# LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression

Huiqiang Jiang [1] Qianhui Wu [1]

[1]Microsoft Corporation.

## Research Gap

Large language models (LLMs) in long context scenarios face higher computational costs, performance reduction due to irrelevant information, and position bias (the **"lost in the middle"** issue) where the placement of key information affects performance. Existing compression methods (e.g., LLMLingua, Selective-Context) do not account for question relevance during compression, leading to ineffective information retention.

## Main Contributions

The paper proposes LongLLMLingua for prompt compression. It aims to improve LLMs' perception of key information **relevant to the question** to simultaneously address the challenges of cost, performance reduction, and **position bias** in long contexts. Key contributions include a **question-aware coarse-to-fine compression** method, a **document reordering strategy**, **dynamic compression ratios**, and a **subsequence recovery strategy**.

## Method Overview

LongLLMLingua builds on the perplexity-based compression of LLMLingua, using a **small language model** to assess token importance. It adds question-awareness at both coarse (document) and fine (token) levels.

- **Coarse-grained compression** evaluates document importance based on the perplexity of the question conditioned on the document ($p(x_{\text{que}}|x_{\text{doc }k})$), aided by a restrictive statement, to select the most relevant documents.
- **Document reordering** places more important documents (based on their score) earlier in the prompt to mitigate position bias ("lost in the middle").
- **Fine-grained compression** uses **contrastive perplexity** ($s_i = \text{perplexity}(x_i|x_{<i}) - \text{perplexity}(x_i|x_{\text{que}}, x_{<i})$) to assess the importance of **individual tokens** based on their association with the question, preserving key information density.
- **Dynamic compression** ratios apply different **compression budgets** in fine-grained compression based on the document importance scores from the coarse stage.
- A subsequence **recovery strategy** post-processes LLM responses to restore the **integrity of entities** (like names or places) by matching original prompt subsequences.

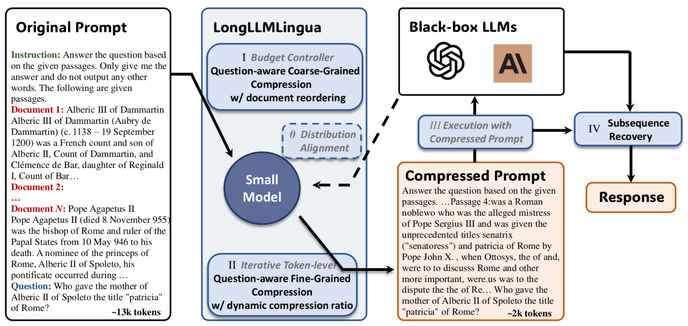



Figure 1. Framework of *LongLLMLingua*. Gray *Italic* content: As in LLMLingua.

## Key Findings

- LongLLMLingua achieves **better performance than original prompts** and other baselines across various long context tasks and compression ratios. For instance, it boosted performance on NaturalQuestions by up to **21.4%**.
- It leads to **substantial cost savings** (e.g., up to 94.0% reduction on LooGLE, **4x fewer tokens** on NaturalQuestions with GPT-3.5-Turbo).
- It significantly reduces **end-to-end inference latency**, accelerating it by 1.4x-2.6x when compressing 10k tokens.
- The method demonstrates strong robustness **across different tasks** and compression constraints compared to other baselines.

## Strengths and Weaknesses

Strengths:

- Effective compression preserves key info better than **entropy-based** methods.
- **Modular** design applicable to **black-box** LLMs.
- **Ablation studies** validate the effectiveness of its individual components.

Weaknesses:

- It is a **question-aware** approach, requiring **re-compression** for each new question, which prevents **context caching** and incurs more computational overhead (twice that of LLMLingua).
- Its effectiveness might be limited when the **context-prompt relationship** is more **complex and subtle** due to its coarse-level question-aware design.

## Future Directions

- Extend from **question-aware** to **task-aware compression**, allowing context reuse and caching.
- Further research into handling complex context-prompt relationships beyond explicit relevance.

## Core Equation

Document Importance Score (Coarse Compression):

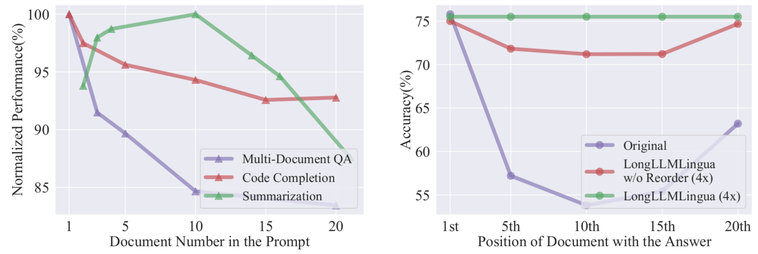
$$r_k = -\frac{1}{N_c}\sum_{i=1}^{N_c} \log p(x_{\text{que, restrict},i} \mid x_{\text{doc},k})$$

- $r_k$: Importance score for document $k$
- $x_{\text{que, restrict},i}$: $i$-th token of the concatenated sequence of the question and the restrictive statement
- $x_{\text{doc},k}$: The $k$-th document in the prompt
- $N_c$: Total number of tokens in $x_{\text{que, restrict}}$
- $p(\cdot \mid \cdot)$: Probability assigned by a small language model

## Ablation study on NaturalQuestions with 2x constraint using GPT-3.5-Turbo.

| | 1st | 5th | 10th | 15th | 20th |
|---|---|---|---|---|---|
| **LongLLMLingua** | **77.2** | **72.9** | **70.8** | **70.5** | **70.6** |
| *Question-aware Coarse-grained* | | | | | |
| - w/o Question-awareness | 42.1 | 40.3 | 39.7 | 40.1 | 40.3 |
| - w/ SBERT | 73.2 | 68.5 | 65.7 | 66.1 | 66.7 |
| - w/ $p(\mathbf{x}_k^{\text{doc}}\|x_i^{\text{que,restrict}})$ | 56.0 | 52.6 | 53.4 | 51.6 | 51.1 |
| - w/o restrict | 75.1 | 72.2 | 70.3 | 70.3 | 70.2 |
| - w/o Question-aware Fine-grained | 75.8 | 71.0 | 68.9 | 68.4 | 69.3 |
| - w/o Dynamic Compression Ratio | 74.4 | 70.7 | 68.7 | 67.9 | 68.1 |
| - w/o Subsequence Recovery | 76.7 | 71.7 | 69.4 | 69.3 | 69.7 |
| - w/ Document Reordering | 76.2 | 76.2 | 76.2 | 76.2 | 76.2 |
| - w/ GPT2-small | 74.6 | 71.7 | 70.1 | 69.8 | 68.5 |
| LLMLingua | 39.7 | 39.5 | 40.4 | 37.1 | 42.3 |
| - w/ Subsequence Recovery | 43.8 | 44.1 | 43.5 | 43.3 | 44.4 |

## Noise, Position Bias




(a) Performance v.s. Document Number

(b) Performance v.s. Key Information Position

Github Pourya Aliannejadi alian.pourya@gmail.com