Prompt Compression for Retrieval-Augmented Generation

Parth Asawa, Shreya Shankar, Aditya Parameswaran

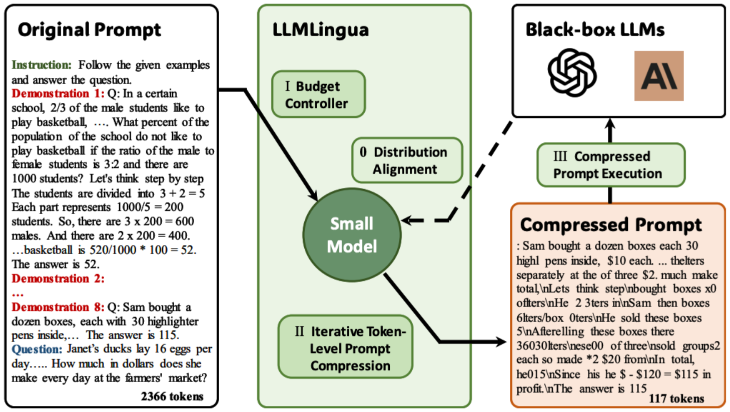
Exploring applications of compression techniques to decrease cost and latency in retrieval-augmented generation.

Background

- **Problem**
  - LLMs are extremely powerful at generative tasks, though are often expensive and high-latency.
  - LLM API inference time scales quadratically and costs scale linearly with input length.
- **Goal**
  - Explore the effectiveness of various types of compression techniques from both the databases and information theory world on reducing input size while maintaining accuracy.
  - Propose methods for compression over both structured and unstructured data corpuses.

Prior Prompt Compression Work

LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Model, Jiang, et al.



Original Prompt

Instruction: Follow the given examples and answer the question.
Demonstration 1: Q: In a certain school, 2/3 of the male students like to play basketball, …. What percent of the population of the school do not like to play basketball if the ratio of the male to female students is 3:2 and there are 1000 students? Let's think step by step The students are divided into 3 + 2 = 5 Each part represents 1000/5 = 200 students. So, there are 3 x 200 = 600 males. And there are 2 x 200 = 400. …basketball is 520/1000 * 100 = 52. The answer is 52.
Demonstration 2:
…
Demonstration 8: Q: Sam bought a dozen boxes, each with 30 highlighter pens inside,… The answer is 115.
Question: Janet's ducks lay 16 eggs per day….. How much in dollars does she make every day at the farmers' market?
2366 tokens

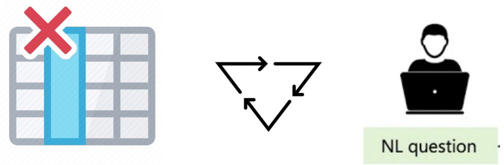
Compressed Prompt

: Sam bought a dozen boxes each 30 highl pens inside, $10 each. ... thelters separately at the of three $2. much make total,\nLets think step\nbought boxes x0 oflters\nHe 2 3ters in\nSam then boxes 6lters/box 0ters\nHe sold these boxes 5\nAfterelling these boxes there 36030lters\nese00 of three\nsold groups2 each so made *2 $20 from\nIn total, he015\nSince his he $ - $120 = $115 in profit.\nThe answer is 115
117 tokens

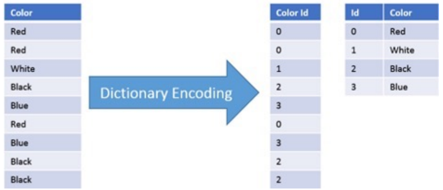
Figure 1: Framework of the proposed approach *LLMLingua*.

Structured Data Corpuses

- **Case Study:**
  - Compression with 3 strategies:



Columnar Compression



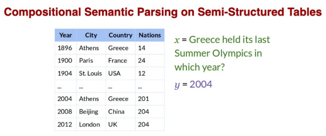
Color
Red
Red
White
Black
Blue
Red
Blue
Black
Black

Color Id
0
0
1
2
3
0
3
2
2

Id	Color
0	Red
1	White
2	Black
3	Blue

Use Cases & Workloads

- We explore use cases of compression in web tables, tabular datasets, unstructured text, and more.



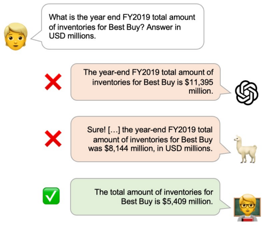
Compositional Semantic Parsing on Semi-Structured Tables

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x = Greece held its last Summer Olympics in which year?

y = 2004

Panupong, et al.



What is the year end FY2019 total amount of inventories for Best Buy? Answer in USD millions.

✗ The year-end FY2019 total amount of inventories for Best Buy is $11,395 million.

✗ Sure! […] the year-end FY2019 total amount of inventories for Best Buy was $8,144 million, in USD millions.

✓ The total amount of inventories for Best Buy is $5,409 million.

Islam, et al. (FinanceBench)

Unstructured Data Corpuses

- General-purpose large language models don't model the importance of words/tokens in specific domains well.
- **Domain-specific compression:** use fine-tuned models.
  - BloombergGPT: A Large Language Model for Finance
  - Med-PaLM: A Medical Large Language Model

Other Considerations & Future Work

- Identifying best compression strategies automatically over long mixed-structure documents.
- Revisiting traditional compression techniques like TF-IDF based stop-word removal.
- What are the most important data workloads or problems for which people would find this work impactful?