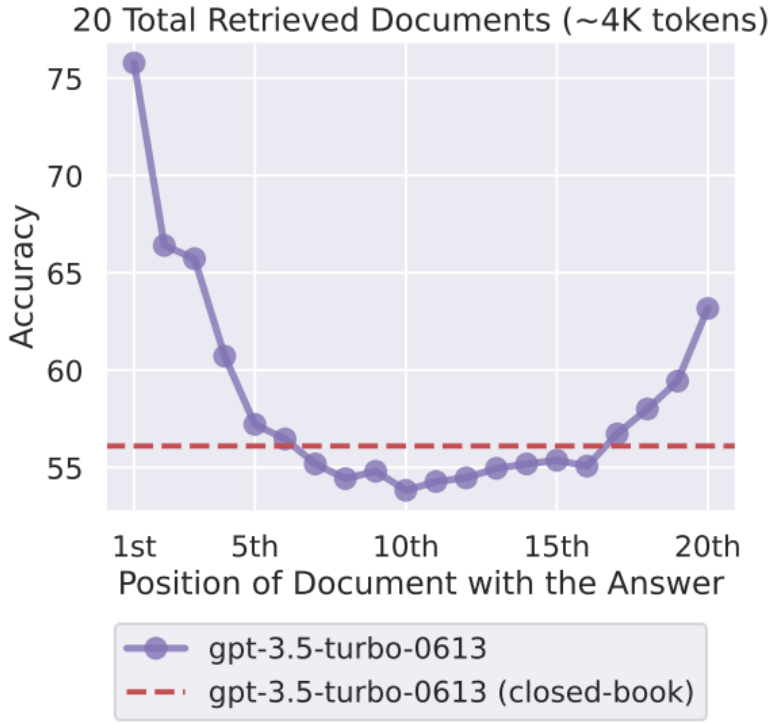
基於LLaMA 3.1 和 BART 語言模型的長文本摘要多層次演算法

動機/目的

研究動機：
- LLM 在長文本情境下 容易因注意力機制分散導致的位置偏差問題。
- 存在 **上下文窗口上限** 與 **幻覺問題**，降低模型在長文本任務的可靠度

研究目標：
- 降低長文本處理中的效能下降問題
- 在有限資源下，同時提升效率與準確率



資料集 & 評估指標

CNN/DailyMail 數據集 :

- 針對新聞文章進行摘要設計的資料集。
- 資料集內容: 唯一識別碼 (id)、文章(article)、摘要 (highlights)。

```
{
  id : 062f78c2922d4050190dbba10f5d65eeff25e1ed
  article : Bayern Munich have an interest in Chelsea defender Branislav Ivanovic ...
  highlights : Branislav Ivanovic's contract at Chelsea expires ...
}
```

ROUGE 評估指標介紹 :

- ROUGE-1 (R-1)：衡量摘要的內容覆蓋率。
- ROUGE-2 (R-2)：評估語言流暢性與結構一致性。
- ROUGE-L (R-3)：評估語義結構與內容順序的連貫性。

主要分析項目: 壓縮率、信息保留度、語言流暢性。

LLaMA3.1 - 壓縮比測試

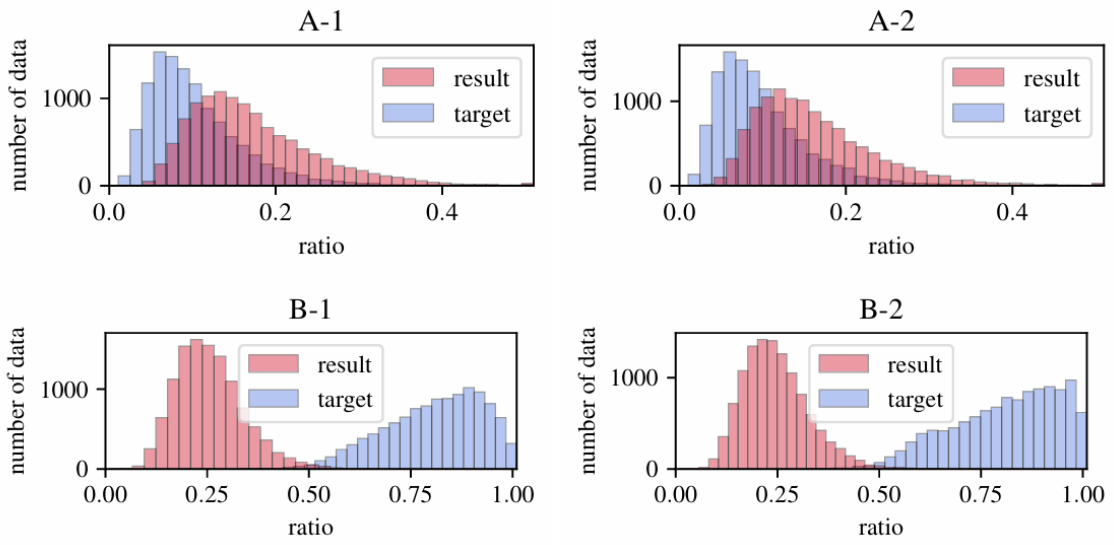
測試結果觀察:

- 嘗試方法:
  - 設定 system prompt
  - 在 user prompt 加入壓縮比限制

User Prompt:

1. Please summarize the following text to approximately **{ratio}** of the original length.
2. The output should be a direct, concise summary without any extra commentary or introductory text.
3. original text: {...}

- 問題:
  - 壓縮比受控不佳，摘要品質不穩定
  - 超過上下文窗口，文章被截斷，摘要失真



多層次摘要框架

layer	
original	$[T_1^{(1)}], \quad [T_2^{(1)}], \quad [T_3^{(1)}], \quad [T_4^{(1)}], \ldots, \quad [T_{m_1}^{(1)}]$
summary 1	$[S_1^{(1)}], \quad [S_2^{(1)}], \quad [S_3^{(1)}], \quad [S_4^{(1)}], \ldots, \quad [S_{m_1}^{(1)}]$
combine 1	$[C_1^{(1)}], \quad [C_2^{(1)}], \quad [C_3^{(1)}], \quad \ldots, \quad [C_{m_2}^{(1)}]$
summary 2	$[S_1^{(2)}], \quad [S_2^{(2)}], \quad [S_3^{(2)}], \quad \ldots, \quad [S_{m_2}^{(2)}]$
combine 2	$[C_1^{(2)}], \quad [C_2^{(2)}], \quad \ldots, \quad [C_{m_3}^{(2)}]$
$\vdots$	$\vdots$
summary n	$[S_1^{(n)}], \quad [S_2^{(n)}]$
combine n	$[C_1^{(n)}]$

表 3.1: 多層次摘要框架

實驗方法

1. LLaMA3.1 壓縮文本:
   - 設定system prompt
   - 設計 4 個壓縮指令

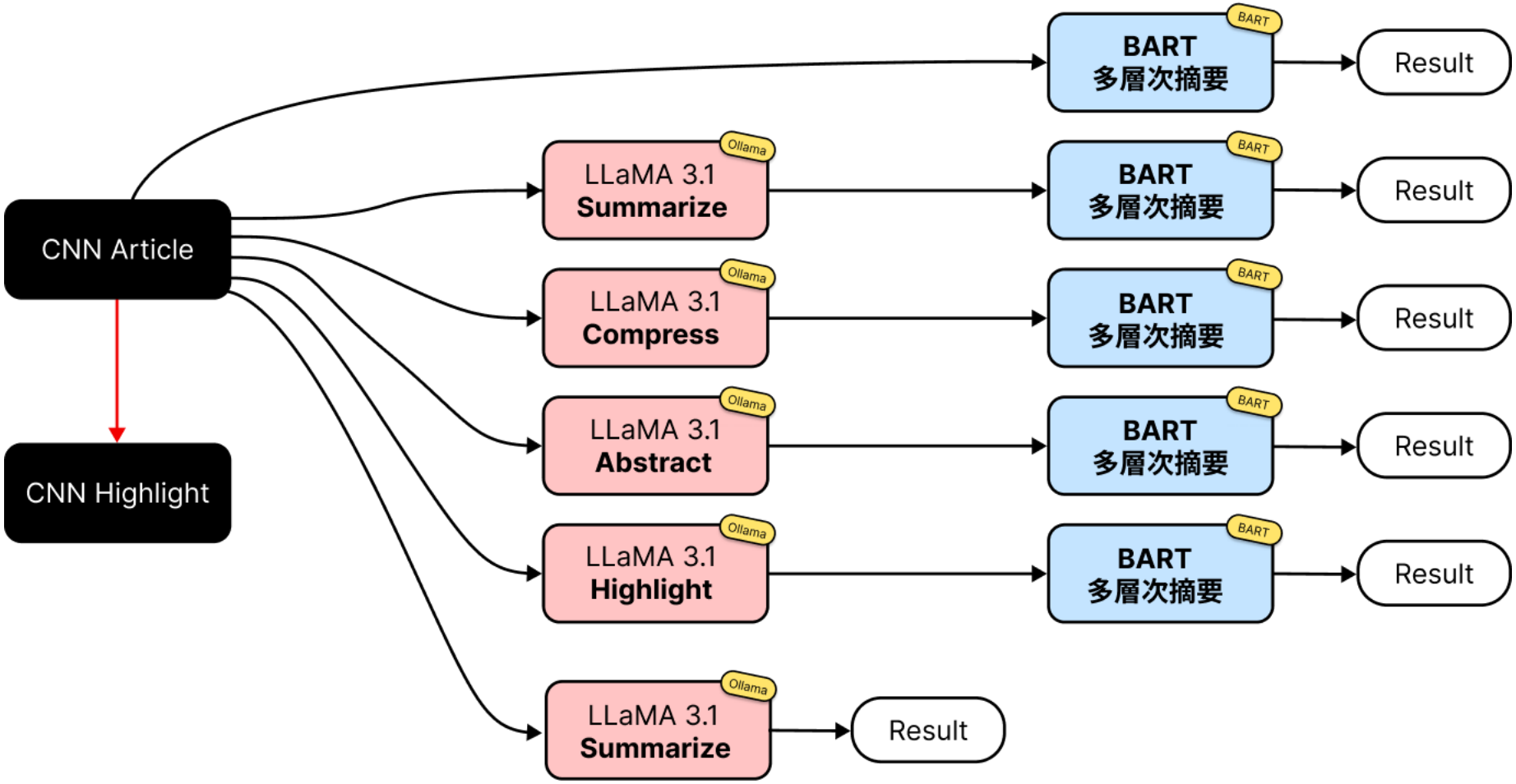
Our compression instructions:

1. You can ONLY remove unimportant words.
2. Do not reorder the original words.
3. Do not change the original words.
4. Do not use abbreviations or emojis.
5. Do not add new words or symbols.
6. Provide ONLY the result text, without any explanations or additional information.

2. BART 多層次摘要

```
original layer-1 (13 parts):    | 500 | 473 | 489 | 500 | 505 | 508 | 497 | 505 | 459 | 507 | 461 | 411 | 478 |
 summary layer-1 (13 parts):    | 245 | 255 | 245 | 246 | 238 | 241 | 242 | 256 | 237 | 240 | 231 | 239 | 256 |
original layer-2 (7  parts):       | 498 | 489 | 477 | 495 | 475 | 468 | 256 |
 summary layer-2 (7  parts):       | 239 | 245 | 240 | 241 | 234 | 253 | 256 |
original layer-3 (4  parts):          | 482 | 479 | 485 | 256 |
 summary layer-3 (4  parts):          | 249 | 254 | 238 | 256 |
original layer-4 (2  parts):             | 501 | 492 |
 summary layer-4 (2  parts):             | 248 | 248 |
original layer-5  (1  parts):                | 494 |
summary  layer-5  (1  parts):                | 286 |
```

流程圖


CNN Article
CNN Highlight
Ollama
LLaMA 3.1
Summarize
Ollama
LLaMA 3.1
Compress
Ollama
LLaMA 3.1
Abstract
Ollama
LLaMA 3.1
Highlight
Ollama
LLaMA 3.1
Summarize
Result
BART
BART
多層次摘要
Result
BART
BART
多層次摘要
Result
BART
BART
多層次摘要
Result
BART
BART
多層次摘要
Result
BART
BART
多層次摘要
Result

Paper with Code 平台表現

Rank	Model	ROUGE-1
1.	Scrambled code + broken (alter)	48.18
2.	**BART 多層次摘要**	**47.54**
3.	PEGASUS + SummaReranker	47.16
⋮	⋮	⋮
8.	MatchSum (BERT-base)	44.22
9.	**LLaMA 3.1 + BART 多層次摘要**	**43.95**
10.	BertSumExt	43.85
⋮	⋮	⋮

表 7.1: **ROUGE-1** in Papers with Code

Rank	Model	ROUGE-2
1.	**BART 多層次摘要**	**23.34**
2.	PEGASUS + SummaReranker	22.55
3.	Fourier Transformer	21.55
⋮	⋮	⋮
12.	BERTSUM+Transformer	20.24
13.	**LLaMA 3.1 + BART 多層次摘要**	**20.02**
14.	Scrambled code + broken (alter)	19.84
⋮	⋮	⋮

表 7.2: **ROUGE-2** in Papers with Code

Paper with Code 平台表現

Rank	Model	ROUGE-L
1.	Scrambled code + broken (alter)	45.35
2.	PEGASUS + SummaReranker	43.87
3.	HAT-BART	41.52
4.	**BART 多層次摘要**	**41.50**
⋮	⋮	⋮
13.	BERTSUM+Transformer	39.63
14.	**LLaMA 3.1 + BART 多層次摘要**	**38.86**
15.	Selector+Pointer Generator	38.79
⋮	⋮	⋮

表 7.3: **ROUGE-L** in Papers with Code

結論

1 避免注意力機制分散與硬體資源消耗問題
- LLaMA3.1 壓縮長文本，減少冗餘內容。
- BART 多層次摘要，逐步提取語義連貫且準確的摘要。

2 實驗結果
- 在CNN/Daily Mail 資料集上，於公開平台上取得良好結果。

3 實際應用上的靈活性
- 適合計算資源有限的環境。
- 可根據不同需求調整壓縮與摘要的層次。