臺北市政府獎狀

府教中字第1133058622號

國立臺灣師範大學附屬高級中學

學生

葉安之　蔡昕翰

利用生成式 AI 實現音樂與故事間之互相轉換

參加「臺北市第 57 屆中小學科學展覽會」

榮獲 高中組 電腦與資訊學科 優等

市長 蔣萬安

中 華 民 國　113　年　5　月　11　日

科學翱翔 創意飛揚

壹、摘要

我們研發了一個模型架構，該架構能夠從音樂中生成情緒相似的故事。這個模型架構不僅能夠產生高品質的故事內容，而且在受試者間獲得了不錯的評價，平均得分為3.75分（滿分5分）。這表明，不僅在客觀標準下，該模型的表現出色，而且在主觀評價上也取得了成功。此外，我們還提出了兩種方法，用於創建一個反向生成函數，使其能夠從故事生成音樂。這項成就標誌著藝術與人工智慧技術的成功融合，我們期望這將為人類帶來更加美好的未來，為文化創意領域帶來前所未有的發展。

貳、動機與目的

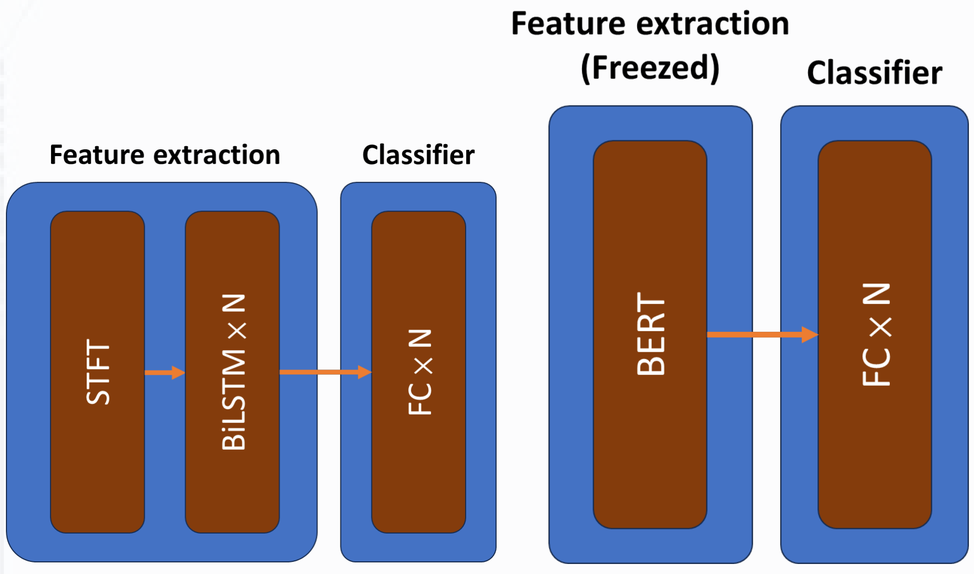
本研究的動機源於音樂和語言在塑造文化方面的固有重要性，它們作為藝術表達和情感的工具。然而，將音樂的情感精髓有效地轉化為文本，或者反之亦然，是一項相當大的挑戰。除了學術目標外，這項研究還具有豐富聽力障礙者生活的潛力，因為它使他們能夠接觸音樂的情感深度，這是失聰後無法體驗到的。此外，它還有能力改變人們與這兩種媒介互動和理解的方式，促進更深入的理解和互動。

研究目的：

- 將音樂的曲調轉換成**情緒模式**判斷
- 嘗試各種方法將音樂情緒轉變為故事
- 建立公眾投票判斷音樂產生出的故事效果並討論結果
- 從音樂到故事的模型訓練故事到音樂的模型

參、研究過程和方法

首先，我們試著找出計算音樂與故事差距的方法，可作為訓練模型的損失函數。而因我們的目標是將音樂轉成情緒上相似的故事，於是我們就從情緒下手，找到**Valence-Arousal**與**Valence-Arousal-Dominance**兩種情緒表示法，並且找到使用前者的**音樂-情緒資料集DEAM**，與使用後者的**句子-情緒資料集EmoBank**。



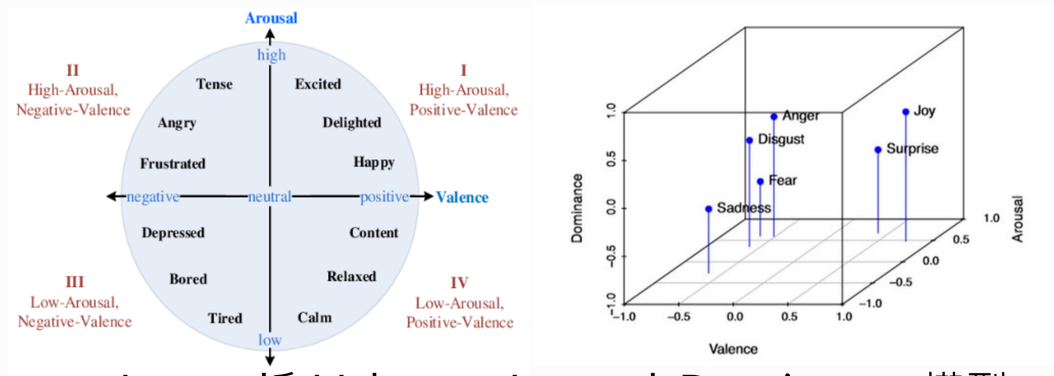


圖1(a) 音樂情緒辨識模型　圖1(b) 文句情緒辨識模型

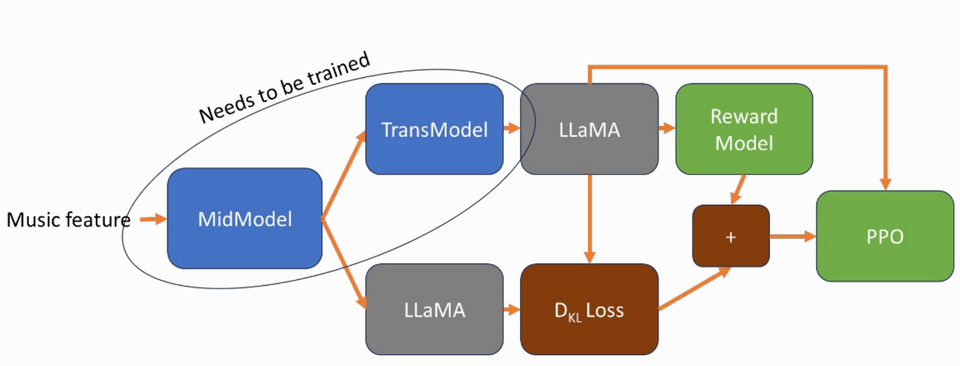
圖2 Valence-Arousal和Valence-Arousal-Dominance模型
(見 Russell, J. A., & Mehrabian, A. ,1977, and Russell, J. A. ,1980)

- 訓練了一個**Bi-LSTM**以識別音樂序列的Valence-Arousal值
- 接受**可變長度的輸入序列**，不如Transformer需padding
- Transformer模型的encoder→給decoder用Attention決定要關注的內容
- 更難訓練Transformer實現**一對一序列**特徵提取。

- 凍結**BERT** BASE，並加一些FC層來創建模型，識別文本（句子）中的情感
- 用EmoBank資料集訓練
- 使用Valence-Arousal-Dominance來表示情感，只取其中的Valence-Arousal值
- 這兩種理論都假設各因子在**情感多維空間中相互垂直**

- 將聲音和文本轉換為情感指標後，可訓練一個能將**音樂轉換為故事**的模型
- 我們想利用**轉移學習**，特別是Meta的**LLaMA**系列開源預訓練模型
- 結合聲音情感識別的Bi-LSTM的輸出（w/o head）與LLaMA

- 測**生成和期望輸出之間的差異**➔結果文本分割成單個句子，提取Valence-Arousal值➔合成情感序列，與Bi-LSTM模型內輸入音樂的情感表示對齊。
- 利用MSE等損失函數改進模型理解和將音樂轉換為敘事結構的能力。
- **BUT**，將文本序列的句子分開**不可微分**➔找更好的方法：**進化演算法** or **人類反饋強化學習**(RLHF w/ PPO)
- ChatGPT和LLaMA都用RLHF微調，達到更好的效果。



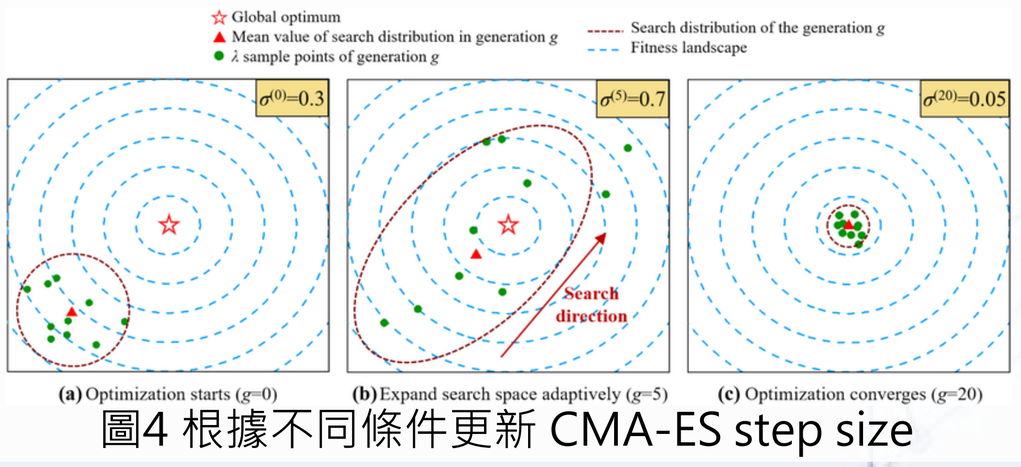


圖3 我們的 RLHF 模型架構和工作流程　圖4 根據不同條件更新 CMA-ES step size

- 發現**參考模型也被更新**➔使用**胡言檢測模型**替代參考模型的功能
- 沒有參考模型➔不必使用RLHF，可以使用純RL算法(PPO)。
- **BUT**，這種方法遇到困難➔進化演算法替代，CMA-ES & DE。
- **BUT**，這個演算法**無法有效率**的訓練模型(e.g.,記憶體分配和每一步消耗的時間)
- 有一方法可實現目標而無需訓練模型：將各種情感**詞彙映射**到它們在情緒二維平面上的坐標，在平面上找到最接近的情感字彙來描述輸出的Valence-Arousal值，我們可以要求LLaMA根據情感序列生成故事。

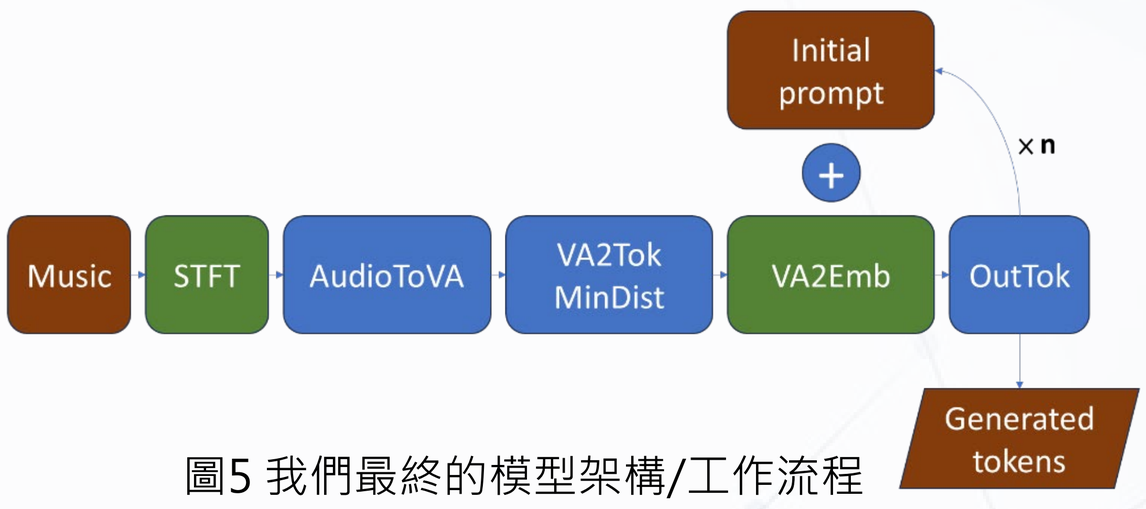


圖5 我們最終的模型架構/工作流程

現在我們有一個可用的音樂到故事模型，我們可訓練一個**反向模型**，使其能夠從故事生成音樂。我們提出了兩種方法。

我們可使用生成對抗網路（GAN）有效地訓練一個反向函數，首先需要對數據進行處理，這包括**從音樂生成多個故事**，然後選擇那些配對效果良好的故事。接下來，只需使用一個GAN架構來訓練一個**生成式AI**。或者，如果GAN的表現不夠好，我們也可以使用擴散模型。

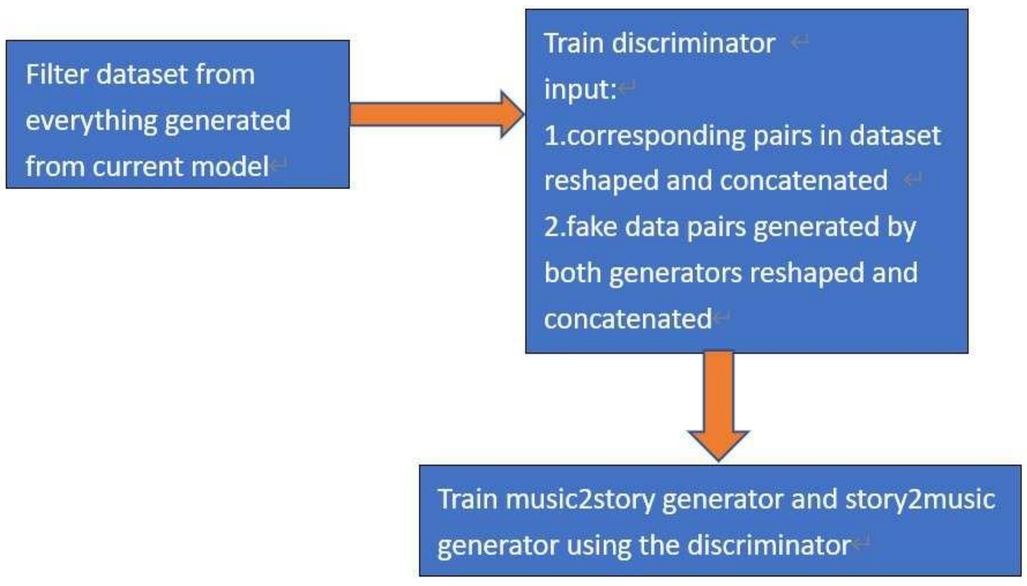


圖6 使用GAN從音樂到故事的模型訓練故事到音樂的模型

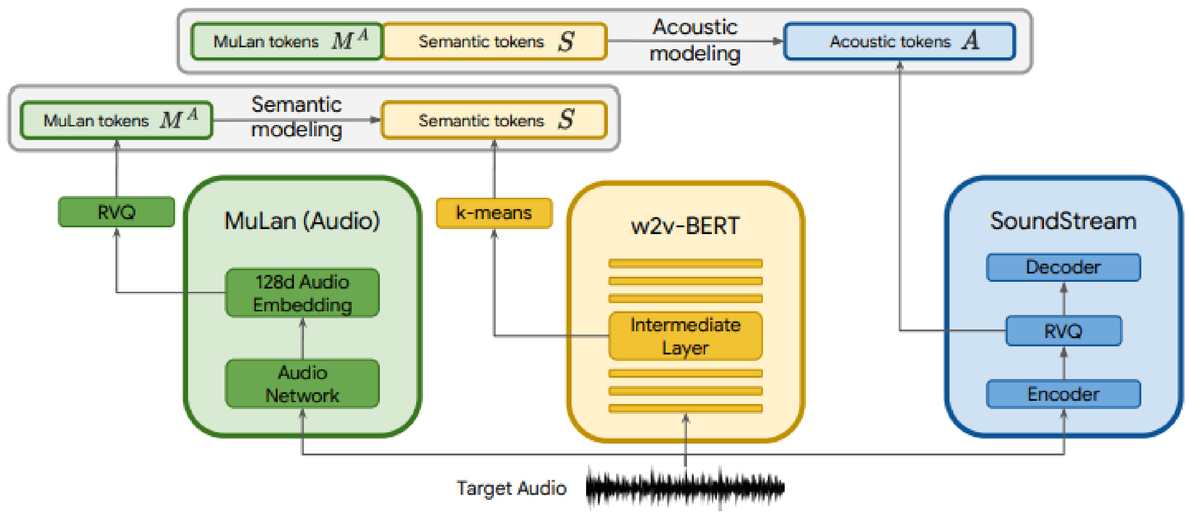


圖7 MusicLM 架構（來自 Andrea 等人，2023年）

- 另一個方法是**遷移學習text-to-music**模型。
- **MusicLM**可從文字描述生成高保真度的音樂
- MusicLM採用了先進的分層序列到序列建模技術，將描述性文本無縫地整合到連貫的音樂作品中。
- Google並**沒有開源模型**，只有社群開源，效果比Google演示的差了一截。
- 改用Meta發表的MusicGen，SoundStream → Encodec，並用一個Transformer免去了並聯多個模型的必要

不論是MusicGen或是MusicLM，他們的輸入都是「**音樂的描述**」，無法提取出故事的情緒並將其製成音樂，因此我們可以有兩種方案:
一、用**參考比較**讓原本的Description encoder變成Story encoder
二、在輸入模型**前面加上**一個text-to-text model（例如T5）
法一會有一個問題：MusicGen的text encoder
並不會把文字壓為一維向量，而是每個字形成的時間序列，
所以我們會用法二。

我們試了兩個資料集製作方法（共花**30小時**人工審核）:
一、使用LLaMA生成故事，寫出適合故事的音樂敘述
二、使用Reddit Short Story資料集的故事，
寫出適合的音樂敘述

最後，我們發現使用**第二種方法**生成之資料集微調T5模型效果較好，生成出的音樂描述與故事情感較匹配。

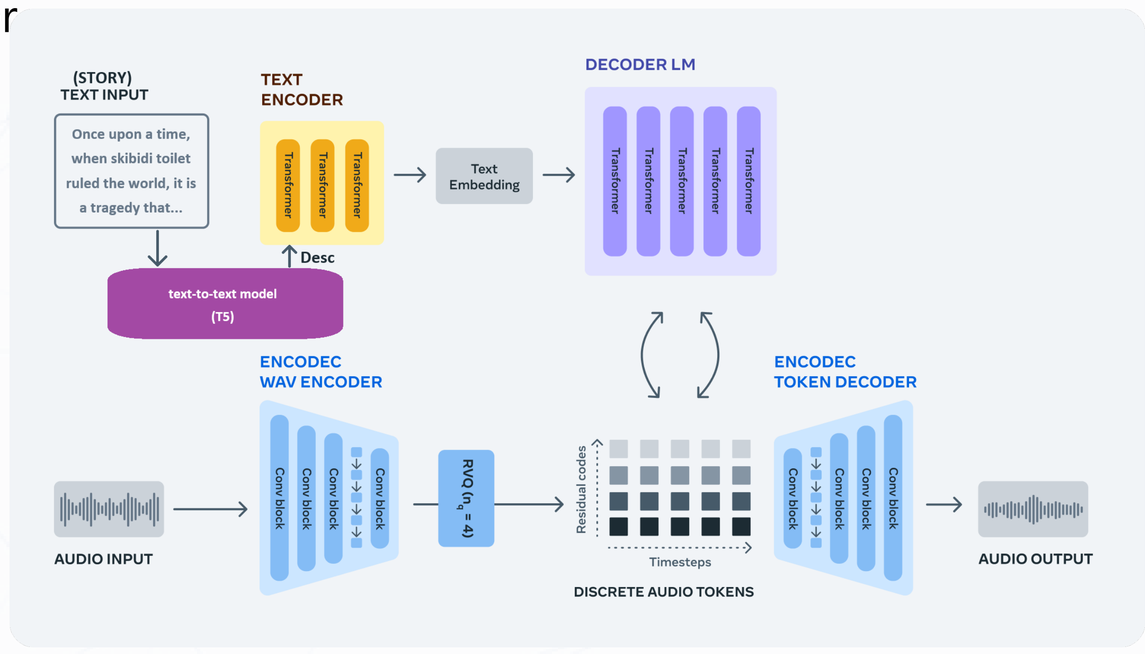


圖8 MusicGen+T5 pre-transcriber

肆、研究結果

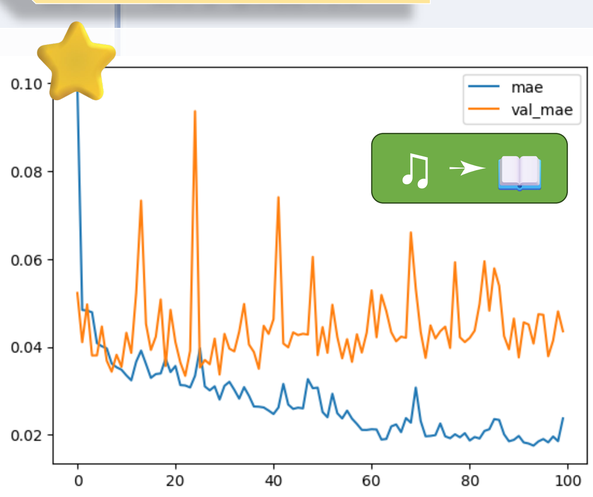


圖9(a) 音樂情緒辨識模型MSE

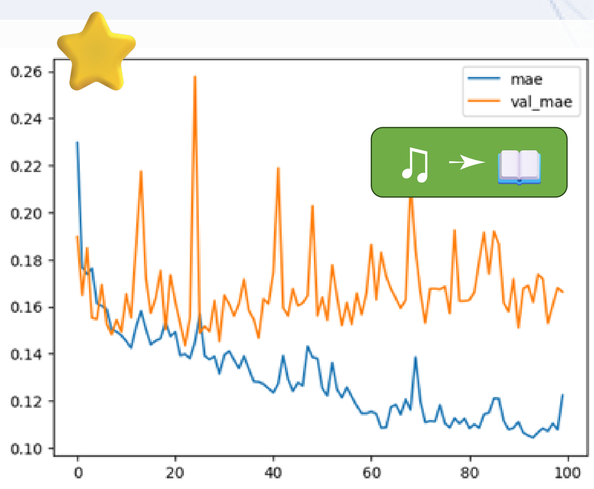


圖9(b) 音樂情緒辨識模型MAE

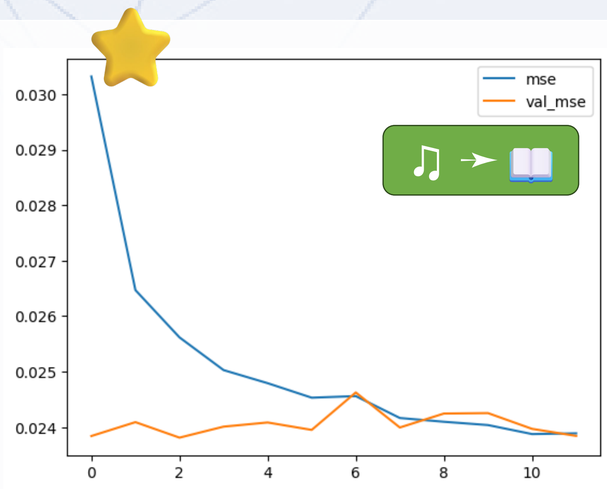


圖10(a) 語句情緒辨識模型MSE

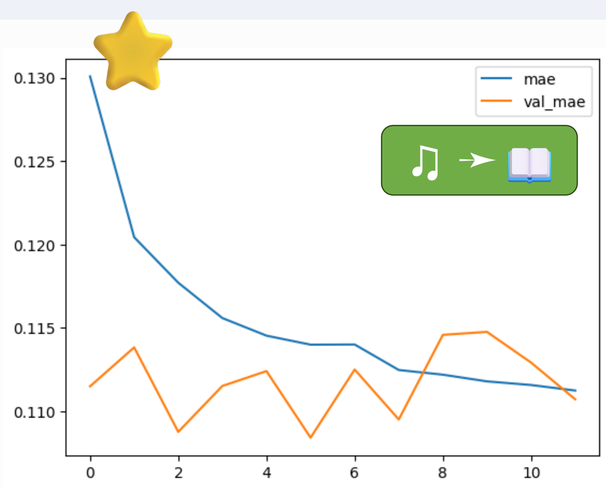


圖10(b)語句情緒辨識模型MAE

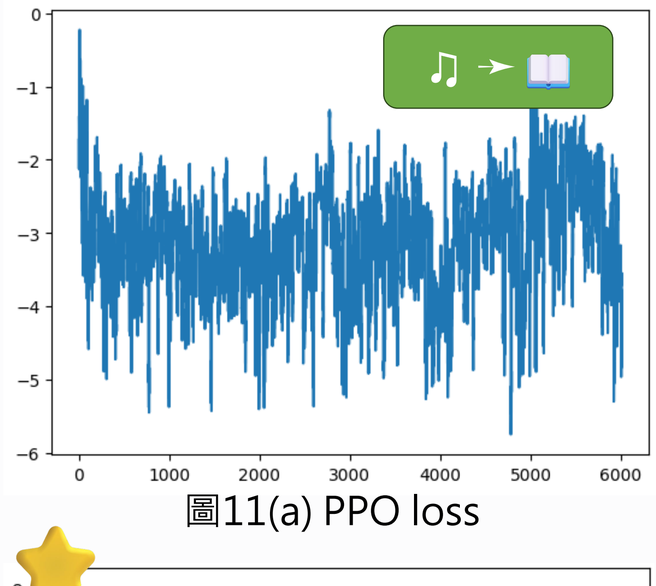


圖11(a) PPO loss

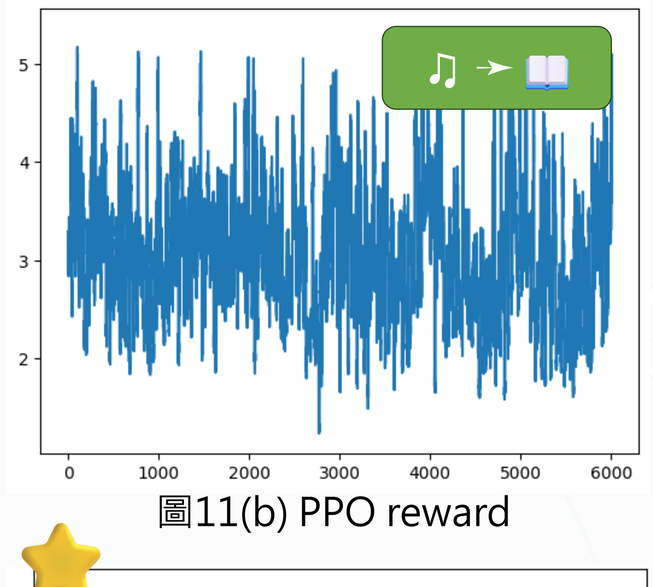


圖11(b) PPO reward

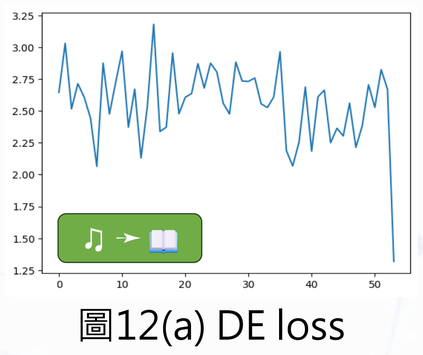


圖12(a) DE loss

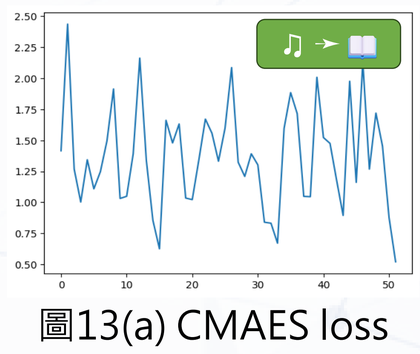


圖13(a) CMAES loss

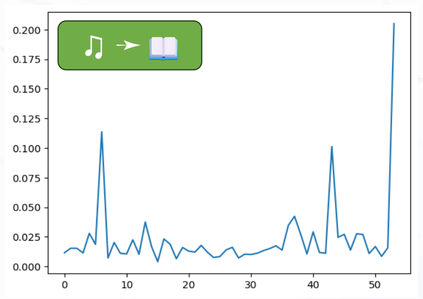


圖12(b) DE gib_score

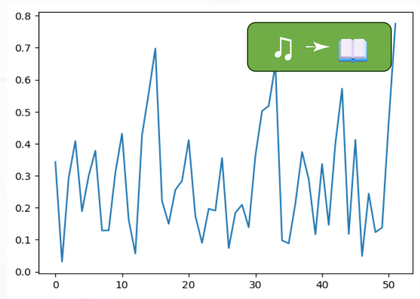


圖13(b) CMAES gib_score

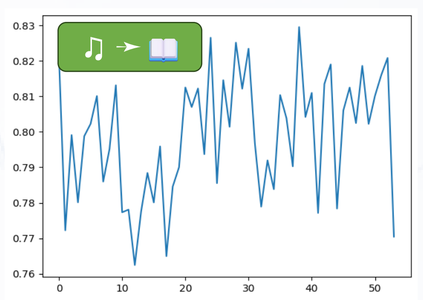


圖12(c) DE distance

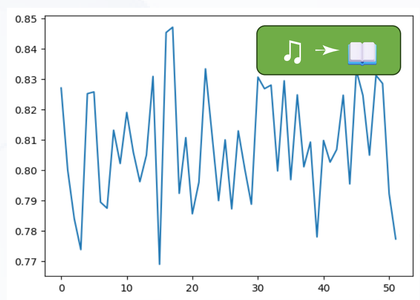


圖13(c) CMAES distance

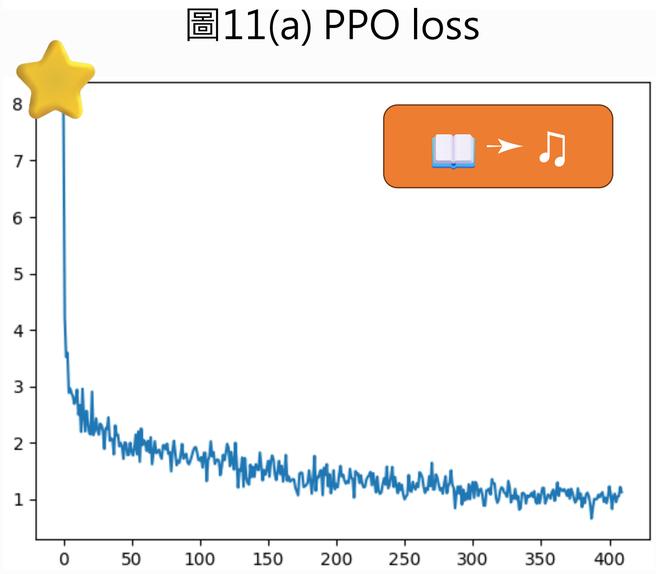


圖14(a) 法一T5模型微調loss曲線

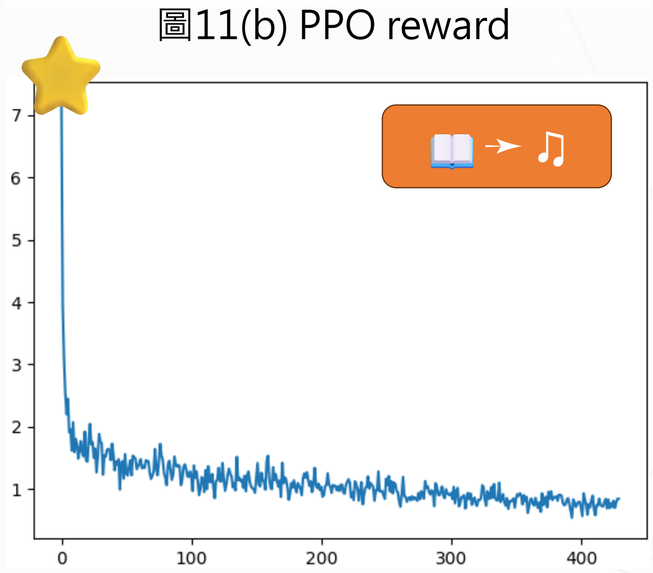


圖14(b) 法二T5模型微調loss曲線

表1 T5 兩模型生成物比較

	Model 1	Model 2
Generated	a haunting melody weaves through a misty landscape, evoking a sense of detachment and self-discovery as the storm passes.	electronic score with a sense of foreboding and unease, gradually building towards a climax of tension and triumph.
Truth	Sun-kissed indie rock strums and soaring vocal harmonies conjure a carefree spirit, basking in the glow of uncharted horizons.	Ethereal, atmospheric piano melody with subtle electronic undertones, building towards a climax with hints of Eastern mysticism and philosophical introspection, evoking a sense of mental sparring and intellectual

- 音樂情感分類模型最低MAE約為0.14～0.15
- 音樂情感分類模型似乎overfitting，但我們嘗試的**正規化**方法（L2 & Higher dropout）都沒有降低validation MAE值（但training MAE → validation MAE）→是模型能夠達到的最佳效果
- 文句情緒辨識的表現良好，MAE為0.11。**準確度約為85%和89%**（值域為-1～1）
- **RL模型訓練的結果不理想**，它優化損失函數，但無法優化reward（loss過度擬合）
- 可能是我們用來訓練LLaMA模型的方法與實際的RLHF算法稍有不同：無“固定”的參考模型→加上gibberish detector調節actor model
- 模型**仍然表現得很糟糕**，它只會輸出**無意義訊息**
- 無梯度優化試了**CMA-ES和DE**，不過他們效果都**不是很好**。
- 微調T5過程中，資料集法一：故事第一句都**非常相似**，模型訓練結果顯示**無法判別**出故事的情緒
- 資料集法二：資料集的**故事比較多元**，模型訓練結果顯示對故事的**情緒理解力較強**，可看出正確答案和生成答案有一定的**相似性**。

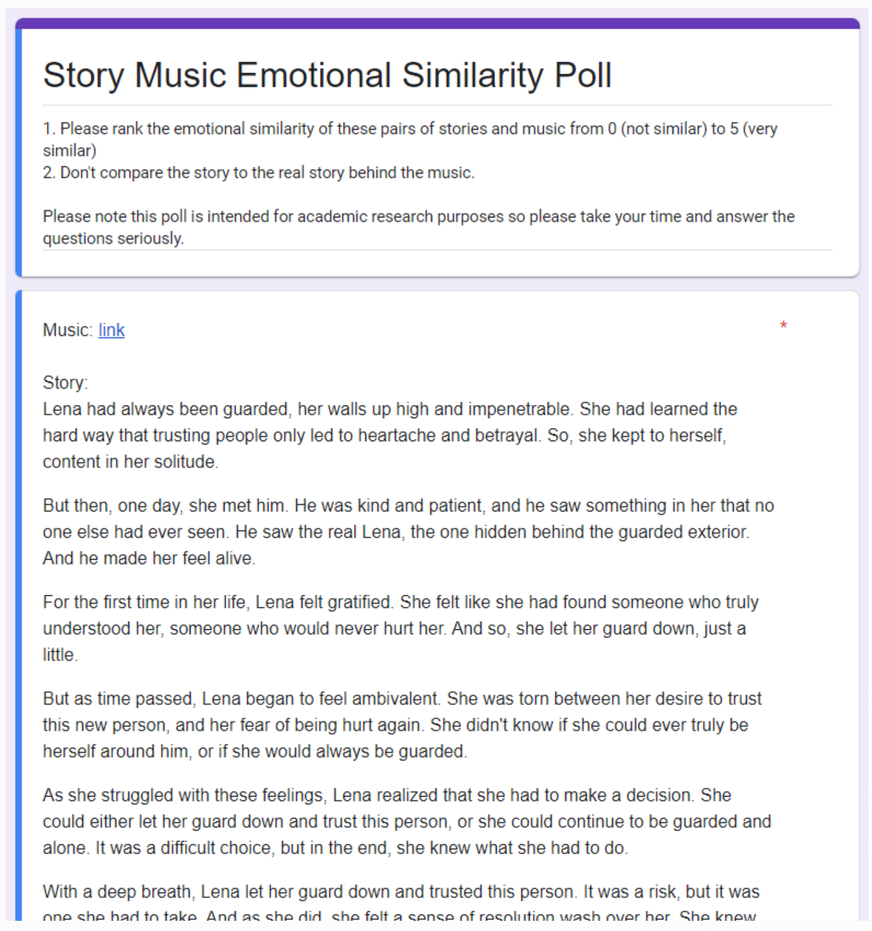


圖15 模型表現得好不好的問卷截圖

表2 每個評分和音樂-故事對的投票統計表

score	0	1	2	3	4	5
Story 1	2	2	1	7	7	8
Story 2	0	2	3	5	9	8
Story 3	0	1	2	4	8	12
Story 4	0	3	4	2	3	15

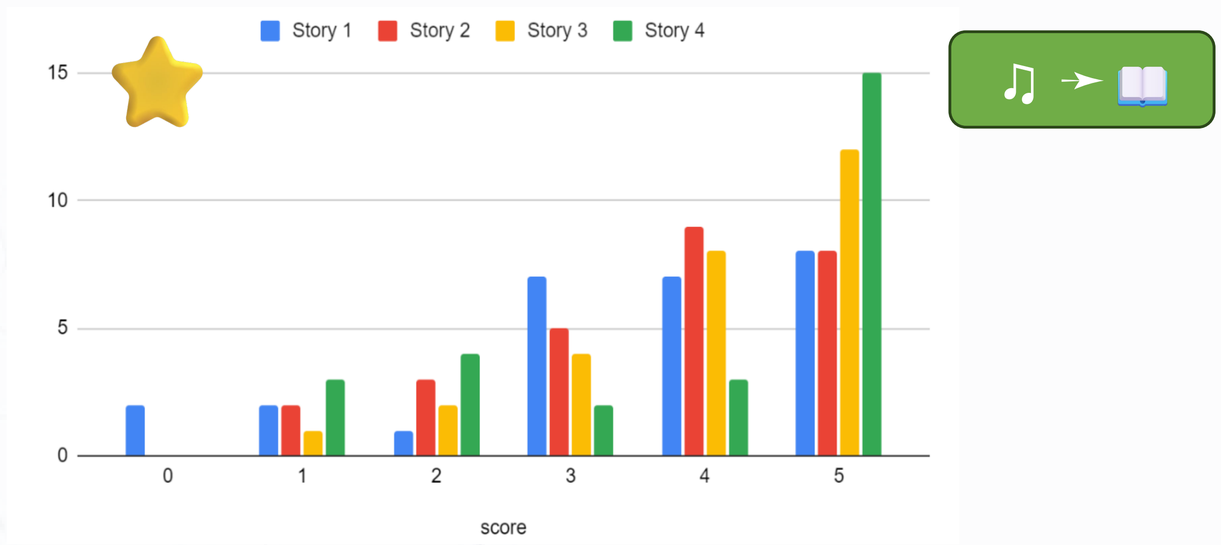


圖16 每個故事和評分人數的條形圖

我們給予**隨機的參與者**音樂的連結，以及由模型生成的相應音樂所生成的故事。此調查最終共有**27位**參與者，以下是最終的統計表。條形圖在較高的分數上傾向於更高，因此我們可以知道它在**主觀指標**方面表現**良好**。

5 討論與結論

表3 按故事評分

	story1	story2	story3	story4	total
avg	3.444444444	3.666666667	4.037037037	3.851851852	3.75
stderr	1.527525232	1.240347346	1.125969035	1.511588663	

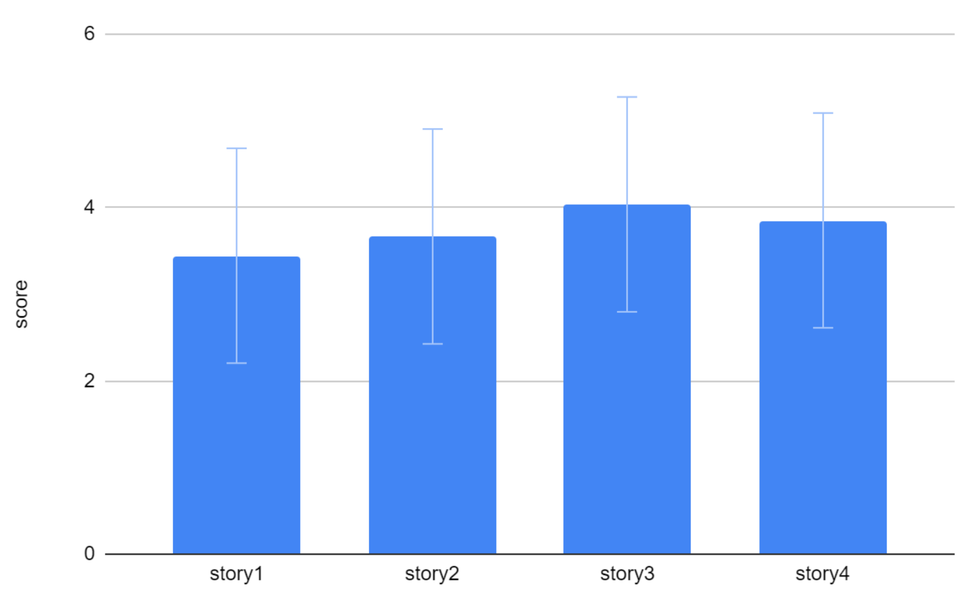


圖17 每個故事-音樂配對的評分

故事1的評分是最差的，第二差的是故事2。故事1之所以糟糕可能僅僅是因為模型的**內部干擾**，我們並不確切知道為什麼它不能生成好的故事。至於故事2，它是從《波西米亞狂想曲》中生成的，原因可能是模型不真正知道如何表達歌曲背後的感受，因為它**太複雜**了。

故事3超越了其他故事，可能是因為它所傳達的情感**直接而容易辨認**——快樂。至於故事4，它是從威爾第的《安魂曲》中生成的，它也獲得了可觀的表現。原因在於它深沉而**易於識別**的情感，一種**極度悲傷**的情感。

我們已經成功開發了一個可以從音樂生成故事的模型，而且故事的**品質相當不錯**。由於社群給出了3.75分（滿分5分）的評分，我們可以知道這個模型在**主觀指標上也表現良好**。我們還提出了兩種方法來創建一個**反向模型**，這個模型可以從故事生成音樂，最終**訓練結果表現優良**。這標誌著藝術與人工智慧技術融合的一個成就，它可能為人類提供一個更美好的未來。

6 參考文獻

[1] Aljanaki et al. (2017, March). Developing a benchmark for emotional analysis of music. PLOS ONE, 12, e0173392. doi:10.1371/journal.pone.0173392

[2] Buechel et al.(2017, April). Emobank: studying the impact of annotation perspective and representation format on dimensional emotion analysis. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (pp. 578–585). Valencia: Association for Computational Linguistics. Retrieved October 21, 2023, from https://aclanthology.org/E17-2092

[3] Copet et al.(2023). Simple and Controllable Music Generation. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems (Vol. 36, pp. 47704–47720). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/94b472a1842cd7c56dcb125fb2765fbd

[4] Meta AI (2024, April 18). Introducing Meta Llama 3: The most capable openly available LLM to date. AI at Meta. https://ai.meta.com/blog/meta-llama-3/

[5] Russell et al.(1980, December). A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1161–1178. doi:10.1037/h0077714

[6] Russell et al.(1977). Evidence for a three-factor theory of emotions. Journal of Research in Personality, 11, 273-294. doi:https://doi.org/10.1016/0092-6566(77)90037-X

[7] Touvron et al.(2023, July). Llama 2: open foundation and fine-tuned chat models. Llama 2: open foundation and fine-tuned chat models. arXiv. doi:10.48550/arXiv.2307.09288