LLMLingua-2: Faithful and Efficient Prompt Compression via Data Distillation

Zhuoshi Pan [1] Qianhui Wu [2]

[1]Tsinghua University [2]Microsoft Corporation

Research Gap

Existing task-agnostic prompt compression methods, which rely on information entropy from causal language models for token removal, face two main challenges:

- First, information entropy may be a suboptimal metric for prompt compression because it only uses unidirectional context and might miss essential information.
- Second, information entropy is not directly aligned with the objective of prompt compression.

Additionally, existing text compression datasets are often abstractive or lack detailed information, which can hurt the performance of LLM inference in downstream applications like question answering.

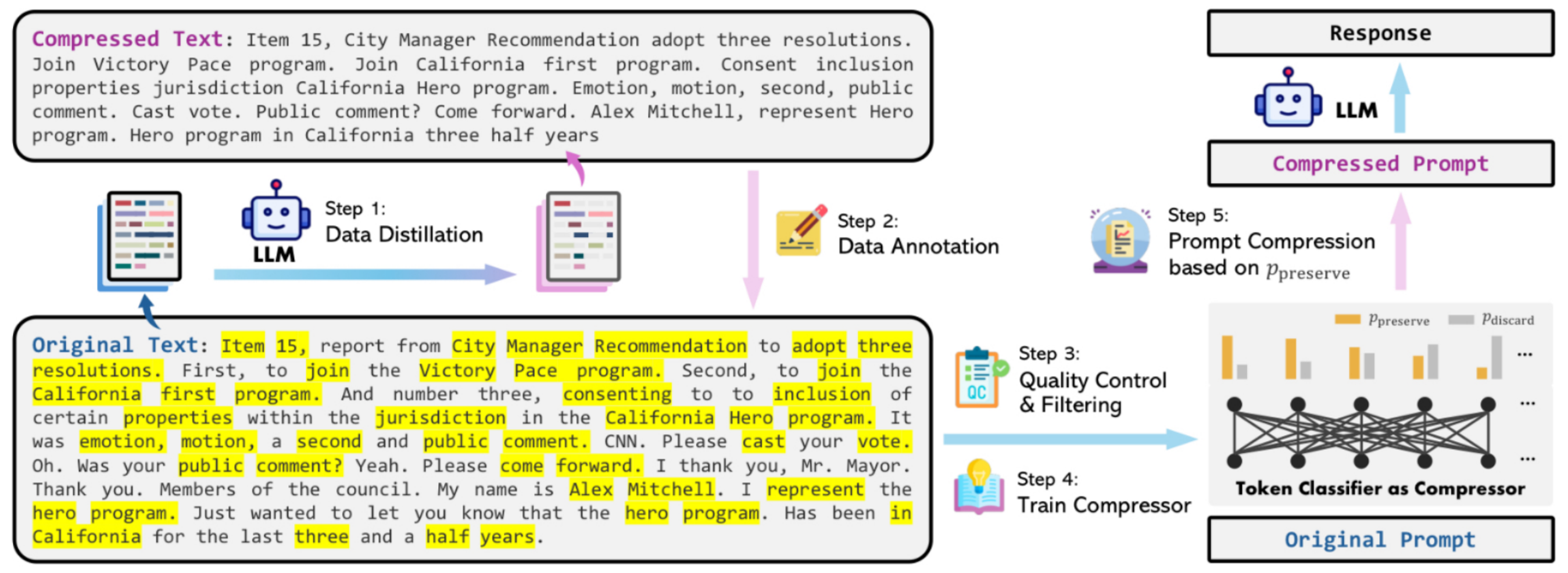
Main Contributions

- A **data distillation** procedure to derive knowledge from a large language model (GPT-4) to compress prompts without losing crucial information.
- Introduces an **extractive compression dataset** (MeetingBank) for training.
- **Token classification formulation** as a token classification problem (preserve or discard) using a Transformer encoder to leverage the full bidirectional context. This approach explicitly learns the compression objective with smaller models, leading to lower latency and guaranteed faithfulness to the original prompt.

Method Overview

The key method involves a data distillation process where GPT-4 is prompted to generate compressed texts from original texts by only removing unimportant words.

1. Chunk long prompts, compress each via GPT-4 using strict no-hallucination rules.
2. Annotate each token: *preserve* or *discard*.
3. Train a classifier (XLM-RoBERTa or mBERT) to predict whether each token in an original prompt should be preserved or discarded.
4. During inference, the model calculates the probability of each token being preserved, and the top tokens based on this probability are retained to form the compressed prompt.



Compressed Text: Item 15, City Manager Recommendation adopt three resolutions. Join Victory Pace program. Join California first program. Consent inclusion properties jurisdiction California Hero program. Emotion, motion, second, public comment. Cast vote. Public comment? Come forward. Alex Mitchell, represent Hero program. Hero program in California three half years

Original Text: 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.

Figure 1. LLMLingua-2 Pipeline Overview.

Key Findings

- The proposed LLMLingua-2 model shows **significant performance gains over strong baselines** (Selective-Context and LLMLingua) on both in-domain (MeetingBank) and out-of-domain datasets (LongBench, ZeroScrolls, GSM8K, BBH) despite being much smaller. In some cases, it even achieves comparable or slightly higher performance than the original prompt.
- LLMLingua-2 is **3x-6x faster than existing prompt compression methods** and accelerates end-to-end latency by **1.6x-2.9x** with compression ratios of **2x-5x**. It also significantly reduces GPU memory costs.
- Experiments show that GPT-4 can effectively reconstruct the original prompt from the LLMLingua-2 compressed prompt, suggesting minimal essential information loss during compression.

Strengths and Weaknesses

Strengths:

- The paper proposes a novel **data distillation approach** using a powerful LLM (GPT-4) to create a high-quality extractive compression dataset.
- Formulating prompt compression as a **token classification problem** allows for leveraging bidirectional context and explicitly learning the compression objective.
- The authors **publicly release the newly created dataset**.

Weaknesses:

- The initial text compression dataset was constructed using only training examples from MeetingBank, which primarily focuses on meeting transcripts, potentially raising concerns about the generalizability of the compressor.
- Task-aware methods can outperform on specific QA tasks.

Future Directions

The authors suggest further work to investigate why the redundancy patterns in text might be similar across different domains, allowing for good transferability of the learned compression knowledge.

- Exploring the effectiveness of this data distillation and token classification approach for **task-aware prompt compression**.
- Investigating the impact of using **different LLMs for data distillation**.
- Applying the LLMLingua-2 approach to compress other parts of the LLM input beyond the initial prompt, such as retrieved documents in RAG.

Core Equation

Training Objective for Token Classification:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \text{CrossEntropy}(y_i, p(x_i, \Theta))$$

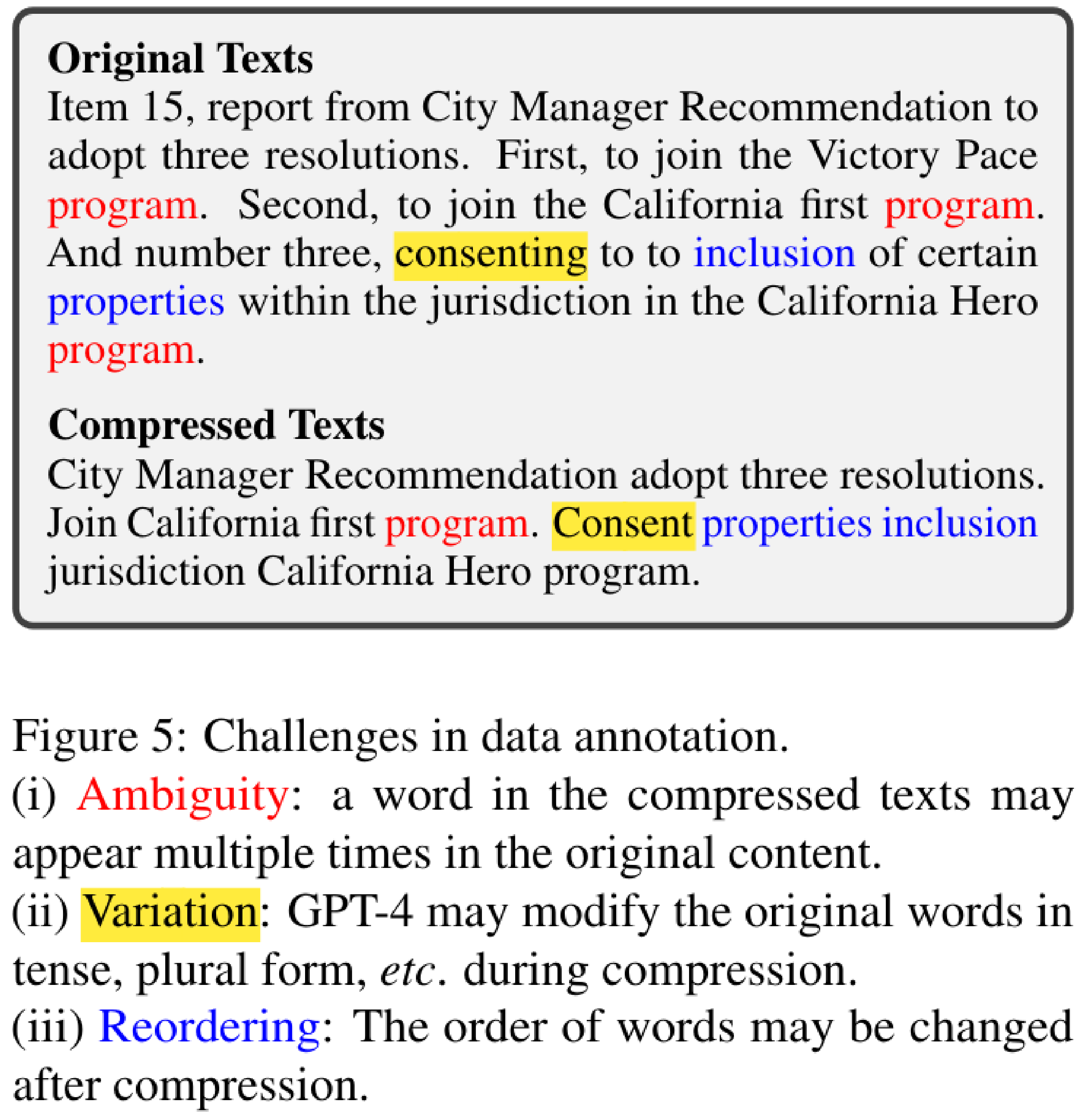
Where $p(x_i, \Theta)$ is the predicted probability of keeping token x_i.

Latency Comparison (V100 GPU)

Method	Latency (2x–5x)
LLMLingua-2	0.5–0.4 sec
LLMLingua	2.9–1.5 sec
Selective-Context	15.9–15.5 sec

Table 1. Prompt Compression Latency (MeetingBank)

Example



Original Texts
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.

Compressed Texts
City Manager Recommendation adopt three resolutions. Join California first program. Consent properties inclusion jurisdiction California Hero program.

Figure 5: Challenges in data annotation. (i) Ambiguity: a word in the compressed texts may appear multiple times in the original content. (ii) Variation: GPT-4 may modify the original words in tense, plural form, *etc.* during compression. (iii) Reordering: The order of words may be changed after compression.

Figure 2. Challenges in data annotation.

References

- Pan et al. 2024, LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Comp ression, arXiv:2403.12968

Github Pourya Aliannejadi alian.pourya@gmail.com