NLP 期末專題報告

張瑞顯　李宜蓁
Institute of Artificial Intelligence Innovation
National Yang Ming Chiao Tung University
{ruei.ii14, azure.ii14}@nycu.edu.tw

摘要

本專案旨在探討大型語言模型（Large Language Models, LLMs）於面向層級情感分析（Aspect-Based Sentiment Analysis, ABSA）任務中處理隱式情感時所面臨的困難，並分析其可能的錯誤來源，進一步研究可行的緩解與改善方法。為解決模型在學習過程中對高頻類別產生偏重的問題，本研究分為兩組實驗方法：在 pipeline base method 中引入 Focal Loss 作為損失函數，以降低常見類別對訓練的主導影響，此外，我們結合 FGM（Fast Gradient Method）對抗訓練技術與 EMA（Exponential Moving Average, 指數移動平均機制，透過加入微小擾動以提升模型的穩定性與泛化能力，同時採用 R-Drop 正則化方法與 Cosine Learning Rate Scheduler，使模型在訓練過程中具備更平滑且穩定的收斂行為；而在 LLM base method 中，我們引入提示工程應用 (Prompt Engineering) 與質採樣增強推論 (Enhanced Inference via Heterogeneous Sampling) 提升模型在推論時的精準度。實驗結果顯示，透過上述方法的整合應用，模型在 ABSA 任務上的整體表現獲得些許提升。

1　專案概述

1.1　研究動機

現今大型語言模型（Large Language Models, LLMs）在理解與回應人類指令方面已展現出相當成熟且穩定的能力，能夠在多數情境下掌握使用者的語意意圖。然而，當模型面對更細緻且結構複雜的情感分析任務時，仍存在值得深入探討的挑戰。在給定單一句評論的情況下，針對複合式情感分析任務——亦即面向層級情感分析（Aspect-Based Sentiment Analysis, ABSA），模型通常能夠對顯式評價做出相對正確的判斷；然而，對於隱式評價，模型是否仍能準確識別其情感傾向，仍有待驗證，此外，於此類情境中，LLM 亦可能出現幻覺（hallucination）或過度推論（over-interpretation）的問題。因此，現階段模型在細緻語意理解與情感判別方面，仍存在相當大的改進空間。

1.2　重大挑戰

1. 在處理隱式評價時，模型對文本語意往往容易產生幻覺或不正確的解讀，進而影響情感判斷的準確性。
2. 既有資料集中標籤分布不均，部分高頻標籤占據主導地位，導致樣本數較少的標籤在訓練過程中容易被模型忽略。

1.3　問題解決

1. 調整模型的學習模式，以降低其在情感判斷過程中對評價內容的過度生成與推論。

2. 針對既有資料集進行分析與處理，以緩解訓練資料分布不均所導致的模型學習偏置。

3. 引入推論階段的投票機制，以過濾並去除模型產生之過度推論評價。

2 相關研究

在提升 Aspect-Based Sentiment Analysis (ABSA) 模型穩定性與穩健性的研究中，前人提出了多項關鍵技術。Ian J 等人提出的快速梯度法（Fast Gradient Method, FGM）[1]，透過在 Embedding 層注入微小擾動，顯著增強了模型對抗文本噪聲的強健性；而 Polyak 等人所開發的權重平均技術 [2]，後經演進為指數移動平均（EMA），被廣泛證明能透過維護影子參數（Shadow Parameters）來平滑訓練過程中的震盪，從而獲取更具泛化能力的模型解。此外，Loshchilov 等人提出的 AdamW 優化器與餘弦退火排程（Cosine Annealing Scheduler）[3]，相較於傳統線性下降法，能使模型在訓練後期以更精細的步長收斂，這對於處理如 Task 3 般複雜的多任務收斂至關重要。

針對模型內部的結構一致性與類別分佈不均挑戰，後續研究提供了有效的解決方案。Wu 等人提出的 R-Drop 技術 [4]，透過最小化同一輸入在不同 Dropout 採樣下預測分佈的 KL 散度，強迫模型內部的子網路達成語義共識，進一步穩固了模型在細粒度抽取任務中的表現。而在處理標籤分佈極度傾斜的屬性分類時，Lin 等人提出的 Focal Loss [5] 藉由動態調整損失權重，使訓練過程能更專注於難學的少數類別。本專案整合上述技術，構建一個能平衡精確度與召回率的增強型 ABSA 系統。

作為 LLM Based Model，已有多項工作針對生成式模型於情感分析任務中的推理與生成行為提出改進方法。Bai 等人透過提示工程的設計 [6]，限制生成器在輸出內容上的自由度，以降低不必要的擴寫與錯誤生成。Jun 等人則提出先生成情感三元組，再進一步預測 VA（Valence–Arousal）數值的策略 [7]，引導模型在做出情感強度判斷前，先完成結構化語意資訊的抽取。另一方面，Wang 等人提出以溫度控制進行多次取樣的方法 [8]，透過整合多條推論路徑來提升模型推理結果的穩定性與穩定性。

3 方法論

在本次專案中，我們分為 Pipeline base 及 LLM base 去做改良：

3.1 Pipeline base method

3.1.1 Focal Loss

在 Task 3（抽取 Aspect-Opinion-Category-VA）中，Category（如食物、價格、服務等）的分布通常極度不均，如表 1統計。

1. **為什麼要用 Focal Loss？**

   - **多數類別 (Majority Classes)**：某些常見類別（如 FOOD）樣本極多，模型容易過度擬合。
   - **少數類別 (Minority Classes)**：某些特定類別（如 LOCATION 或 AMBIENCE）樣本極少，傳統的交叉熵損失（Cross Entropy）會讓模型忽略這些難學的少數類。

表 1: 訓練資料中各個類別的數量

類別	數量
FOOD	7317
RESTAURANT	378
AMBIENCE	235
DRINKS	307
LOCATION	37
SERVICE	249

- **簡單與困難樣本**：模型對於容易區分的類別會產生很高的置信度，Focal Loss 主要就在降低這些「簡單樣本」的權重，讓模型專注於「困難樣本」。

2. **核心公式與運作機制**

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t)$$

- p_t：模型預測正確類別的機率。
- γ (`focal_gamma`)：預設為 2.0。當模型對某個樣本很有把握（p_t 很大）時，$(1-p_t)^{\gamma}$ 會變得很小，大幅削減該樣本對總損失的貢獻。
- 動態權重：如果一個樣本很難區分（p_t 很小），權重就會相對較大，強迫模型去學習這個困難樣本。

3. **進階修改**

程式碼我們不單純只用 Focal Loss，還整合了 CategoryLossEnhanced，同時做了三件事：

- **Focal Loss (γ = 2.0)**：解決難易樣本不均。
- **Label Smoothing (0.1)**：防止模型對特定標籤過於自信，增加泛化能力。
- **Class Weights**：根據訓練集中各類別出現的頻率自動計算權重，給予出現次數少的類別更高的損失權重。

4. **小結**

在程式中，Focal Loss 扮演著「類別預測優化器」的角色。它確保模型在判斷 Aspect 屬於哪個分類（如：是「食物」還是「環境」）時，不會因為「食物」樣本太多就亂猜，也不會因為「環境」樣本太少就學不會，進而提升整體的 F1 分數。

3.1.2 FGM (Fast Gradient Method)

FGM (Fast Gradient Method) 是一種對抗訓練 (Adversarial Training) 技術。它的核心邏輯是在輸入的 Embedding 層中加入微小的擾動，模擬「對抗樣本」，迫使模型在面對雜訊時依然能保持正確的預測，進而提升模型的穩固性 (Robustness) 與泛化能力。

1. **FGM 的數學原理**

FGM 的概念源於 Fast Gradient Sign Method (FGSM)，其目標是尋找一個擾動 r_{adv}，使得模型的損失函數 L 最大化：

$$r_{adv} = \epsilon \cdot \frac{g}{||g||_2}$$

其中：

- g 是損失函數對輸入 Embedding 的梯度。
- $||g||_2$ 是梯度的 L2 範數（正規化）。
- ϵ (`adv_epsilon`) 是擾動的大小，預設為 1.0。

2. **FGM 特殊設定**

- **梯度累積 (Gradient Accumulation)**：
  程式碼中判斷了 $(\text{batch\_index} + 1)\%$ `args.gradient_accumulation_steps` $= 0$。這代表 FGM 不是在每一個小 batch 都攻擊，而是在即將更新模型參數的那一刻才進行對抗訓練，這樣可以節省計算資源。
- **局部攻擊策略**：
  它只針對 `forward_aspect`（前向 Aspect 抽取）這個核心任務進行對抗訓練。這是因為 Aspect 抽取是整套模型（Task 2 & 3）的基礎，如果 Aspect 抓不準，後續的 Opinion 和 Category 都會錯。

3. **小結**
加入 FGM 有以下幫助：

- **防止過擬合**：FGM 相當於一種強大的正則化手段，防止模型過度依賴訓練集中的特定字眼。
- **應對數據噪聲**：在多語言或餐廳/飯店評論中，用戶的用語往往不規範。FGM 讓模型學會忽略那些不影響語義的微小特徵變化。
- **提升 F1 分數**：在 ABSA 競賽中，加入 FGM 通常能穩定提升 0.5% - 1.5% 的 F1 值。

3.1.3 EMA 平滑

EMA (Exponential Moving Average, 指數移動平均) 扮演的是一個「影子模型」的角色。它不是直接改變模型的訓練方式，而是透過平滑模型參數，來提升模型在測試集上的泛化能力和穩定性。

1. **EMA 的基本原理**
在訓練過程中，模型的權重(W)會在每個 Step 劇烈波動。EMA 會維護一組「影子參數」(Shadow Parameters, W_{ema})，它是歷史權重的加權平均：

$$W_{ema}^{(t)} = \text{decay} \cdot W_{ema}^{(t-1)} + (1 - \text{decay}) \cdot W^{(t)}$$

- **Decay(衰減率)**：預設為 0.999。這代表新參數只佔 0.1%，而過去的累積經驗佔 99.9%。
- **平滑效應**：可以濾除訓練過程中的隨機噪聲，使模型最終停留在一個更平緩、更寬大的損失函數局部最小值中，通常代表泛化更好。

2. **程式碼中的三個關鍵階段**
EMA 的運作貫穿了整個 train() 函數，主要分為：**更新、應用、恢復**。

- **初始化與更新 (訓練期間)**
  在每個 optimizer.step() 之後，EMA 都會被調用來更新影子參數。

- **應用影子參數 (驗證/推理期間)**
  在進行 Dev 驗證或最終 Inference 時，我們不使用當時正在訓練的那套權重，而是臨時換成「影子權重」。
- **恢復原始參數 (繼續訓練前)**
  評估完畢後，需要把模型換回原本的訓練權重，否則會破壞優化器的狀態。

3. **小結**
   加入 EMA 有以下幫助：

   - **防止過擬合 (Overfitting)**：ABSA 數據集（尤其是中文餐廳或筆電）通常規模較小，模型很容易在訓練集上產生波動。EMA 透過回顧歷史參數，防止模型過度擬合最後幾個 Batch 的特徵。
   - **提高預測穩定性**：我們程式碼中包含多個任務（Aspect, Opinion, Category, VA），損失函數非常複雜。EMA 能確保模型在不同任務之間取得平衡，不會因為某個 Batch 的 Loss 偏高就讓模型參數跑偏。

3.1.4 R-Drop 正則化

R-Drop (Regularized Dropout) 是一種非常強大的正則化技術。它的核心思想非常簡單：讓模型在進行兩次不同的 Dropout 隨機失活時，對於同一個輸入，產生的預測結果盡可能一致。

1. **R-Drop 的數學原理**
   在 Transformer 模型中，Dropout 是隨機的。如果我們把同一個輸入 x 送進模型兩次，因為 Dropout 掉的神經元不同，會得到兩個略有差異的輸出分佈 $P_1(y|x)$ 和 $P_2(y|x)$。
   R-Drop 的目標就是最小化這兩個分佈之間的 KL 散度 (Kullback-Leibler Divergence)：

$$\mathcal{L}_{R-Drop} = \frac{1}{2}(KL(P_1||P_2) + KL(P_2||P_1))$$

   最終總損失函數為：

$$\mathcal{L}_{Total} = \mathcal{L}_{Task} + \alpha \cdot \mathcal{L}_{R-Drop}$$

   其中 α (`rdrop_alpha`) 是控制正則化強度的係數。

2. **為什麼 R-Drop 對的 ABSA 任務有幫助？**

   - **結構一致性**：Dropout 會導致神經網路在訓練時結構不斷變化。R-Drop 強迫這些隨機生成的網路子集（Sub-networks）在語義層面達成共識，這對於需要精確判斷 Aspect 和 Opinion 邊界的任務非常有幫助。
   - **減少過擬合**：ABSA 的標籤（如 Category）分佈稀疏，模型容易死記硬背。R-Drop 增加了一個約束，讓模型必須學習到更穩健的特徵，而不是依賴特定的神經元通路。
   - **輔助 Focal Loss**：
     - Focal Loss 解決了類別不平衡（讓模型關注難學的樣本）。
     - R-Drop 解決了預測不確定性（讓模型對難學樣本的預測更穩定）。這兩者結合通常能讓 Task 3 的 F1 分數有感提升。

3. **小結**
   R-Drop 就像是讓模型進行「自我對話」：針對同一句話，模型內部的兩個隨機分身如果意見分歧（KL 散度大），就會受到懲罰。最終，模型會學會產生一個不論 Dropout 如何隨機，都始終如一的穩健預測。

3.1.5 Cosine Scheduler

Cosine Scheduler（`get_cosine_schedule_with_warmup`）是用來動態調整學習率 (Learning Rate) 的策略。它讓學習率不再是一個固定的數值，而是隨著訓練進度呈「餘弦函數」曲線下降。

1. **運作的三個階段**
   這套策略將整個訓練過程（由 `training_steps` 決定）分為兩個主要階段：
   - **第一階段：線性預熱 (Warmup)**
     - **動作**：學習率從 0 開始，在 `warmup_steps` 內線性增加到你設定的初始學習率（例如 3e-5）。
     - **目的**：在訓練初期，模型的權重是隨機初始化的，梯度可能非常不穩定。預熱可以防止模型在第一步就因為過大的梯度而「跑偏」或產生梯度爆炸。
   - **第二階段：餘弦退火 (Cosine Decay)**
     - **動作**：達到頂峰後，學習率會按照餘弦波形的後半段曲線緩慢下降，直到訓練結束時接近 0。
     - **目的**：在訓練後期，模型接近最優解，此時需要微小的步長來精細調整參數。餘弦曲線在中間下降較快，在末尾下降非常平緩，這有助於模型跳過局部最小值並最終穩定在一個更好的最優點。
2. **為什麼 Cosine 比 Linear 更好？**
   在原版的程式中通常使用 Linear（線性下降），但我們使用了 Cosine，原因如下：
   - **更平滑的收斂**：線性下降在末尾下降得太「急」，而餘弦曲線在最後階段（Cosine Annealing）非常平滑，這讓模型有更多時間在最優解附近進行「微調」，通常能提升最終的 F1 分數。
   - **逃離局部最優**：餘弦函數中段下降較快，這種動能有時能幫助模型衝過一些差勁的局部極小值（Local Minima）。
   - **配合多任務學習 (Task 3)**：由於 Task 3 需要同時學習 Aspect、Category 和 VA，損失函數非常複雜。穩定的學習率下降曲線能確保不同任務的 Loss 都能協調地下降，而不會互相干擾。
3. **與其他組件的協作**
   - **與 EMA 協作**：Cosine 讓參數在後期變動非常微小，這與 **EMA (指數移動平均)** 的平滑特性完美契合，兩者結合產生的「影子模型」會極度穩定。
   - **與 AdamW 協作**：AdamW 負責處理每個參數的自適應縮放，而 Cosine Scheduler 則負責控制整體的「節奏」。

3.2 LLM base method

3.2.1 提示工程應用 (Prompt Engineering)

在將生成式模型實際應用於系統時，提示工程對生成結果的穩定性、可控性與專業一致性具有關鍵影響。因此，在本專題中，我們參考 Is Compound Aspect-Based Sentiment Analysis Addressed by LLMs? [6] 所提出的核心概念，透過設計明確且結構化的指令，引導模型在推理與生成過程中維持一致的輸出邏輯，進而提升整體生成品質。

在目標設計方面，我們亦借鑑 Dynamic Order Template Prediction for Generative Aspect-Based Sentiment Analysis [7] 的研究成果。該研究指出，在生成式 ABSA 任務中，目標序列的排列順序會直接影響模型在解碼階段的條件依賴關係，進而左右其推理效能。適當的生成順序有助於模型更有效地捕捉不同元素之間的語意關聯。

基於上述觀察，我們捨棄傳統隨機或任意的輸出排列方式，並提出「先證據、後結論(Evidence-to-Conclusion)」的生成策略。具體而言，模型需先生成具體且可驗證的情感三元組（Aspect–Category–Opinion Triplet），再依據這些已產生的文本證據，進一步預測較為抽象的 VA（Valence–Arousal）數值。

System Prompt 設計：

```
Task: Extract a nested JSON list of sentiment quadruplets from the input text.

Output Format Requirements:
1. 'triplet':
   - 'aspect_category': Choose strictly from the constraint list.
   - 'aspect_term': The target entity (use 'NULL' if implicit).
   - 'opinion_term': The sentiment phrase
     (Must be an EXACT substring of the input).

2. 'valence_arousal':
   - Format: 'V.VV#A.AA' (1-9 scale)
   - Valence: 1(Negative) <-> 5(Neutral) <-> 9(Positive)
   - Arousal: 1(Calm) <-> 5(Moderate) <-> 9(Excited)

Constraint:
The 'aspect_category' is formed by combining an Entity
and an Attribute (Format: Entity#Attribute).

Valid Entities: {entities}
Valid Attributes: {attributes}
```

此外，我們在提示詞中引入 few-shot learning 架構，以示範模型的推理與輸出方式，範例如下：

```
Input: 這家餐廳服務很好，但價格有點貴。
```

```
Output: [
  {
    "triplet": {
      "aspect_category": "SERVICE#GENERAL",
      "aspect_term": "服務",
      "opinion_term": "很好"
    },
    "valence_arousal": "8.00#6.50"
  },
  {
    "triplet": {
      "aspect_category": "RESTAURANT#PRICES",
      "aspect_term": "價格",
      "opinion_term": "有點貴"
    },
    "valence_arousal": "3.50#4.00"
  }
]
```

小結：透過上述提示設計，我們能夠觀察到以下重點：

- 生成過程中透過明確的輸出約束，使模型僅能在指定的意見詞 (Opinion) 與面向 (Aspect) 層級範圍 (Constraint List) 內進行生成，有效降低不必要的擴寫與過度推論。
- 模型被引導先生成具體的情感三元組，再基於已生成的結構化結果預測對應的 VA 數值，使推理流程更為清晰且具可解釋性。
- 透過引入 few-shot 指令範例，模型得以學習情感三元組與 VA 數值之間的生成關係，進一步提升輸出格式與語意的一致性。

3.2.2 質採樣增強推論 (Enhanced Inference via Heterogeneous Sampling)

在推論階段中，我們參考 Self-Consistency Improves Chain of Thought Reasoning in Language Models [8] 所提出的 Self-Consistency 方法。該方法通常於推論時採用固定的溫度參數進行多次取樣。然而，我們認為 Aspect 抽取任務在本質上需要一種雙重策略：對於明確提及的項目，應著重於高精確度（High Precision）；而對於隱晦或間接表達的情感，則需兼顧較高的召回率（High Recall）。

基於強化學習中探索與利用（Exploration–Exploitation）的核心原則，我們實作了一種異質採樣策略（Heterogeneous Sampling Strategy），並採用溫度排程

$$T = \{0.1, 0.5, 0.5, 0.5, 0.7\}$$

於推論階段進行多樣化取樣。具體而言：

- $\tau = 0.1$（利用 / Exploitation）：透過最大化生成結果的概似性（Likelihood），精準抽取語意明確且具高置信度的 Aspect。
- $\tau = 0.7$（探索 / Exploration）：提升生成結果的多樣性，以捕捉在貪婪式解碼（Greedy Decoding）或低溫取樣下容易被忽略的隱晦或結構較為複雜之情感表達。

小結：透過此設計有效防止了模式崩潰 (Mode Collapse)——即多條推論路徑同時產生完全相同錯誤的現象，進而顯著提升了多數決投票機制 (Majority Voting) 的穩健性 (Robustness)。

4 實驗設置

4.1 Pipeline base method

在此次實驗中，我們使用 Pre-Trained Models for Chinese Natural Language Processing (Cui et al., 2020)[9] 所提出的 RoBERTa-wwm-ext-large 模型，模型使用中文維基百科、新聞文本和各式網路語料進行訓練，使其語言理解能力強於原始英文 RoBERTa 的中文版本。我們訓練相關配置、介紹如以下：

- Hidden Size: 1024
- Layers: 24
- Batch size: 2(restaurant)、1(laptop)
- Epoch: 40
- Parameters: 約 325M
- Whole Word Masking (WWM)
  選用全詞遮蔽策略：例如「國際化」不會被拆成國 + 際 + 化來獨立遮罩，而是整個詞一起遮住，優點是：
  - 增加語義學習的完整性
  - 更適合中文多字詞語義的特性
- EXT：Extended Training
  除了 WWM，該模型也進行更多訓練步數與更大 batch size，因此語言知識更加紮實，有助於下游任務表現提升。

基於上述的模型和先前提出的方法，有以下參數為我們著重觀察和調整的方向:

- 訓練週期：從 3 到 50 個 epoch 不等，考察模型收斂特性。
- 學習率範圍：在 1e-3 到 5e-5 之間調整，平衡訓練速度與穩定性。
- 正則化強度：系統性地調整各項正則化參數組合。

4.2 LLM base method

本專案採用 Qwen3 Technical Report（Alibaba Cloud, 2024）[10] 提出的 Qwen-3-14B 作為基礎大型語言模型，並透過 QLoRA 進行參數高效微調。作為目前開源社群中同量級模型中表現最為突出的模型之一，Qwen 系列在 MMLU、C-Eval 等多項基準測試中展現了優異的通用推理能力。特別是在中文語境理解與複雜指令遵循（Instruction Following）方面，Qwen 顯著優於同級模型（如 LLaMA-3-8B），使其非常適合應用於本專案所關注的細粒度語意分析與結構化 JSON 生成任務。

模型微調所使用的相關參數如下：

- `LORA_R = 16`
- `LORA_ALPHA = 32`
- `LORA_DROPOUT = 0.05`
- `BATCH_SIZE = 1`
- `GRADIENT_ACCUMULATION_STEPS = 16`
- `WARMUP_RATIO = 0.03`
- `NUM_TRAIN_EPOCHS = 3`
- `LR_SCHEDULER_TYPE = cosine`
- `MAX_GRAD_NORM = 0.3`
- `WEIGHT_DECAY = 0.01`
- `LEARNING_RATE` = $5 \times 10^{-4} \sim 1 \times 10^{-3}$
- `target_modules = [q_proj, k_proj, v_proj, o_proj, up_proj, down_proj, gate_proj]`
- `inference_thereshold = 2`

4.3 評估指標

Evaluation Metrics 為評估模型於 ABSA 任務中的表現，本專案採用以下指標進行分析與比較：

- **True Positive (TP)**
  TP 表示模型正確預測為正例，且與標註結果一致的樣本數量。在本任務中，意指模型成功且正確地抽取出實際存在的 Aspect 或情感標註。
- **False Positive (FP)**
  FP 表示模型錯誤地將不存在或不正確的項目預測為正例的數量，反映模型產生多餘或錯誤預測的情形，常用於衡量過度推論（over-generation）的程度。
- **False Negative (FN)**
  FN 表示模型未能預測出實際存在之正例的數量，代表模型在抽取或判斷過程中遺漏正確標註的情況。

- **Correct True Positive (CTP)**
  CTP 用於表示在較嚴格評估條件下，被視為完全正確的預測數量。於複合式 ABSA 任務中，CTP 通常要求模型在 Aspect、Opinion 及情感類別等層面皆與標註結果完全一致。

- **Category Precision (cPrecision)**
  cPrecision 衡量模型所有預測為正例的結果中，有多少比例在類別層級上是正確的，其定義如下：
  $$\text{cPrecision} = \frac{TP}{TP + FP}$$
  此指標反映模型預測結果的準確性，當 FP 增加時，cPrecision 會隨之下降。

- **Category Recall (cRecall)**
  cRecall 衡量模型成功找回實際存在正例的比例，其定義如下：
  $$\text{cRecall} = \frac{TP}{TP + FN}$$
  此指標用以評估模型對正確樣本的涵蓋能力，FN 越高，cRecall 越低。

- **Category F1-score (CF1)**
  CF1 為 cPrecision 與 cRecall 的調和平均，用以綜合評估模型在準確性與完整性之間的平衡，其定義為：
  $$\text{CF1} = \frac{2 \times \text{cPrecision} \times \text{cRecall}}{\text{cPrecision} + \text{cRecall}}$$
  CF1 能有效避免模型僅偏重 Precision 或 Recall，特別適合用於標籤分布不均的 ABSA 任務。

5 實驗結果

5.1 Pipeline base method

表 2: Performance of our improved pipeline base method on restaurant dataset

Epoch	Learning rate	adv_epsilon	label_smoothing	ema_decay	focal_gamma	beta	inference_beta	drop_alpha	未輸出數量	cF1
3	1e-3	x	x	x	x	1	0.9	4.0	16	0.5757
20	2e-5	1.0	0.2	0.999	2.5	1.5	0.82	4.0	5	0.5393
30	2e-5	1.0	0.2	0.999	2.5	1.5	0.82	4.0	5	0.5561
50	2e-5	1.0	0.2	0.999	2.5	1.5	0.82	4.0	2	0.5595
40	2.5e-5	1.8	0.1	0.999	2	2	0.78	4.0	5	0.5471
40	2e-5	1.4	0.15	0.9995	2.5	2	0.83	4.0	5	0.5587
40	2e-5	1.5	0.2	0.9998	3	2	0.88	4.0	6	0.5640
40	5e-5	1.5	0.2	0.9998	3	2	0.9	1.0	9	**0.5825**

表 3: Detailed performance of our improved pipeline base method on restaurant dataset

cF1	CPERCISION	CRECALL	CTP	TP	FP	FN
0.5757	0.6049	0.5492	417.9769	437	254	324
0.5561	0.5386	0.5747	437.3369	456	356	305
0.5595	0.5316	0.5903	449.2394	468	377	293
0.5471	0.5126	0.5866	446.4345	465	406	296
0.5587	0.5370	0.5822	443.0308	462	363	299
0.5640	0.5440	0.5855	445.5330	464	335	297
0.5825	0.5609	0.6059	461.0831	481	341	280

表 4: Performance of our improved pipeline base method on laptop dataset (using the setup of best score on restaurant dataset)

cF1	CPERCISION	CRECALL	CTP	TP	FP	FN	備註
0.4106	0.4208	0.4010	220.9240	233	292	318	baseline
0.3402	0.3557	0.3260	179.6125	205	300	346	lr=5e-4
0.3441	0.3287	0.3209	198.8772	211	394	340	lr=5e-5

1. **訓練時長的影響**
   從表 2前四行的對比可以看出，在相同參數配置下 (2e-5 學習率)，隨著訓練 epoch 從 20 增加到 50，cF1 分數穩步提升 (0.5393→0.5561→0.5595)，但增幅逐漸趨緩，顯示模型在 40-50 epoch 之間趨於收斂。對應表 3的混淆矩陣分析顯示，真陽性 (TP) 從 456 增加至 468，假陰性 (FN) 從 305 降至 293，證實了訓練時長確實改善了模型的召回能力。

2. **基礎配置的重要性**
   表 2第一行實驗使用較少的正則化技術和較短的訓練時間 (3 epoch)，儘管如此仍獲得了相對競爭力的分數 (0.5757)，顯示基礎模型架構本身已具備一定能力。從表 3可見，該配置雖然 TP 較低 (437)，但 FP 控制最佳 (254)，精確率達到最高的 0.6049，展現出保守但穩健的預測策略。

3. **正則化策略的權衡**
   表 2比較第 2-7 行可見，引入對抗訓練 (`adv_epsilon`)、標籤平滑 (`label_smoothing`) 和 EMA 等正則化技術後，雖然提升了模型穩定性 (未輸出數量從 16 降至 5 左右)，但初期反而導致性能下降，需要通過細緻的參數調優才能超越簡單配置。表 3顯示，這些正則化技術雖然增加了 FP(從 254 增至 356-377)，但同時也顯著提升了 TP 和召回率，體現了精確率與召回率之間的權衡。

4. **召回率與精確率的演化趨勢**
   表 3展現了明確的性能改變過程：召回率從 0.5492 穩步提升至 0.6059(+10.3%)，同時精確率從 0.6049 下降至 0.5609(-7.3%)。特別值得注意的是，假陰性 (FN) 從 324 降至 280(改善 13.6%)，顯示模型逐漸減少了對正樣本的遺漏。

5. **參數敏感性分析**
   表 2對比第 5-8 行的漸進調整，可以發現 `inference_beta` 從 0.78 提升到 0.9、`drop_alpha`

從 4.0 降至 1.0、以及學習率從 2e-5 提升到 5e-5 是性能突破的關鍵因素，這表明推理階段的閾值策略和訓練動態調整對最終性能有顯著影響。從表 3的數據驗證，這些調整使 TP 從 462 增加至 481，CTP 從 443.03 提升至 461.08(+4.1%)，證實了參數優化的有效性。

6. **假陽性控制的挑戰**
   表 3顯示了性能優化過程中的代價：為了提高召回率，FP 從基線的 254 增加到最終的 341(+34.3%)。其中第 4 行配置 (Epoch 50) 的 FP 高達 406，顯示過於激進的參數設置會導致誤報失控。最佳配置 (第 7 行) 成功將 FP 控制在 341，相比 Epoch 50 減少了 65 個誤報，在高召回率下實現了相對合理的精確率平衡。

7. **EMA Decay 的關鍵作用**
   表 2第 6-8 行的對比顯示，將 `ema_decay` 從 0.9995 提升至 0.9998 帶來了顯著改善。結合表 3分析，這一調整使 cF1 從 0.5587 提升至 0.5640，最終達到 0.5825，召回率也從 0.5822 提升至 0.6059。這表明更高的 EMA 衰減率能更好地穩定模型參數，減少訓練波動，是性能提升的重要貢獻因素之一。

8. **最佳配置組合與綜合性能**
   最高性能 (cF1=0.5825) 出現在 40 epoch 時，採用較高學習率 (5e-5)、中等對抗訓練強度 (`adv_epsilon`=1.5)、較高的 `focal_gamma`(3.0) 和 `inference_beta`(0.9)，以及較低的 `drop_alpha`(1.0)，這組配置在探索與穩定性之間取得了良好平衡。表 3證實該配置在所有指標上表現最優：TP 達 481(最高)、FN 降至 280(最低)、召回率 0.6059(最高)，同時精確率維持在 0.5609 的可接受水平，實現了真正意義上的性能突破。

9. **遷移學習的挑戰**
   從表 4可觀察到，直接將 restaurant 數據集的最佳參數配置遷移到 laptop 數據集並未取得成功，所有改進配置的 cF1 分數（0.3402-0.3441）均顯著低於 baseline（0.4106），降幅達 17-20%。這表明不同領域數據集需要針對性的參數調優，restaurant 數據集上的最優策略（高學習率 5e-5、高 `inference_beta` 等）在 laptop 數據集上可能過於激進，導致模型欠擬合或不穩定。

10. **應用場景的配置選擇建議**
    綜合兩表分析，針對不同應用需求可選擇不同配置：

    (a) 高精確率建議採用表 2第 1 行配置，精確率 0.6049、FP 僅 254。

    (b) 高召回率建議採用表 2第 8 行配置，召回率 0.6059、FN 最低 280。

    (c) 平衡型應用可考慮表 2第 7 行配置，F1=0.5640 且各項指標相對均衡。

5.2 LLM base method

表 5: Performance of our LLM method on laptop dataset

Laptop	Learning rate	temperature	cF1	CPERCISION	CRECALL	CTP	TP	FP	FN
baseline	1e-3	0.7	0.4221	0.4169	0.4275	235.5471	246	319	305
Prompt engeering + Heterogeneous Sampling Stragegy	1e-3	[0.1,0.5,0.5,0.5,0.7]	0.4281	0.3995	0.4612	254.1051	265	371	286

表 6: Performance of our LLM method on restaurant dataset

Restaurant	Learning rate	temperature	cF1	CPERCISION	CRECALL	CTP	TP	FP	FN
baseline 1	5e-4	0.7	0.5652	0.5645	0.5660	430.7052	455	308	306
baseline 2	5e-4	0.15	0.5809	0.5786	0.5832	443.8096	467	300	294
Prompt engeering + Heterogeneous Sampling Stragegy	1e-3	[0.1,0.5,0.5,0.5,0.7]	0.5722	0.5791	0.5654	430.2750	451	292	310
Prompt engeering + Heterogeneous Sampling Stragegy	5e-4	[0.1,0.5,0.5,0.5,0.7]	0.5741	0.5876	0.5613	427.1623	448	279	313

從上述實驗結果中，我們可觀察到以下幾點關鍵現象：

1. 在表 5 Laptop 資料集上，加入所提出的方法後，TP（True Positive）數量相較於 baseline 提升近 20 筆，顯示模型對正確 Aspect 的捕捉能力有所提升；然而，FP（False Positive）數量亦同時增加約 50 筆。推測此現象可能源於多溫度取樣所帶來的預測多樣性，使模型生成較多候選結果。未來可透過引入更合適的決策閾值或投票策略，以進一步抑制 FP 的產生。

2. 在表 6 Restaurant 資料集上，加入所提出的方法後，主要觀察到 FP 值顯著下降，而 TP 的變化相對有限。此結果顯示，目前所採用的溫度設定與投票機制，對於以顯式評價為主的 Restaurant 資料集並不完全適用。未來將嘗試偏向低溫取樣的投票策略，使模型在推論階段更著重於高置信度的顯式評價預測。

3. 在表 6 Restaurant 資料集上，我們觀察到使用較高學習率進行訓練時，模型表現反而劣於採用較低學習率的設定。推測其原因在於 Restaurant 資料集相較於 Laptop 資料集具有較高的規律性與顯式評價比例，整體任務難度較低；在此情況下，過高的學習率可能導致模型在訓練階段過度擬合訓練資料，進而影響其在測試集上的泛化表現。

小結：整體而言，針對本研究所提出之方法，其在部分評估指標上已展現出性能提升的趨勢，顯示仍具有相當程度的調整與優化空間。由各項細部實驗結果可觀察到，所提出的設計確實能對模型的生成行為產生正向影響，提升其在目標任務中的判斷能力與結構化輸出品質。然而，該方法在不同資料集與設定條件下的表現仍存在差異，未來仍有必要進一步針對推論策略與參數配置進行更細緻的分析與改進，以充分發揮其潛在效益。

6 結論

6.1 Pipeline base method

這次實作針對中文情感元素抽取任務，採用 RoBERTa-wwm-ext-large 模型，並比較多種訓練強化策略之效果。實驗結果顯示，整合對抗訓練、label smoothing、EMA 與 Focal Loss 等方法，能有效提升模型對困難樣本與資料不均衡問題的處理能力。最佳實驗組合之 restaurant cF1 達 0.5825，相較基準模型有顯著提升。此結果證明，訓練流程與推論參數的精細調整，對於提升模型整體表現具有重要貢獻。

6.2 LLM base method

對於 LLM base method，我們使用 Qwen-3-14B 作為基底，結合提示工程設計與多溫度取樣投票機制進行改良，在部分評估指標上觀察到些微但一致的效能提升。此結果顯示，透過推論階段的策略調整，即使不改變模型參數本身，仍能在一定程度上改善模型於 ABSA 任務中的表現。

實驗結果亦顯示該方法的成效會隨資料集特性而有所差異。對於以意見層面詞較少的資料集，過度引入高溫取樣可能增加誤判風險；相反地，多樣化取樣與投票機制則有助於提升模型的召回能力。此現象突顯了推論策略需依任務與資料特性進行調整的重要性。

綜合而言，這次實作驗證了在生成式 ABSA 任務中，透過結構化提示設計與異質取樣策略，可作為提升大型語言模型推論穩定性的一種可行方向。未來工作將進一步探討更具適應性的溫度調度與動態投票機制，探討在不同資料分布下取得更穩定且一致的效能表現。

參考文獻

[1] Ian J. Goodfellow and Jonathon Shlens and Christian Szegedy. *Explaining and Harnessing Adversarial Examples.* arXiv preprint arXiv:1412.6572.

[2] Boris Polyak and Anatoli B. Juditsky. *Acceleration of stochastic approximation by averaging* `https://api.semanticscholar.org/CorpusID:3548228`.

[3] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization* arXiv preprint arXiv:1711.05101.

[4] Xiaobo Liang and Lijun Wu and Juntao Li and Yue Wang and Qi Meng and Tao Qin and Wei Chen and Min Zhang and Tie-Yan Liu. *R-Drop: Regularized Dropout for Neural Networks* ヌ arXiv preprint arXiv:2106.14448.

[5] Tsung-Yi Lin and Priya Goyal and Ross Girshick and Kaiming He and Piotr Dollár. *Focal Loss for Dense Object Detection* arXiv preprint arXiv:1708.02002.

[6] Bai, Yinhao and Han, Zhixin and Zhao, Yuhua and Gao, Hang and Zhang, Zhuowei and Wang, Xunzhi and Hu, Mengting. *Is Compound Aspect-Based Sentiment Analysis Addressed by LLMs?* `https://aclanthology.org/2024.findings-emnlp.460/`.

[7] Yonghyun Jun and Hwanhee Lee. *Dynamic Order Template Prediction for Generative Aspect-Based Sentiment Analysis* arXiv preprint arXiv:2406.11130.

[8] Xuezhi Wang and Jason Wei and Dale Schuurmans and Quoc Le and Ed Chi and Sharan Narang and Aakanksha Chowdhery and Denny Zhou. *Self-Consistency Improves Chain of Thought Reasoning in Language Models* arXiv preprint arXiv:2203.11171.

[9] Cui, Yiming and Che, Wnxiang and Liu, Ting and Qin, Bing and Wang, Shijin and Hu, Guoping. *Revisiting Pre-Trained Models for Chinese Natural Language Processing.* arXiv preprint arXiv:2004.13922.

[10] An Yang and Anfeng Li and Baosong Yang and Beichen Zhang and Binyuan Hui and Bo Zheng and Bowen Yu et al. *Qwen3 Technical Report.* arXiv preprint arXiv:2505.09388.