}$, the first-order approximation gives:

$$D_{\mathrm{KL}}(\pi^*\|\pi_\theta) \sim \frac{1}{\beta}\mathbb{E}_{\pi_c}\left[r_c(x,y)\right] - \mathbb{E}_{\pi_c}\left[\log\pi_\theta\right], \tag{19}$$

hence the rate $O\left(\frac{R_{\max}}{\beta}\right)$.

□

Theorem 2 provides a rigorous justification for the convergence properties of the PC-RLFT framework. The result establishes that our KL-regularized optimization not only yields a stable training dynamic but also guarantees that the learned policy asymptotically matches the optimal one as the regularization becomes sufficiently strong. Furthermore, the convergence rate being inversely proportional to the regularization coefficient provides actionable guidance for tuning β in practical implementations. These findings demonstrate the theoretical soundness of PC-RLFT and reinforce its effectiveness in balancing reward maximization with distributional alignment, a crucial property for safe and controllable reinforcement learning.

B Dataset Details

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 ZeroSCROLLS.

Arxiv-March23. (Clement et al., 2019) This dataset comprises recent academic papers from March 2023 on arXiv. We use 500 samples collected by Li as our test set. Because some articles are extremely long, we take only the first five sections of each article and limit each section to 10,000 characters. We then concatenate these sections to form the original prompt and use GPT-3.5-Turbo to generate a summary as our reference.

LongBench. (Bai et al., 2023) A multi-task long-context benchmark containing 3,750 English problems in six categories: single-document QA, multi-document QA, summarization, few-shot learning, synthetic tasks, and code completion. The average prompt length is 10,289 tokens.

ZeroSCROLLS. (Shaham et al., 2023) This multi-task long-context benchmark has 4,378 problems across four categories: summarization, question answering, aggregated sentiment classification, and information reordering. The average prompt length is 9,788 tokens.

Novel Test: We select a novel with nearly 250K context. We test the novel on Summarisation and Question Answering (QA). The Summarisation task aims to evaluate whether selective context affects the model's overall understanding of the input context. The Question Answering task aims to assess the model's understanding of specific queries. We compare compression time, compression quality, similarity to the original text, and end-to-end time on these tasks. Additionally, we propose the Compression Semantic Efficiency (CSE) metric, calculated through the compression ratio and similarity.

C Other Implementation Details

We ran our experiments on two machines—an A800-80GB GPU server and a 3090ti machine—ensuring that each experiment used the same hardware setup for consistency. We used tiktoken11 and the GPT-3.5-Turbo model to calculate all tokens. We open-sourced an early version of the system's preset instructions, part of the core intermediate process: a set of text simplification rules that help users select the appropriate level of simplification based on reading and comprehension needs.

D Rules

D.1 Basic Rules

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

4. **Deep Simplification:** Apply both basic and advanced rules, retaining key sentence meaning.

5. **Very Deep Simplification:** Extremely reduce details, retaining only the main parts of the sentence (subject, verb, object).

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 that 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.

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.

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, Motheris it true? Can’t you really come with us on our holidays? Youalways 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 ratherfun for you to have a holiday on your own too. But now that you can’t go to Polseath, I don’t reallyquite 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, frowningman, a clever scientist who spent all his time studying. He lived by the sea but that was about allthat the children knew of him! "Quentin?" said Mother, pursing up her lips. "Whatever made you think of him? I shouldn’t think he’dwant 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 matterand I don’t think things are going too well for them. Fanny said that she would be quite glad if shecould hear of one or two people to live with her for a while, to bring a little money in. Their house isby the sea, you know. It might be just the thing for the children. Fanny is very niceshe 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 seewhat’s her name something funny yes, Georgina! How old would she be? About eleven, I shouldthink." 2 "Same age as me," said Dick. "Fancy having a cousin we’ve never seen! She must be jolly lonely allby herself. I’ve got Julian and Anne to play with but Georgina is just one on her own. I shouldthink she’d be glad to see us." "Well, your Aunt Fanny said that her Georgina would love a bit of company," said Daddy. "Youknow, I really think that would solve our difficulty, if we telephone to Fanny and arrange for thechildren 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.