小模型，大貢獻：準確且高效的中文新聞摘要模型訓練研究

Small Model, Big Impact: A Study on Training for Accurate and Efficient Chinese News Summarization Models

指導教授：陳響亮 教授 專題成員：王郁豪

一、 動機

在資訊爆炸與節奏快速的現代社會，大眾往往無暇細讀完整新聞報導，並可能因聳動標題而造成資訊判斷偏差。因此，若能開發一款可在手機等資源有限的設備上運行、具備高理解能力的小型中文摘要模型，不僅能快速提供精煉且具意義的新聞摘要，亦可大幅提升資訊獲取的效率與品質。

二、 目的

本專題旨在結合模型蒸餾（Knowledge Distillation）與課程式訓練（Curriculum Training）等策略，有效壓縮模型參數，同時提升模型泛化能力與語言理解效果，最終實作一個在行動裝置端部署模型的中文新聞摘要程式，達成「小模型、高品質」的實用目標。

三、 實驗設計

1. 資料蒐集：
   自行撰寫爬蟲程式，收集聯合新聞網新聞，並整理 YeungNLP/firefly-pretrain-dataset 作為模型預訓練的資料。
2. 模型選擇：
   - 學生模型選用 Qwen2.5-0.5B-Instruct（以下簡稱為 0.5B 模型）。
   - 教師模型選用 Qwen2.5-32B-Instruct（以下簡稱為 32B 模型），作為生成資料、資料處理以及評分指引的模型。
3. 課程式訓練：
   訓練共分為五個階段（S1-S5）：
   - S1 繁體中文轉換：使用 OpenCC 轉換成繁體的輸出作為訓練資料。
   - S2 要素提取：給定新聞內容，提取出關鍵要素（essential aspect）。
   - S3 推理建構：給定新聞內容與關鍵要素，建構代表要素間關聯的三元組（triple）。
   - S4 新聞摘要：給定新聞內容、關鍵要素與三元組，生成新聞摘要。
   - S5 直接摘要：僅給予新聞內容，直接生成新聞摘要。

4. 資料生成策略：

實驗比較以下四種資料建構流程：

- V1: 一次性生成關鍵要素、三元組與摘要。
- V2: 分階段先生成關鍵要素，後生成三元組，最後生成摘要。
- V3: 先生成摘要，再推導關鍵要素與三元組。
- V4: 在 V3 的基礎上進行人工修正（錯字、用語、格式等）。

資料集統計如下：

表 1：訓練資料數量（單位：篇文章）

	繁中訓練集	繁中驗證集	訓練集	驗證集	測試集
篇數	23848	5960	19386	4847	6052

5. 參數設置：

實驗測試了不同的參數設置，並比較其對模型訓練結果的影響，以探討不同訓練策略對性能的貢獻及其背後的意涵：

- 預設：使用遞減學習率（Learning rate decay），這是最常見的訓練策略，能幫助模型在初期快速收斂，後期穩定微調。
- lr_adj：在 S5 訓練階段中，設定學習率與 S1 相同，即不使用遞減策略，試圖觀察是否能避免模型收斂過早或陷入局部最小值。此舉可探討遞減學習率在本任務中是否必要，或是否過早抑制了模型的學習能力。
- only_attn：僅訓練 attention head，凍結其他參數。這樣做可檢驗模型是否主要依賴注意力機制進行任務學習，並觀察在資源受限的微調條件下，是否仍能取得合理表現。
- only_mlp：僅訓練 attention head 後的全連接（MLP） 層，凍結 attention 層與其他參數。這項設定用在檢視模型的非注意力部分是否具有足夠的表達能力，並探討 MLP 層對最終預測的貢獻。
- lora：使用 LoRA（Low-Rank Adaptation）進行參數高效微調，將 MLP 層的 rank 設為 160，attention head 的 rank 設為 32。此做法旨在不修改大部分原始參數的前提下，透過加入少量可訓練參數提升模型能力，期望在效能與計算資源之間取得良好平衡。

6. 課程階段比較：

實驗比較了三種訓練流程：

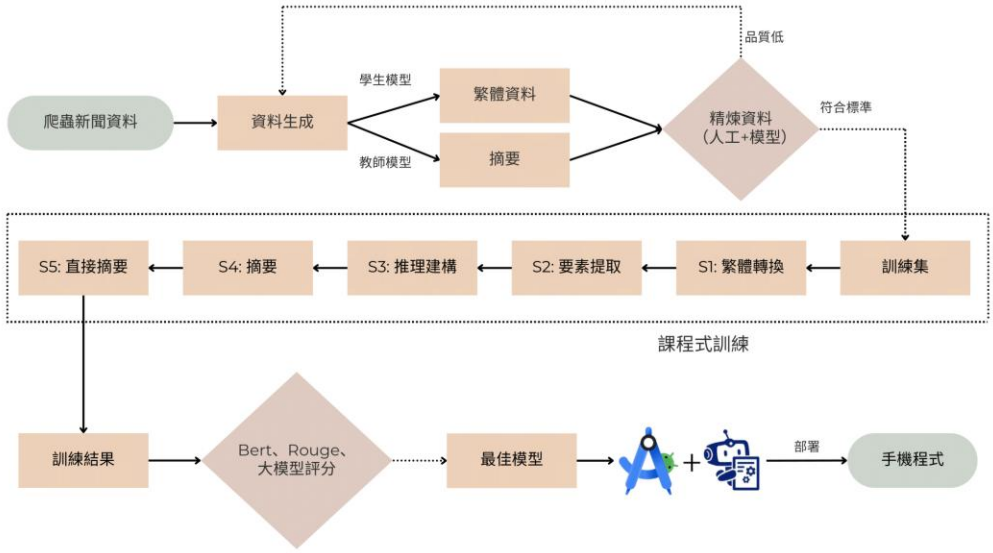
- 1-stage：僅使用 S5（無課程式學習）。
- 4-stage：省略 S1（不翻譯），使用 S2-S5。
- 5-stage：完整課程式訓練 S1-S5。

7. 模型評估：

- 自動評分指標：使用 ROUGE-1/2/L、BERTScore

- 模型評分：由 32B 教師模型針對生成摘要進行自然性與資訊涵蓋度評分（Judge 分數）。
- 以下實驗會主要以 R-1（Rouge1-F）、B-F1（BERTScore F1）以及 Judge（模型評分）作為品質判斷依據，並使用 R-2（ROUGE-2 F）與 R-L（ROUGE-L F）作為輔助資訊。

以下為實驗流程圖



圖一：實驗流程圖

四、 實驗結果與討論

1. 資料生成策略
   - 實驗發現教師模型不擅長以推理方式直接生成摘要，在 V1 中，產出的關鍵要素與三元組彼此雖然僅有少部分不連貫，但涵蓋的摘要內容過少，無法僅靠要素和三元組組合出好的摘要，因此改用 V2 分三次產生資料，希望能減少部連貫性。
   - 然而，V2 產生的要素和三元組更加的不連貫，且內容常常失焦，為了修正這問題，我將生成方式改為先生成摘要，再推導關鍵要素與三元組的 V3。
   - 從實驗結果可見，V3 的表現最佳，整體評分指標均勝過 V1、V2。V2 最差的表現直接反映出「失焦」和「不連貫」問題，使得模型無法順利的學習到推論的過程。（為節省訓練時間，省略翻譯的 S1 並使用 OpenCC 套件翻譯）

表 2：使用不同資料生成方式的訓練表現

MODEL	R-1	B-F1	Judge	R-2	R-L
Qw2.5-0.5B_4stg_v3	**45.5**	**77.9**	**70.3**	**24.3**	**37.6**
Qw2.5-0.5B_4stg_v1	43.8	76.8	64.0	22.1	35.5
Qw2.5-0.5B_4stg_v2	37.6	69.4	65.1	17.5	23.4

2. 不同訓練策略

實驗發現：

- 使用不使用遞減學習率能使學生模型表現更好，整體優於傳統遞減策略。
- 固定訓練 Attention 或 MLP 仍具不錯表現：例如 only_mlp 可保持高 B-F1（78.6%），說明模型的注意力與非注意力部分都具有足夠的摘要能力。
- 對於摘要的任務使用 LoRA 訓練效果不好，幾乎沒有進步。此外，引入額外參數也違背本研究「小模型」原則。

表 3：使用不同訓練策略的訓練表現

MODEL	R-1	B-F1	Judge	R-2	R-L
4stg_v3-lr_adj	**48.4**	**79.3**	**72.8**	**25.7**	**40.1**
4stg_v3-lr_adj-only_mlp	46.6	78.6	71.5	24.2	38.4
4stg_v3-lr_adj_lora	45.6	78.0	73.6	23.3	37.4
4stg_v3	45.5	77.9	70.3	24.3	37.6
4stg_v3-lr_adj-only_attn	45.2	77.8	69.1	23.0	37.1

3. 不同訓練階段

從 S1-S5 分段實驗可見：

- 在不使用遞減學習率的條件下，1-stage 模型在 R-1 上稍高於 5-stage，但 5-stage 能顯著提高繁體中文生成比例與輸出品質穩定性。
- 在遞減學習率或在自定義預訓練模型下，階段越多越能幫助模型掌握摘要結構與繁體中文風格。

表 4：不同階段數在遞減學習率策略的訓練表現

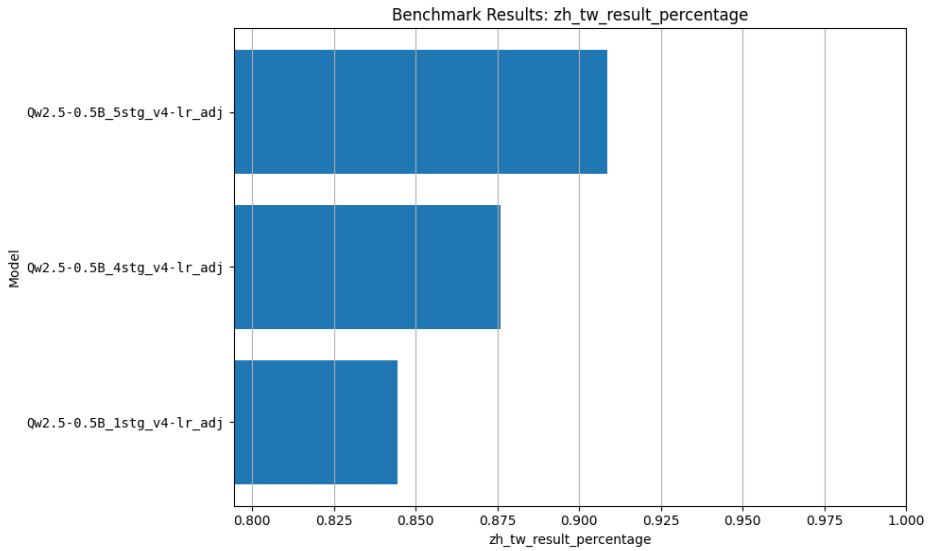
MODEL	R-1	B-F1	Judge	R-2	R-L
Qw2.5-0.5B_4stg_v3	45.5	**77.9**	**70.3**	**24.3**	37.6
Qw2.5-0.5B_1stg_v3	**45.9**	77.9	68.8	23.7	**37.7**

表 5：不同階段數在不使用遞減學習率策略的訓練表現

MODEL	R-1	B-F1	Judge	R-2	R-L
Qw2.5-0.5B_1stg_v4-lr_adj	**47.0**	**78.8**	**75.9**	**24.5**	**38.9**
Qw2.5-0.5B_4stg_v4-lr_adj	46.8	78.6	75.2	24.2	38.6
Qw2.5-0.5B_5stg_v4-lr_adj	46.5	78.6	75.1	23.9	38.3

表 6：不同階段數在自定義預訓練模型的訓練表現

MODEL	R-1	B-F1	Judge	R-2	R-L
Custom-pre-4stg_v3-lr_adj	**13.3**	**50.6**	**0.9**	**1.8**	4.0
Custom-pre-1stg_v3-lr_adj	11.8	50.5	0.6	1.5	**5.0**



圖二：不同階段數訓練的繁體中文輸出的比例

4. 最終模型性能與現有模型比較

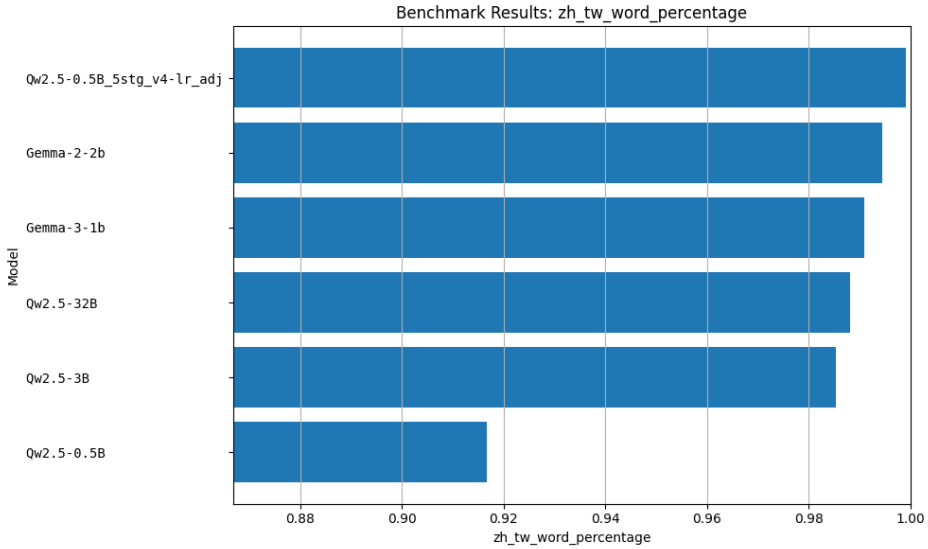
引入教師模型與同級別學生模型比較：

- 0.5B 模型最終版在 R-1/B-F1 處理上接近 3B 同參數模型，生成品質提升明顯。並顯著優於相近參數的 Gemma 系列模型， 0.5B 模型提升約 5-7% R-1 分數，達到實用水平。
- 相比訓練前，模型在 Judge 分數提升超過 0.16，顯示訓練策略有效的改善輸出的自然程度與資訊涵蓋度。

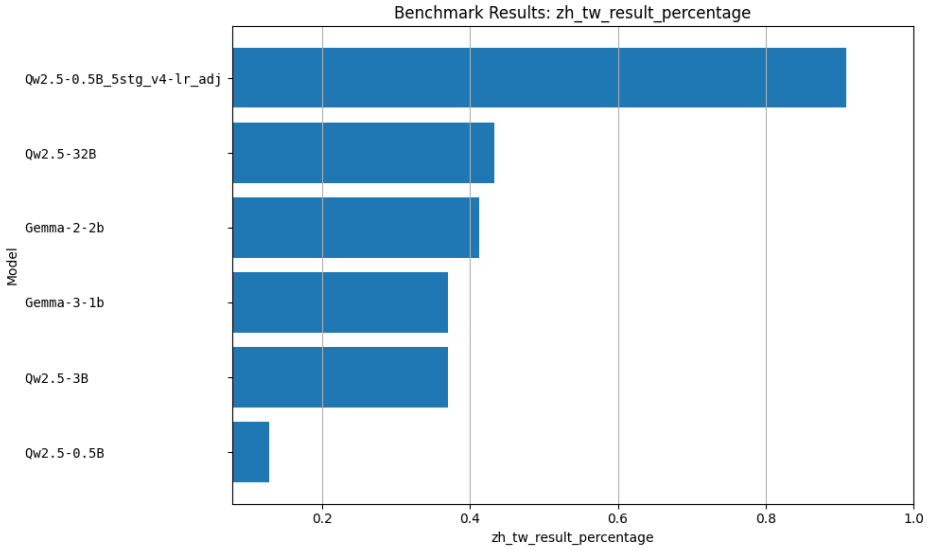
表 7：最終模型與其他相近參數量的模型的評分表現

MODEL	R-1	B-F1	Judge	R-2	R-L
Qw2.5-32B	**55.6**	**82.2**	81.7	**34.9**	**48.2**
Qw2.5-3B	49.5	79.8	**83.8**	26.2	40.1
Qw2.5-0.5B_5stg_v4-lr_adj	46.5	78.6	75.1	23.9	38.3
Llama-3.2-3B	43.3	76.8	74.3	20.7	34.7
Gemma-3-1b	41.8	76.0	72.6	18.3	31.2
Qw2.5-0.5B	39.7	74.8	58.7	18.5	30.9
Gemma-2-2b	39.3	71.9	83.4	18.3	26.2
Llama-3.2-1B	38.7	74.1	63.4	16.9	29.5
DeepSeek-R1-1.5B	32.2	71.3	65.8	12.0	21.4

另外，在訓練後模型簡體中文文字的比例最高，甚至超越了教師模型本身。完全不含簡體中文的回答比例，更是高出其他模型不少。（如下圖）



圖三：回答中簡體中文文字的比例



圖四：回答完全不含簡體中文的比例

五、 結論

1. 資料生成策略方面，實驗發現教師模型不擅長以推理方式直接生成摘要，易產生不連貫或失焦內容；相較之下，先生成摘要，再利用摘要生成關鍵要素與三元組效果較佳，這可能是因為在 Qwen2.5 模型訓練時，模型沒有學習到這種推論的過程。
2. 在訓練策略方面，採用遞減學習率訓練效果效果較差；此外，直接微調全部參數的模型表現效果更好，沒有必要使用 LoRA 或凍結部分參數的微調

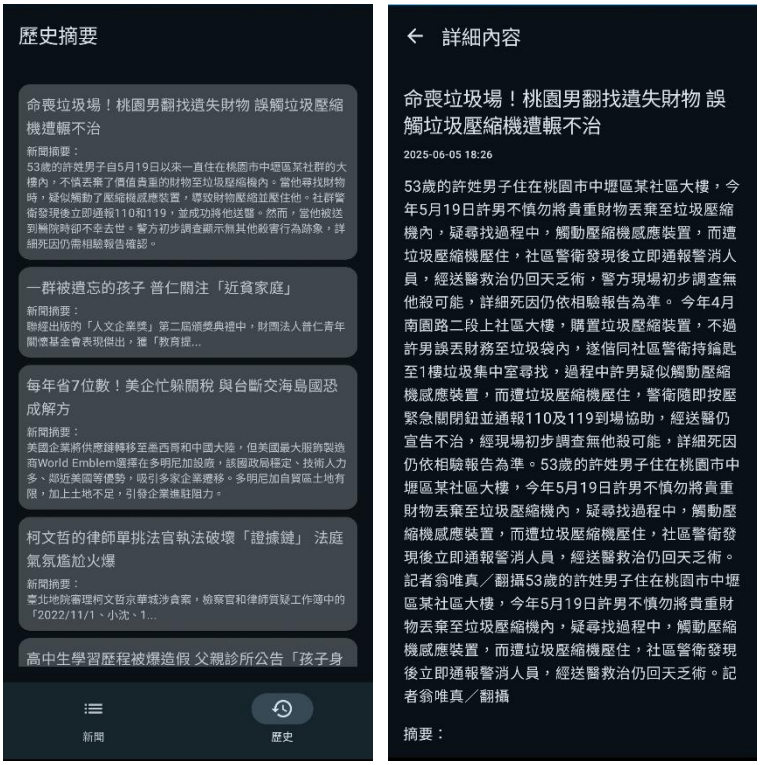
策略。

3. 在大規模訓練資料下，課程式學習對生成品質提升有限，但能提升繁體中文生成的準確率；而在僅使用約 10 萬篇新聞預訓練、並以少量資料微調的自訂義小模型中，課程式學習顯著提升模型的摘要準確度。這可能是因為在這樣的訓練量與參數之下，無法大幅改變原模型理解的方式，讓模型學習到推論過程。
4. 0.5B 模型透過蒸餾與課程訓練已可提供接近 3B 模型水準的摘要能力，透過精煉訓練集，也避免了出現已知教師模型的問題（如異常 token、錯誤譯字等）。
5. 總結：本研究成功展示了結合模型蒸餾與課程式訓練策略，可有效訓練出在效能、品質與資源消耗之間取得平衡的中文新聞摘要小模型。實驗結果表明，透過合理設計的資料生成方式與階段式訓練流程，即使在參數規模僅 0.5B 的條件下，亦能達到接近大型模型的輸出品質，並優於其他同級模型，實現「輕量＋高品質」目標，具有高潛在應用市場，也為中文新聞摘要領域的輕量化模型應用提供了可行且具潛力的實作框架。

六、 手機部署

我使用 MLC-LLM 作為開發的主要套件，使用其模板與 Android Studio 進行手機部署與開發，以下為程式部分畫面。

圖五、六：手機程式畫面：使用者可選取文章 → 手機端生成摘要



圖七、八：手機程式畫面：使用者也瀏覽歷史摘要以及新聞全文

七、 參考文獻

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. Advances in neural information processing systems, 30.

Pengcheng Jiang, Cao Xiao, Zifeng Wang, Parminder Bhatia, Jimeng Sun, Jiawei Han. 2024. TriSum: Learning Summarization Ability from Large Language Models with Structured Rationale. arXiv preprint arXiv: 2403.10351.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv: 2106.09685.

八、 附錄

專題程式碼: https://github.com/BennyWang1007/Individual-studies

應用程式程式碼: https://github.com/BennyWang1007/Individual-studies-app

應用程式下載: https://github.com/BennyWang1007/Individual-studies-app/releases/

生成 V1 的 prompt：

System prompt:
給定一份文章，完成以下任務：
(1) 提取新聞的關鍵要素，關鍵要素應為關鍵短句、名詞或事實。
(2) 對於每個關鍵要素，檢索詳細的三元組，格式為 [實體 1| 關係 | 實體 2]，這些三元組用於構成摘要。
(3) 使用檢索到的三元組撰寫一份摘要。
核心要素、三元組和撰寫的摘要應該在同一份回應中，並以換行符分隔。所有三元組 [實體 1| 關係 | 實體 2] 的長度必須為 3（以 "|" 分隔）。
範例：
==========範例==========
提示：
[新聞]: [新聞]

更新：
核心要素：
[關鍵要素 1]、[關鍵要素 2]、[關鍵要素 3]、...

三元組：
[實體 1_1| 關係_1| 實體 1_2]
[實體 2_1| 關係_2| 實體 2_2]
[實體 3_1| 關係_3| 實體 3_2]
...

生成摘要：
[摘要]

User prompt:
[新聞]: {news}

生成 V2 關鍵要素的 prompt：

System prompt:
請根據以下新聞內容，提取新聞的關鍵要素，關鍵要素應為關鍵短句、名詞或事實，請用中文回答，並且不要使用任何標點符號。請將每個關鍵要素用[]與、分隔。例如：
關鍵要素：
[關鍵要素 1]、[關鍵要素 2]、[關鍵要素 3]

User prompt:
新聞：
{news}

生成 V2 三元組的 prompt：

System prompt:
請根據以下新聞內容與關鍵要素，檢索詳細的三元組，格式為 [實體 1| 關係 | 實體 2]，這些三元組用於構成摘要，請用中文回答，並且不要使用任何標點符號。所有三元組用[]與、分隔，且長度必須為 3。
例如：
三元組：
[實體 1_1| 關係_1| 實體 1_2]、[實體 2_1| 關係_2| 實體 2_2]、...

User prompt:
新聞：
{news}

關鍵要素：
{essential_aspects}

生成 V2 摘要的 prompt：

System prompt:
請根據以下新聞內容與檢索到的關鍵要素以及三元組，為新聞生成一份摘要，請用繁體中文回答。
例如：
生成摘要：

User prompt:
新聞：
{news}

關鍵要素：
{essential_aspects}

三元組：
{triples}

生成 V3 摘要的 prompt：

System prompt:
請根據以下新聞內容，為新聞生成一份 100 字內精簡的摘要，請用繁體中文回答。
例如：
生成摘要：

User prompt:
新聞：
{news}

生成 V3 關鍵要素與三元組的 prompt：

System prompt:
請根據以下新聞內容以及摘要，提取新聞的關鍵要素與三元組，關鍵要素應為關鍵短句、名詞或事實，三元組應為[實體 1｜關係｜實體 2]的格式，這些三元組用於構成摘要，請用繁體中文回答。請將每個關鍵要素與三元組用[]與、分隔。例如：
關鍵要素：
[關鍵要素 1]、[關鍵要素 2]、[關鍵要素 3]、...

三元組：
[實體 1_1｜關係_1｜實體 1_2]、[實體 2_1｜關係_2｜實體 2_2]、...

User prompt:
新聞：
{news}

摘要：
{summary}

範例新聞、關鍵要素、三元組、摘要（V3）：

新聞：

母愛不分物種。動保組織接獲民眾通報發現草叢有一窩胖嘟嘟奶汪，沒想到牠們的母親為了照顧這些孩子把自己餓成皮包骨，還有嚴重營養不良跟脫水狀況，對比起小狗們都肥碩健康，狗媽媽更是令人心疼。

根據 The Dodo 報導，美國密蘇里州(Missouri)聖路易斯流浪動物救援組織(Stray Rescue of St. Louis,SRSL)日前接獲民眾通報，說草叢裡面發現一窩肥胖的奶汪，但卻找不到狗媽媽，希望他們可以派人來協助一下。動物救援組織工作人員湯姆森(Natalie Thomson)表示，當他們趕往民眾通報的現場，的確真的看到一窩被照顧得好好的奶汪，很像是被人飼養後遺棄在附近。

但令動保人員意外的一幕出現了，他們過沒多久在附近的草叢找到了狗媽媽，可是這隻渾身骨瘦如柴、幾乎可以用皮包骨形容的黃狗顯然非常營養不良，跟牠一窩肥壯幼崽形成強烈對比，這隻狗媽媽顯得有些害怕人類，但牠並沒有逃跑或是圖保護孩子，反而是將目光投射在小狗身上，希望眼前的人類不要傷害牠的孩子。

每一隻狗寶寶都相當健康可愛甚至還有點肥。 (圖/取自 Stray Rescue of St. Louis 官網)

後來這一窩共 10 隻奶汪跟牠們孱弱的母親都被聖路易斯流浪動物救援組織帶回收容所，經過健康檢查後反而讓獸醫跟動保人員更難過了，因為這 10 隻小奶狗除了有點寄生蟲問題之外，沒有任何營養不良的狀況，甚至還有些過重，但狗媽媽卻嚴重營養不良還脫水，可以說為了照顧孩子鞠躬盡瘁。

目前這一窩小奶汪都受到良好的照顧，不日將可開放認養，而狗媽媽則因為身體虛弱還需要靜養一段時間才可以考慮出養。湯姆森說，這隻狗媽媽其實是很溫柔的，只是個性比較慢熟，所以需要有愛心跟耐心的飼主陪伴，牠就會慢慢敞開心房願意相信人類。

摘要：

美國密蘇裡州一窩小狗被發現肥胖健康，但牠們的母親卻因照顧孩子而嚴重營養不良、脫水。這 11 隻狗已被救出，小狗們即將開放認養，母親則需繼續靜養。救援組織尋找有耐心的飼主，以陪伴這位溫柔但慢熟的母親。

關鍵要素：

[美國密蘇裡州, 一窩小狗, 肥胖健康, 母親嚴重營養不良, 脫水, 已被救出, 小狗們即將開放認養, 母親需繼續靜養, 救援組織尋找有耐心的飼主, 溫柔但慢熟的母親]

三元組：

[美國密蘇裡州 | 發現 | 一窩小狗], [小狗 | 肥胖健康 | 狗媽媽], [狗媽媽 | 照顧孩子 | 嚴重營養不良], [狗媽媽 | 照顧孩子 | 脫水], [一窩小狗 | 被 | 救出], [小狗們 | 即將 | 開放認養], [狗媽媽 | 需 | 繼續靜養], [救援組織 | 尋找 | 有耐心的飼主], [狗媽媽 | 是 | 溫柔但慢熟的]

模型評分 prompt

System prompt:
你是一位語言評估專家。你的任務是根據文章與標準摘要，評估模型生成的摘要品質。
請根據以下評分標準，從 0 到 20 為其打分：
0：格式不正確或無意義的文字。
1：完全無關，與文章毫不相干。
2：虛構內容，語意不明。
3：嚴重誤解，包含重大錯誤。
4：幾乎無法反映原文，非常不完整。
5：文法錯誤，缺乏連貫性與相關性。
6：內容不完整且部分離題。
7：遺漏關鍵要點，有輕微虛構。
8：摘要過於模糊，缺乏具體性。
9：簡潔，涵蓋大部分重點。
10：可理解但可能遺漏細節。
11：忠實但略有遺漏。
12：大致正確但稍顯冗餘。
13：準確、結構良好，但有輕微風格問題。
14：涵蓋完整、清晰，語氣尚可改進。
15：清楚、忠實且具風格。
16：簡潔優雅，涵蓋所有重點。
17：非常接近理想摘要，僅有些微瑕疵。
18：優秀的摘要，易讀且內容完整。
19：幾近完美，僅可做細微風格潤飾。
20：完美——清楚、忠實、完整且優雅。
請回傳"分數："加一個整數分數（0–20），接著是一句簡短的理由（例如：「分數：17—— 非常接近理想摘要，僅有些微瑕疵」）。

User prompt:
文章：
{article}

標準摘要：
{ground_truth}

模型生成摘要：
{response}