

國立臺灣大學管理學院資訊管理學系



碩士論文

Department of Information Management

College of Management

National Taiwan University

Master Thesis

以憂鬱症病患為對象之多標籤對話生成

Multi-label Dialogue Generation for Depression Patients

蘇成恩

Cheng-En Su

指導教授：莊裕澤 博士

Advisor: Yuh-Jzer Joung, Ph.D.

中華民國 111 年 8 月

August 2022

國立臺灣大學（碩）博士學位論文
口試委員會審定書



中文題目：以憂鬱症患者為對象之多標籤對話生成

英文題目：Multi-label Dialogue Generation for Depression

Patients

本論文係 蘇成恩 君（學號 R09725065）在國立臺灣大學資訊管理學系、所完成之碩（博）士學位論文，於民國 111 年 8 月 22 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

蘇成恩 (簽名)
陳連新 盧紹鈞

系主任、所長

陳連新

(簽名)

致謝



首先，必須先感謝指導教授莊老師在這兩年中的指導，從莊老師這邊我也學習到許多做研究該有的觀念，莊老師注重的思維邏輯對我的影響也非常深刻，幫助我了解到思維邏輯的重要性，從碩一開始，每週的報告莊老師都不斷提點我們，不能只是為了報告而讀論文，必須從自身的角度去理解整篇論文，了解作者進行研究時的思維邏輯，也和自己的想法進行比對，試著去找出論文或自身想法的盲點，這樣才能幫助自己進步，讓我在閱讀論文時能夠收穫更多東西。而在進行研究時，並非只是將實驗結果報告出來，需要針對結果進行直觀的解釋，要去思考整體的研究方法與結果是否合乎邏輯，也要找結果能夠證明或是有什麼意義，才能夠確立研究結果的價值。總而言之，非常感謝莊老師這兩年的諄諄教誨，讓我體會到做研究的辛苦與樂趣，也讓我學到許多在業界或是課堂上所學不到的知識或觀念，讓我受益良多。

同時，也非常感謝實驗室的同學們，感謝郁云與我共同合作，每次都不斷討論論文的方向，也一起解決老師提出來的問題，也感謝人傑不斷為整個實驗室的付出，讓我們能擁有良好的環境做研究，最後也感謝翔岳與安浩，每次都會一起討論報告或是論文相關的問題。感謝大家這兩年的陪伴與互相督促，讓我們都變得越來越好，也祝福大家畢業後都能有好工作，能夠賺大錢。

最後，要特別感謝我的親人們一路上的支持與幫助，讓我能夠考上研究所，也付出許多讓我能夠安心讀完碩士，在過程中也不斷關心我，也感謝學長姐、朋友們和我的伴侶這兩年來的協助與建議，讓我這一路上走來非常開心也順利，祝福所有的人都能有美好的未來。

蘇成恩謹誌
國立臺灣大學資訊管理學系
民國一百一十一年八月

中文摘要



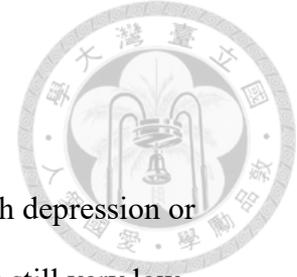
在現今社會中，儘管憂鬱症病患或是有憂鬱傾向的人越來越多，但憂鬱症病患的求助率卻十分低迷，主要是由於憂鬱症病患缺乏一個良好且即時的聊天管道，加上許多應用在醫療領域的對話系統都著重在解決特定任務上，無法解決憂鬱症病患缺乏聊天管道的難題，而此時能夠理解使用者情緒，並且幫助使用者紓解負面情緒或是給予正面積極的回覆的對話系統就顯得極為重要。

因此本論文與本研究室另一位成員 Huang(2022)共同設計一個對話系統，能夠分辨使用者輸入文字的態度、意圖與主題，並且透過這些標籤來進行對話生成，希望藉此來控制模型產生出正面積極且相關的回覆給使用者，來與使用者聊天或是給予建議。

本論文使用強大的 GPT-2 預訓練模型(pre-trained model)並結合多種標籤來訓練模型，分別對英文對話資料集 DailyDialog 與由本論文與 Huang(2022)所共同建構的中文對答資料集 ChinesePsyQA 進行實驗，將多種維度標籤加入到對話文本前，藉此讓模型產生出相對應的回覆，而最後在人工評估與自動評估兩種評估方式上，本論文所提出的方法皆比其他生成模型有著更好的表現，說明將多種標籤加入到對話中，可以有效控制模型的生成文本。

關鍵字：憂鬱症、對話生成、預訓練模型、多標籤、深度學習

英文摘要



In today's society, although there are more and more people with depression or depression tendencies, the help-seeking rate of depression patients is still very low, mainly due to the lack of a good and real-time chat channel for depression patients. In addition, many dialogue systems applied in the medical field focus on solving specific tasks, and cannot solve the problem of lack of chat channels for patients with depression. At this time, a dialogue system that can understand the user's emotions and help users relieve negative emotions or give positive responses is extremely important.

Therefore, this paper and another member of our laboratory, Huang (2022), jointly designed a dialogue system that can distinguish the attitude, intentions and topics of the user's input text, and use these tags to generate dialogues, hoping to control the model to produce positive and relevant replies to users, to chat with users or give advice.

This paper uses the powerful GPT-2 pre-trained model and combines various labels to train the model. Experiments were carried out on the English dialogue dataset DailyDialog and the Chinese QA dataset ChinesePsyQA jointly constructed by this paper and Huang (2022). Various dimension labels were added before the dialogue text, so that the model can generate corresponding responses. Finally, in the human evaluation and automatic evaluation, the method proposed in this paper has better performance than other generative models, indicating that adding multiple tags to the dialogue can effectively control the generated text of the model.

Keywords: Dialogue Generation, Pre-trained Model, Multi-label, Deep Learning, Depression

目 錄



	Page
口試委員審定書	i
致謝	ii
中文摘要	iii
英文摘要	iv
圖目錄	vii
表目錄	viii
第一章 緒論	1
1.1 研究背景與動機	1
1.2 研究目的	3
1.3 論文架構	4
第二章 文獻探討	5
2.1 憂鬱症治療	5
2.2 對話系統	6
2.2.1 自然語言理解	7
2.2.2 對話生成	8
2.3 基於 GPT 模型之對話系統	9
2.4 對話系統評估方法	11
2.4.1 自動評估指標	11
2.4.2 人工評估指標	12
2.4.3 現有評估指標之問題	13
2.5 總結	15
第三章 研究方法	16
3.1 研究架構	16



3.2	資料集.....	17
3.2.1	DailyDialog.....	17
3.2.2	ChinesePsyQA.....	19
3.3	模型架構.....	21
3.3.1	模型輸入格式.....	21
3.3.2	GPT-2 DoubleHeads Model.....	22
3.4	研究驗證方法.....	25
3.4.1	自動評估.....	25
3.4.2	人工評估.....	25
第四章	研究結果.....	27
4.1	Baseline Models.....	27
4.2	訓練參數.....	28
4.3	自動評估結果.....	28
4.4	人工評估結果.....	30
4.5	回覆內容分析.....	31
4.6	對話系統串連.....	35
4.7	小結.....	37
第五章	結論.....	39
5.1	研究成果.....	39
5.2	研究貢獻.....	40
5.3	研究限制.....	40
5.4	未來研究方向.....	41
參考文獻	43
附錄	47

圖目錄



圖 1 全台使用抗憂鬱症藥物人數歷年統計圖	1
圖 2 聊天系統流程圖	16
圖 3 DailyDialog 意圖維度分佈	18
圖 4 DailyDialog 情緒維度分佈	18
圖 5 DailyDialog 主題維度分佈	19
圖 6 ChinesePsyQA 意圖維度分佈	20
圖 7 ChinesePsyQA 態度維度分佈	20
圖 8 ChinesePsyQA 主題維度分佈	21
圖 9 模型輸入格式圖	22
圖 10 模型架構圖	23

表目錄



表 1 人工評估指標種類	14
表 2 Daily Dialog 對話範例	18
表 3 ChinesePsyQA 對答範例.....	20
表 4 自動評估結果表	29
表 5 人工評估詢問方式	30
表 6 人工評估結果表	30
表 7 不同態度標籤之生成範例	32
表 8 不同意圖標籤之生成範例	33
表 9 有無主題標籤之生成範例	34
表 10 單輪對話系統串連範例	35
表 11 多輪對話評估詢問方式	36
表 12 多輪對話系統串連範例	36
表 13 多輪對話評估結果	37
表 14 模型不良生成範例	41



第一章、緒論

1.1 研究背景與動機

在現今社會中，憂鬱症是極度需要重視的疾病，因為憂鬱症會使得人無法工作、生產力下降，進而對家庭與社會造成嚴重的經濟負擔。根據世界衛生組織所統計的全球高負擔疾病中，憂鬱排名第三名，並預測在 2030 年將攀升至第一名。根據(宋偲嘉, 2021)憂鬱症在臺灣一年所造成的社會經濟損失，已逾 350 億元新台幣。同時，也因為社會快速進步，民眾在面對快速變遷的社會時難以調適，使得臺灣的憂鬱症病患人數逐年上升，如：全台使用抗憂鬱症藥物人數歷年統計圖（圖 1.）所示。

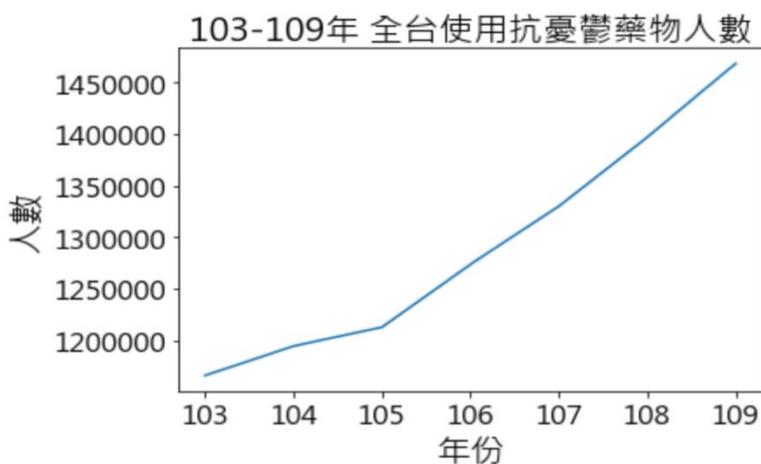


圖 1. 全台使用抗憂鬱症藥物人數歷年統計圖

(資料來源：衛生福利部中央健康保險署)

儘管憂鬱症病患人數不斷上升，國內的憂鬱症就診率卻依舊低迷。根據臺灣憂鬱症防治協會調查(張家銘, 2020)，歐美的憂鬱症患者 57.4%有求助行為，而臺灣則僅有 20%，顯示在臺灣相較歐美而言，憂鬱症病患的求助率是較少的。而未接受治療的原因主要為不承認自身患有憂鬱症、抗拒就醫、不認為就醫對症狀有效用與就醫管道匱乏。此外，許多憂鬱症病患不願與他人講述自身患有憂鬱症或出



現憂鬱症的症狀，也不願讓自身情緒狀況影響到家人或朋友，醫療幫助也並非隨時可以取得，使得病症無法即時得到良好的抒發管道，造成病症不斷加重與身心功能逐漸下降，甚至導致患者有自殺傾向。由上述可知，臺灣憂鬱症病患主要缺乏一個良好且即時的聊天與抒發管道來幫助他們，而自然語言處理領域中的對話系統便是一個良好的解決辦法。

自然語言處理 (Natural Language Processing) 是人工智慧中極為熱門的領域之一，人類的語言是具有極高複雜性的，透過字詞、符號來傳達資訊與其意涵，人們從小到大不斷使用語言藉此學習，而語言事實上是極度龐大的非結構性資料，而這樣的資料對於電腦來說是難以學習的，但隨著人工智慧的發展，自然語言處理也隨之盛興。在我們的日常生活中其實有許多任務都藉由自然語言處理來達成，語言翻譯、搜索引擎及文字分析，這些都是已經應用在我們日常生活中的例子。同時，能夠與人類進行對話互動的對話系統也因此成為一個熱門議題，對話系統可以裝置在手機或筆記型電腦等攜帶式裝置，且對話系統不受時間上的限制，讓使用者可以隨時隨地使用，這樣的特性是有利於解決憂鬱症病患缺乏對話管道的問題，且與系統對話同時也解決憂鬱症病患不願與真人分享的問題。

對話系統現在已廣泛運用在各個領域，在醫療照護領域也有許多應用。但多數的醫療應用都是著重在客服問答、預約掛號或是衛教知識傳遞，這樣的對話系統在面對憂鬱症病患的效用較低，因為此類對話系統缺乏情緒與個性，也缺乏理解使用者情緒的元件，但理解情緒並帶有情緒的回覆在面對憂鬱症病患是重要的，如心理治療中的一般支持性心理治療 (Supportive psychotherapy) 提到，以積極樂觀的態度與支持鼓勵的意圖來回應憂鬱症病患使是對病患有助的，因此讓對話系統帶有正向態度與意圖是設計以憂鬱症病患為對象的對話系統的關鍵之一。

此外，生成符合使用者期待的語句並保持語句順暢與相關性可以增進使用者的使用意願與增加談話長度。根據(Li et al., 2016)、(Zhang et al., 2018)的研究指出，



以往的對話生成主要的問題在於缺乏回答的一致性與無法展現生成的獨特性，因為電腦藉由大量的對話範本訓練，而這些範本都是由不同人所回答的，使得訓練後會混雜所有回答者的個性，而無法展現特定的個性，因此加入個人特徵來控制機器人。同時，在(Lin et al., 2019)、(Zandie & Mahoor, 2020)的研究中也指出，在對話生成中加入情緒維度可以讓回答更有同理心，可以使使用者更有參與感與持續使用的動力，但對於將多個維度加入對話，目前較缺乏相關研究。

除此之外，預訓練語言模型（Pre-trained Language Model）的出現讓聊天機器人在對話生成有著卓越的進步，預訓練語言模型是指先藉由大量文本來訓練一個模型，使此模型學習到通用的語言表示，之後只需要小樣本的資料便能讓模型適用於特定領域的應用。其中，GPT-2（Generative Pre-trained Transformer 2）在文本生成有著卓越的效能，GPT-2 由 OpenAI 團隊(Radford et al., 2019)所提出的語言模型，GPT-2 在訓練時是以預測下個字為目標，使得 GPT-2 善於藉由上文來生成下文。(Lin et al., 2019)、(Zandie & Mahoor, 2020)，也嘗試使用 GPT-2 來進行聊天對話的生成，也都取得超越以往研究的成果，然而上述研究皆是直接將文本輸入模型來產生對話，對於先將態度、意圖與使用者資訊等標籤判斷出來後再加入模型的相關研究也較為缺乏。同時，目前的研究也都以英文為主，在中文對話生成上的研究較少。

因此本篇論文將以憂鬱症患者為對象並使用預訓練模型建立一個對話系統，並在對話中加入態度、使用者意圖與主題等標籤來建立對話生成的一致性與獨特性，提供憂鬱症患者一個即時的抒發與聊天管道，作為初步的引導工具，幫助病患減輕壓力與提供他們正向的回覆。

1.2 研究目的

本研究的主要目的旨在解決目前臺灣憂鬱症病患缺乏即時的聊天管道來減緩症



狀或紓解負面情緒的問題。因此，本研究設計一個對話系統架構，主要包含對話判斷與對話生成兩個部分，與本研究室另一位成員 Huang(Huang, 2022)共同合作並由 Huang 來研究對話分析的部分，而本論文則專注於對話生成部分。首先，系統會先主動關心使用者情況，並與使用者進行交談，在交談過程中確認使用者是否需要幫助，若需要幫助則使用 BERT (Bidirectional Encoder Representations from Transformers) 將文本分類並獲得文本在態度、意圖與主題三者的維度標籤，並將標籤與文本送入 GPT-2 (Generative Pre-trained Transformer 2)，讓生成的對話可以帶入標籤，讓生成的對話可以帶有正向的態度並保持一致性與幫助性，希望在交談中帶給病患正面積極的態度，幫助病患減緩症狀或負面情緒，作為憂鬱症病患的初步引導者。

1.3 論文架構

本論文將於第二章探討過去對話系統的相關文獻，包含過往對話系統中可學習之處與可改善之處、如何評估對話系統以及使用預訓練模型來產生對話，在第三章則會描述整體實驗架構、資料集與實驗模型，在第四章說明實驗結果及與其他模型之比較，於第五章說明結論及未來研究方向。

第二章、文獻探討



本論文旨在將預訓練語言模型應用在對話系統以建立出以憂鬱症病患為對象之對話判斷與生成，透過帶入態度、意圖與主題等維度來提升對話品質。在本章節根據探討與分析過往研究，進而找出可以突破與嘗試的研究方法。以下將根據本論文所需探討的技術分為四個段落。

於本章第一節本論文將會探討目前主要的憂鬱症治療方法與技術，基於這些研究技術延伸出本研究的對話架構與帶入維度的原因。第二節，將會探討的本研究的核心架構：對話系統，並著重探討對話系統中的對話生成，並探討帶入各種標籤的對話生成的相關研究，基於這些研究結果延伸出本論文之實驗架構。第三節將探討使用現今熱門的預訓練模型(Pre-trained Model)來進行對話生成的相關研究，將成為本論文所使用的實驗模型並加以延伸。第四節將會探討現有的對話生成評估指標，探討各種不同的分數指標與其不足之處，藉此找出能作為本實驗成果的評估指標，用來驗證此研究的結果。最後，於第五節總結本論文所探討的文獻，整理出結論與可突破及嘗試的方法。

2.1 憂鬱症治療

目前對於憂鬱症有許多的治療方法，根據文獻(李昭慶, 2000)指出，憂鬱症目前的治療方法可以分為兩大類：藥物性治療與非藥物性治療，藥物性治療主要是以服用抗憂鬱症藥物為主，可以減緩病患因憂鬱症而出現的身體症狀，非藥物性治療有心理治療 (Psychotherapy) 以及電擊治療 (Electroconvulsive therapy)。而憂鬱症治療通常會以並行的方式來執行，在 Thase et al. (1997)的研究中指出，心理治療可以協助抗憂鬱症藥物達到更好的效果，而單獨進行藥物或非藥物治療的效果都沒有兩者同時實施高，但比起無治療，心理治療依舊可以顯著改善病況，且根據研究(Carr, 2008)，患有憂鬱的青少年在接受心理治療後，有 63%的病患其憂

鬱症病況減輕是優於沒有經過治療的，可見心理治療對於憂鬱症病況改善依舊有顯著的影響力。



心理治療主要針對病患的心理以及思考進行治療，主流的心理治療有一般支持性心理治療 (Supportive psychotherapy)、認知行為心理治療 (Cognitive-behavioral therapy)、精神動力心理治療 (Psychodynamic Therapy) 以及團體心理治療 (Group psychotherapy)。由於精神動力治療需要對病患進行長時間的分析來讓病患了解自身的潛意識與思想，團體心理治療則需要多名病患共同參與，互相協助與支持來達到治療，因此支持性心理治療是最簡易達成的。

支持性心理治療是目前最常使用的心理治療方式，根據醫學百科(A+醫學百科, 2011)的定義，支持性心理治療主要透過鼓勵、勸導、同情、支持與建議的方式，幫助病患認識問題、改善心境與提高信心，進而處理症狀與促進心身恢復，而其中進行此治療時，醫師必須富含情緒，熱情對待病患與尊重病患。認知行為治療是目前應用最廣泛的心理治療方式，是基於個人的看法與認知會決定自身的心情與行動的假設下進行的，如：悲觀的想法會產生悲觀的情緒與消極的行動，以此假設為出發點進而以改變認知，幫助病患認識會影響自身情緒的想法，並改正病患對於事物的不良看法，藉此改善病患情緒與行為。因此，根據本研究的目的，較為適合使用支持性心理治療的模式來幫助病患，使用正向與積極的態度去支持病患，給予病患正向積極的回覆，藉此改善病患狀況。

2.2 對話系統

隨著人工智慧的發展，人類一直在找尋如何藉由機器來幫助人類完成任務，而在自然語言領域中，智慧聊天機器人 (Intelligent Chat Bot) 是近年來極度熱門的項目，智慧聊天機器人可以輔助人們去完成許多事項，如：蘋果 Siri、Google 語音助理以及 Amazon Alexa。然而，要教導機器如何去和人類溝通、理解人類的語

言並給予回覆是件具有挑戰性的事，以下將會探討對話系統的兩個重要元件：自然語言理解和對話生成，並著重在對話生成的部分。



2.2.1 自然語言理解

為了使對話生成與說話者所輸入的文字有所連結，自然語言理解 (Natural Language Understanding, NLU) 便是重要的第一步，我們必須要理解使用者所想要表達的語義才能給予相對應的回覆。根據研究(Zhou et al., 2020)指出，為了讓使用者能夠更有參與感並持續使用，在聊天機器人中加入 EQ 是很有幫助的，EQ 指的是讓機器人擁有偵測並理解使用者的情緒、意圖、意見、對話行為與話題，並不只是單純取得語句的語義，並且加入 NLU 的元件可以幫助後續的對話管理與對話生成的表現。另外，Cahn (2017)也指出，在設計聊天機器人時，應加入情緒、意圖與對話行為的辨識系統，透過得到這些意涵可以幫助機器人更了解使用者想要表達的語義。而目前有許多的學者針對 NLU 做出研究，並隨著預訓練模型的出現，現今的 NLU 研究都是以 BERT 模型(Devlin et al., 2018)來進行，如在 Huang et al. (2019)的研究中，藉由 BERT 模型來進行連續對話的情緒偵測，模型可以更好的捕捉到情緒標籤。此外，在 Chen et al. (2019)的研究中，同樣使用 BERT 模型來進行文本意圖偵測，同樣得到超越以往 Seq2Seq 模型 (如 LSTM 與 RNN) 的效能，可以得知 BERT 模型在文本分類上的效能是非常高的。然而，目前的研究中，鮮少使用 BERT 模型同時進行多個標籤的文本分類，但在聊天機器人中，我們必須同時取得多種標籤來幫助機器人了解使用者的語義。因此本研究將延續上述研究，在對話系統中加入 NLU 的元件，並使用在文本分類有強大效能的深度學習模型-BERT 模型，先得到使用者輸入的文字的語義表達，再進行多標籤的分類。



2.2.2 對話生成

對話系統的核心在於如何生成回覆，而這就屬於自然語言領域中的自然語言生成 (NLG)。根據文獻(Cahn, 2017)，目前的回覆生成模型有以下數種：基於規則的模型 (Rule-based model)、基於資訊檢索的模型 (Information retrieval-based model)、統計機器翻譯生成模型 (Statistical machine translation generative model)、Seq2Seq 的模型 (Sequence to Sequence model) 以及基於預訓練模型 (Pre-trained model)。基於規則的生成模型的主要概念為模型內含有一個知識庫，知識庫內含多個文件，每個文件有問題和回答的語言對，而當系統接收到含有特定問題的語句，就會回覆此問題的回答，如問題為”你的名字是?”時，就會回覆”我的名字是...”，而通常這樣的 pattern 和 template 都是由人工制定的，因此十分耗時但可以確保模型生成符合期望的答案，最早的聊天機器人 ELIZA (Joseph Weizenbaum, 1966) 便是用這樣的模型打造的。

儘管基於規則的生成模型有諸多研究使用或以此為基準發展不同的種類，但此模型依舊有許多的限制，設計基於規則的生成模型時，問題與答案的組合必須由設計者一一寫入知識庫內，造成龐大的時間成本。而由於有許多大型聊天資料集，如 Twitter 與 FaceBook，這使得模型在生成時可以去搜尋現有答案來進行回覆，也稱之為基於資訊檢索的模型，藉由資料要集含的的大量對話，模型可以去尋找說話者的輸入與相對應的回答，就不必再自行設計對話。

而在 Ritter et al. (2011)的研究中，使用大量的 Twitter 貼文與回覆，以資料驅動 (data-driven)的 SMT 模型來訓練回覆生成，其表現優於傳統的資料檢索模型。同時，隨著深度學習的出現，Cho et al. (2014)提出以統計機器翻譯生成 (SMT) 模型為基礎，並以深度學習模型 RNN (Recurrent neural networks) 為底來得到更高的準確率，以編碼器 (encoder) 來對輸入做編碼，並以解碼器 (decoder) 來產生輸出，這樣的模型稱之為 Seq2Seq 模型，Seq2Seq 模型(Sutskever et al., 2014)現在



也被廣泛運用在對話生成上。許多學者以 Cho 等人的研究為基礎衍伸出許多研究，如(Asghar et al., 2018)使用 LSTM (Long short-term memory) 模型進行對話生成，來解決一般 RNN 模型會出現的梯度消失與梯度爆炸問題。

即便 Seq2Seq 模型在實務上被廣泛運用，但此類模型有一個缺陷，由於 Seq2Seq 模型都是藉由最大概似機率 (Maximum likelihood probability) 來產生出回應，使得機器人會受到資料集內頻繁出現的語句影響，產生較為通用且重複的回應，如：“我不知道”與“好的”這類語句。這樣並不能達到聊天機器人想要模仿人類的目的，同時多樣性低、不有趣且沒有意義的回覆也會降低使用者使用的意願。而(Li et al., 2016)的研究中對此提出解決辦法，嘗試將個人資訊轉換成向量後加入到 Seq2Seq 模型訓練中，藉由特定個人資訊來控制聊天機器人，讓機器人的回應可以保有一致性及多樣性。在(Lin et al., 2019)的研究中，則是將 32 種情緒加入到 Seq2Seq 模型中進行訓練，以編碼器來分類出語句的情緒，並以多個解碼器來訓練不同的情緒回應。同時在(Harilal et al., 2020)的研究中，使用了 LSTM 模型進行情緒分類與並將情緒與上文一起送入生成模型來生成回覆，兩個研究的結果評估上也確認到機器人的回覆確實包含特定情緒且呼應到使用者輸入，表明將個人資訊或情緒加入到模型中可以增加模型的回覆的資訊量、一致性與多樣性。因此，本研究將延續上述研究結果，將態度、主題與意圖三種維度加入到模型中進行生成訓練，提供更多的額外資訊給模型，讓模型藉由額外資訊抓取到正確的詞語或語句，確保生成的一致性與穩定性。

2.3 基於 GPT 模型之對話系統

由於硬體上限逐漸提升，電腦可以負荷大量計算資源，使得需要大量文本與參數量的預訓練模型逐漸盛行，語言預訓練模型如 BERT (Bidirectional Encoder Representations from Transformers) 與 GPT-2 (Generative Pre-trained Transformer 2) 在自然語言領域任務上，都有超出以往研究的成績。GPT-2 是由 OpenAI 團隊



於 2019 年時所提出，是 GPT 模型的升級版，基於 Transformer 模型的解碼器所構成的模型，最大的版本擁有高達 15 億的參數量，且由 800 萬頁網頁文本（約 40GB）所預訓練而成，也使得 GPT-2 模型對於文本有一定的語言理解與泛化能力，因此在面對多面向的下游任務時只需少量的特定資料便能訓練而成，不必為了特殊任務訓練一個模型。GPT-2 的核心架構是 Transformer 模型的解碼器，Transformer 模型是由(Vaswani et al., 2017)所提出的序列生成模型，與一般的序列生成模型不同，Transformer 使用了自注意力機制（self-attention mechanism），Transformer 透過此機制可以對整個序列進行平行運算，並且讓模型在學習時，可以同時關注到上下文的字詞，讓當前字詞可以藉由上下文來產生語義資訊，產生有語義資訊的輸出，這樣的模型架構在自然語言理解的任務上有著顯著的進步。

由於 GPT-2 是基於 Transformer 的解碼器(decoder)所構成，且在訓練時是採用單向的方式訓練，這個方式是指模型會藉由上文來預測下一個字，而不會根據上下文來做計算，這使得 GPT-2 在文本生成領域的效果較好，而 GPT-2 在文本生成已有許多熱門的應用上，如生成文章、小說與新聞，都有卓越的成果。同時，開始有人嘗試將預訓練模型應用於對話系統，在(Wolf et al., 2019)的研究中，使用了 GPT 模型來進行對話生成，將 GPT 模型用由(Zhang et al., 2018)所提出的 PERSONACHAT 資料集進行微調，將資料集的个人資訊資料轉換成詞嵌入（Word Embedding）並加入在文本之中，藉此將個人資訊融入到文本內供機器學習，在結果上可以勝過以往所有使用此資料集來做對話生成的研究，也透過案例分析表明對話確實有融入個人資訊，說明將額外資訊融入對話中，可以幫助模型生成更符合實際情況的回覆。而(Zhang et al., 2019)將大量的 Reddit 資料訓練在 GPT-2 模型上，訓練出以對話為基底的 DialoGPT，並顯示出其生成結果在相關性、內容性與一致性都勝過基準模型，由此可見 GPT 模型與 GPT-2 模型在對話生成領域是適用的。



而在(Zandie & Mahoor, 2020)的研究中，嘗試將情緒與使用者意圖加入 GPT 模型的訓練中，幫助模型生成的文本是帶有情緒並穩定模型輸出，同樣將情緒與使用者意圖轉為詞嵌入後加入到文本中，在自動評估指標上一樣有著超出其他模型的成果。儘管上述研究表明了 GPT 模型在模型生成的成功，但這些研究卻無法直接應用在實務上，原因在於這些研究在訓練模型時都是使用資料集中已標注好的資料來進行訓練，如個人資訊、情緒與使用者意圖，而這些資訊在與人實際對話時是不會事先提供的。而面對這樣的瓶頸，(Lin et al., 2020)提出了一個端到端 (end-to-end) 的對話系統，但可惜的是此研究是直接將資料集中的文字送入 GPT 模型做訓練，並未將情緒標籤融入文本再送入模型。因此，端到端且含有維度的預訓練對話系統的研究是較為缺乏的，且上述研究都是以英文為主，在使用預訓練模型生成中文對話的研究較少。故本研究會將 BERT 模型與 GPT-2 模型做結合，先使用 BERT 模型對文本進行分類，並將分類出來的態度、意圖、主題三種標籤融入文本中，再送入 GPT-2 模型進行中文對話生成，達到端到端模型的目的。

2.4 現有對話生成評估方法

自從對話生成的概念出現後，許多人開始研究各式各樣的對話生成方式，使用不同的模型架構來進行訓練，也因此需要評估標準來評估對話生成的性能，為了更加了解目前主流使用的對話生成評估方法，下方將探討(Finch & Choi, 2020)，所統整的近年來 20 篇對話系統研究的評估方法，點出其中的優劣之處，接著說明現有評估方法的問題，最後講述本論文將會用來評估實驗結果的指標。

2.4.1 自動評估指標

根據(Finch & Choi, 2020)，自動評估 (Automated Evaluation) 是指透過數學公式將對話的內容進行計算，提供對對話生成的客觀定量測量，使得自動評估是最

客觀、最直接與成本低的評估方法。以下內容將探討目前主流使用的自動評估方法，作為本論文對話生成評估方法的選擇參考。



BLEU (Bilingual evaluation understudy) 是由(Papineni et al., 2002)所提出一種自動評估指標，採用 N-gram 的方式並藉由數學公式來評估機器所生成的文字與人類所標記的文字之間的正確性。在許多相關研究中，如(Lin et al., 2019)、(Zandie & Mahoor, 2020)、(Su et al., 2021)，皆有使用 BLEU 來評估對話生成，可見 BLEU 是目前現今常用的評估指標之一。

Perplexity (PPL) 是由(Jelinek et al., 1977)所提出，是用評估語言模型的指標，以詞為單位來估計一句話出現的機率，並使用句子長度來進行標準化，降低句子長度的影響，PPL 的概念可以理解成預測下一個字會有多少選擇，舉例來說，若是模型的 PPL 為 90，表示模型在生成語句時下一個詞會有 90 個選項可供選擇，因此當 PPL 越低，模型的可選詞越少，模型更加準確。同時，根據文獻 (Adiwardana et al., 2020)指出，PPL 與此研究所提出的 SSA 人工評估指標有一定的相關性，若是 PPL 數值低的模型，在人工評估的分數表現上會較佳，說明 PPL 數值的高低與會影響到人工對於生成結果的評估。

Distinct 由(Li et al., 2016)所提出，Distinct 的出現是由於近年來 Seq2Seq 模型在對話生成領域使用上較為頻繁，而 Seq2Seq 模型是一種最大概似估計 (Maximum Likelihood Estimate) 的模型，容易產生一些通用且簡短的詞語。因此，Distinct 用來判斷模型生成的詞語的多樣性高低，以 n-gram 的方式切割字詞。Distinct 計算不重複的 unigrams 或 bigrams 數量，並除以總體文字數量，藉此計算整體生成的不重複率，Distinct 值越高，代表模型生成文字的多樣性越高。

2.4.2 人工評估指標

人工評估意指由人類來對對話生成文本進行評估，這類評估並不會有準確的數



學公式或方法來評估，而是人類以使用者的角度進行主觀評論，在邏輯上比起基於標準答案做評分的自動評估較為適用。人工評估的指標通常以不同對話生成的目標而有所不同，且根據(Finch & Choi, 2020)指出，人工評斷的方式即使使用相同的方法名稱，其定義也有所不同，不同名稱的方法可能類似的定義，因此將各種指標依照其功能與敘述將其分為三大類：Grammatical Capability、Turn Coherence 與 Response Informativeness。Grammatical Capability 主要注重生成文本的文法正確性，文法正確性是指生成文本是否符合文法，流暢度是指生成文本是否自然且閱讀流暢，可讀性是指生成文本對於使用者可否順利閱讀。Turn Coherence 主要注重生成文本與輸入的相關性與一致性，判斷生成文本對於歷史對話是否有關與是否符合主題，如說話者在講述運動主題，而機器人卻回答電影主題，則在此類別的分數就會較低。Response Informativeness 注重生成文本對於輸入是否含有特定資訊或知識，是否屬於特定領域回答，如：電影相關、醫療相關或運動相關，與生成文本的文字多樣性程度，若所生成的文本都是“是喔”或“我不知道”這種較為一般的回答，則代表多樣性程度低。

2.4.3 現有評估指標之問題

從上述可知，不論是自動評估或人工評估都被廣泛使用，然而這兩種評估方法都有存在缺陷。在(Liu et al., 2016)的研究中指出，自動評估方法與人工評估的相關性較低，說明自動評估方法不應使用於對話系統的評估上，因為在實務上，與他人對話時是沒有標準答案的，一種輸入可能有多種回答，這些回答都可以是合理的，因此以標準答案為準的自動評估並不是個合適的評估方式。而人工評估在各個研究所制訂出來的標準有所不同，名稱相同但定義卻不一致，即使名稱不相同但定義卻雷同，如在文獻(Finch & Choi, 2020)中就指出，流暢度這個指標在同一年（2019）就有五篇研究使用，但這五篇研究對於流暢度的定義皆不相同。此外，人工評估是由人類進行評分，而目前沒有統一機關來執行，使得各研究的評



斷人員的素質不一，分數會受到嚴重的主觀影響。

然而，即便兩種評估方式都有其缺陷存在，但由於沒有統一且合適的評估方式，目前的研究還是以這兩種方式為主。而本研究是以憂鬱症病患為對象來設計對話系統，因此在自動評估指標上會以 Perplexity (PPL) 這個指標作為評估方式，使用 PPL 的原因是要確定語言模型的流暢度與效能，同時 PPL 與人工評估分數也有一定的相關性。而不使用其他自動評估指標是因為針對問題或上文，並非只有一種回覆，因此使用以真實回覆為主的評估方法並沒有辦法確切評估生成文字的好壞，因此主要的評估會著重在人工評估上。而在人工評估上，本論文根據文獻(Finch & Choi, 2020)，所整理出來的評估指標，人工評估指標種類 (表 1.) 來進行選取。

Dimension	Definition
Grammaticality	Responses are free of grammatical and semantic errors
Relevance	Responses are on-topic with the immediate dialogue history
Informativeness	Responses produce unique and non-generic information that is specific to the dialogue context
Emotional Understanding	Responses indicate an understanding of the user's current emotional state and provide an appropriate emotional reaction based on the current dialogue context
Engagingness	Responses are engaging to user and fulfill the particular conversational goals implied by the user
Consistency	Responses do not produce information that contradicts other information known about the system
Proactivity	Responses actively and appropriately move the conversation along different topics
Quality	The overall quality of and satisfaction with the dialogue

表 1. 人工評估指標種類

(資料來源：(Finch & Choi, 2020))

本論文根據上表 (表 1.) 與研究目的選取下列指標：相關性 (Relevance)、情緒理解性 (Emotional Understanding) 與參與性 (Engagingness) 這三種指標來做評估。希望評斷對話系統的回答與使用者的輸入有一定的相關性，同時希望回答是可以針對問題並以正向積極的態度來回覆使用者，並且是有與使用者互動並給予使用者建議或意見來達到減緩病患壓力與舒緩情緒等目的。



2.5 總結

根據文獻探討的結果，本論文整理出以下結論：

1. 心理治療中的支持性心理治療與認知行為治療已被廣泛運用在治療憂鬱症上，且都有文獻說明對於憂鬱症治療有一定的效用。因此，本研究認為理解病患意圖與情緒，並以正向積極的態度和病患對話與並給予有幫助的建議可以緩解憂鬱症症況，這就是本研究欲將態度、意圖與主題導入對話系統的主要原因。
2. 在對話系統架構中，Seq2Seq 模型是現今最廣泛使用的，但有許多學者意識到 Seq2Seq 模型有通用且重複性生成的問題，並提出將個人資訊、情緒與主題等維度加入到對話生成中，都有顯著的改善。GPT 與 GPT-2 預訓練模型也開始被應用在對話生成領域，也有許多學者嘗試將標籤融入到對話中做訓練，成果皆超出 Seq2Seq 模型。然而，對於先將對話分類得到多個標籤再融入對話送入 GPT-2 模型此類端到端的研究較為缺乏，因此本研究欲將對話先用 BERT 模型分類得到維度再送入 GPT-2 模型訓練，來確保標籤可以融入到對話中，同時控制模型產生出與標籤相關的回覆，來達成本研究想要回覆給使用者正面積極的語句的目的。
3. 根據我們的調查，不論是自動評估方法和人工評估方法都有其缺陷存在，但由於目前沒有統一機構或方法來評估對話生成，大部分的研究都還是以這兩種方式來進行評估。因此本論文欲以注重模型效能且與人工評估指標有一定相關程度的 Perplexity (PPL) 指標作為自動評估方法。此外，人工評估方法則從表 1. 中選取相關性、情緒理解性與參與性這三種指標來做評估，藉由人工評估確保對話系統實際與人類對話時也有穩定的生成。

第三章、研究方法



3.1 研究架構

本論文與 Huang(2022)提出一個以憂鬱症為導向所設計的聊天系統架構，聊天系統架構將實驗架構分成三大階段，第一階段由 Huang 首先對使用者所輸入的對話進行初步判斷，確認使用者是否需要幫助，若沒有則進行簡單對話並結束對話，若需要幫助則進入第二階段，第二階段 Huang 對使用者所輸入的對話進行自然語言理解 (NLU)，將對話進行分類，分類出態度、意圖與主題三種維度的標籤，作為第三階段的輔助輸入。在第三階段本論文會將第二階段所分類出來的標籤與對話轉換成 token 並進行合併，送入深度學習模型 GPT-2，得到回覆來回應使用者。而本論文主要針對第三階段進行嘗試與實驗。聊天系統流程圖如圖 2.所示。

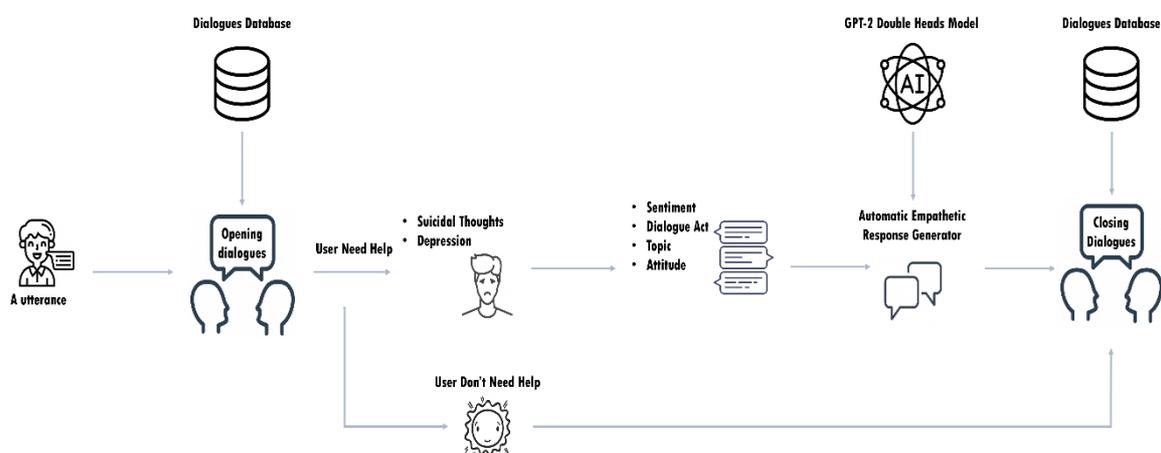


圖 2. 聊天系統流程圖

本論文在第三階段希望能夠藉由對話的方式來舒緩使用者情緒與陪伴使用者，因此藉由深度學習模型 GPT-2 來產生與對話相符合的回答，並將態度、意圖與主題融入文本來幫助生成模型能夠產生出帶有正面積極態度的回覆來回覆給使用者。



3.2 資料集

本論文將會使用兩個對話資料集來對 GPT-2 模型進行微調 (Fine-Tuning)，英文資料集 DailyDialog(Li et al., 2017)與由本論文與 Huang(2022)所共同整理出來的中文對話資料集 ChinesePsyQA，使用 DailyDialog 主要是實驗在大量文本資料下，本論文的研究方法的效果如何，而 ChinesePsyQA 則是由於目前大多數的對話資料集都是以日常對話為主，在心理問題方面的對話資料集十分缺乏，加上大多數的對話資料集都是以英文為主，因此為了符合研究目的，本論文與 Huang 自建一個中文心理對答資料集來進行實驗。

3.2.1 DailyDialog

DailyDialog 是一個高質量的多輪對話資料集，資料集中的對話都是人工編寫，使得資料集內的雜訊較少，主要以人類日常的對話方式來呈現。資料集由兩個部分所組成：

1. 多輪對話：對話內容是從多個網站中爬蟲下來的，這些網站中的對話都是英語學習者在平時練習英文對話所寫下的，這些對話通常都有特定的主題與相對應的內容，將多於兩個人的對話內容過濾掉並糾正錯字。此對話集共有 13,118 個對話，平均每個對話大約有 8 輪，平均每個對話大約有 115 個字，平均每句話大約有 15 個字。
2. 標籤：此資料集也對對話內容進行人工標記，將每句對話都標記三種標籤，情緒、主題與意圖。各標籤詳細維度如下：
 - 甲、情緒：憤怒、厭惡、恐懼、喜悅、悲傷、驚喜。
 - 乙、主題：日常、學校、文化與教育、態度與情緒、關係、觀光、健康、工作、政治、財經。
 - 丙、意圖：告知、詢問、指令、承諾。

多輪對話與標籤範例如下表 2.所示:

Topic : <Relationship>		
Utterances	Emotion	Intent
We are through .	anger	inform
Why ?	surprise	question
You impose too much on me , and I'm tired away .	anger	inform
I'm very sorry for that	sadness	inform
I don't want to see you anymore .	anger	inform
Please clam down .	sadness	directive

表 2. Daily Dialog 對話範例

資料集中各標籤的分布如下圖 3.、圖 4.、圖 5.所示：

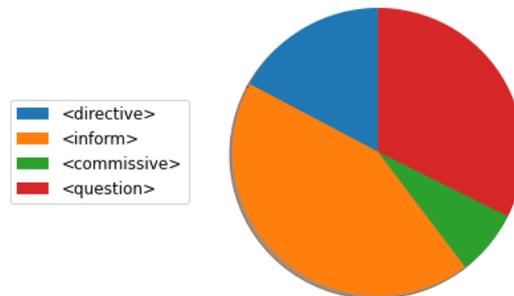


圖 3. DailyDialog 意圖維度分佈

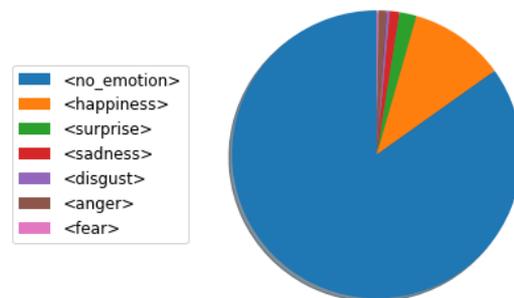


圖 4. DailyDialog 情緒維度分佈

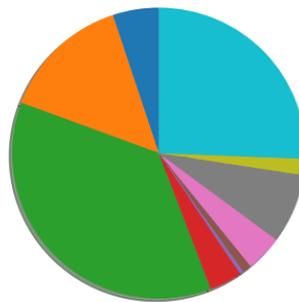


圖 5. DailyDialog 主題維度分佈

3.2.2 ChinesePsyQA

ChinesePsyQA 是由本論文與 Huang 所共同建構的資料集，資料集中的對話都是由真實的心理問答所組成，並且都經過人工的過濾，將不合理的文字或符號去除，並由兩人以上來針對每句語句進行標記，為資料集建構出主題、意圖與態度三種標籤，而資料集的組成如下所示：

1. 單輪對話：對話是由壹心理網站(<https://www.xinli001.com/>)中爬取貼文，每則貼文都會含有問題以及回覆，而每組問答都會有相對應的主題，在此資料集中共有約 15,489 句，9,882 組問答，平均每組問答有 100 字，貼文與回覆的範例請見附錄一。
2. 標籤：將問題與回覆整理好後，對這些對話進行標籤標記，由兩人共同對給予語句態度、意圖與主題這三種標籤，若兩人標記的標籤相同，則給予該語句標籤，若兩人標記的標籤不同，則參考第三人的意見來進行標記，並選取較多人標記的標籤，各標籤的詳細維度如下：

甲、態度：正向、負向、中性。

乙、主題：存在議題、人際、性格、親密關係、成就、生心理問題、童年/家庭。

丙、意圖：告知、詢問、指令。

單輪對話與標籤範例如下表 3.所示:

主題: <人際>		
問題	態度	意圖
怎樣和那種很愛比較的人相處?	中性	詢問
回覆		
1 每個人都是不同的,大家互相包容吧。	正向	告知
2 保持一點距離,能不理就別理吧。	中性	指導
3 果斷和他絕交吧,這種人沒必要和他相處。	負向	指導

表 3. ChinesePsyQA 對答範例

資料集中各標籤的分佈如下圖 6.、圖 7.、圖 8.所示:

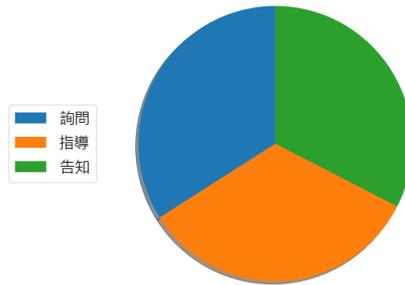


圖 6. ChinesePsyQA 意圖維度分佈

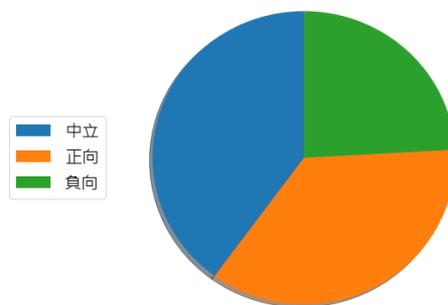


圖 7. ChinesePsyQA 態度維度分佈

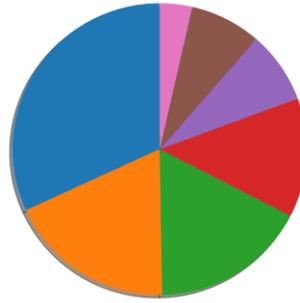
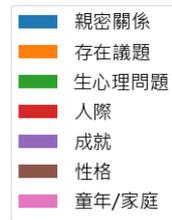


圖 8. ChinesePsyQA 主題維度分佈

3.3 模型架構

在此小節，我們會針對生成模型進行詳細的說明，包含模型的架構以及模型輸入的格式。

3.3.1 模型輸入格式

GPT-2 預訓練模型雖然有輸入長度限制，但由於對話資料集內的每個對話平均長度大約為 100 字，這樣的長度是在 GPT-2 模型可以接受的範圍內，因此可以將整個對話同時送進去模型中。而在將文本送入模型前，我們必須先對文本與標籤進行轉換，將文本與標籤轉換成 token，而標籤則會被設定為特殊 token。同時，我們也必須把轉換好的維度標籤與文本相嵌起來，我們會將這些標籤放置在文本前，將問題的主題、態度與意圖放置在問題前，接著將回覆的態度與意圖放置在回覆前，最後將問題與回覆相接在一起。此外，我們會隨機從對話資料集中選取一句回覆當作混淆語句，加上標籤後將此語句與問題合併，共同與真實問答送入模型藉此來混淆模型，整體轉換過程如圖 9.所示。

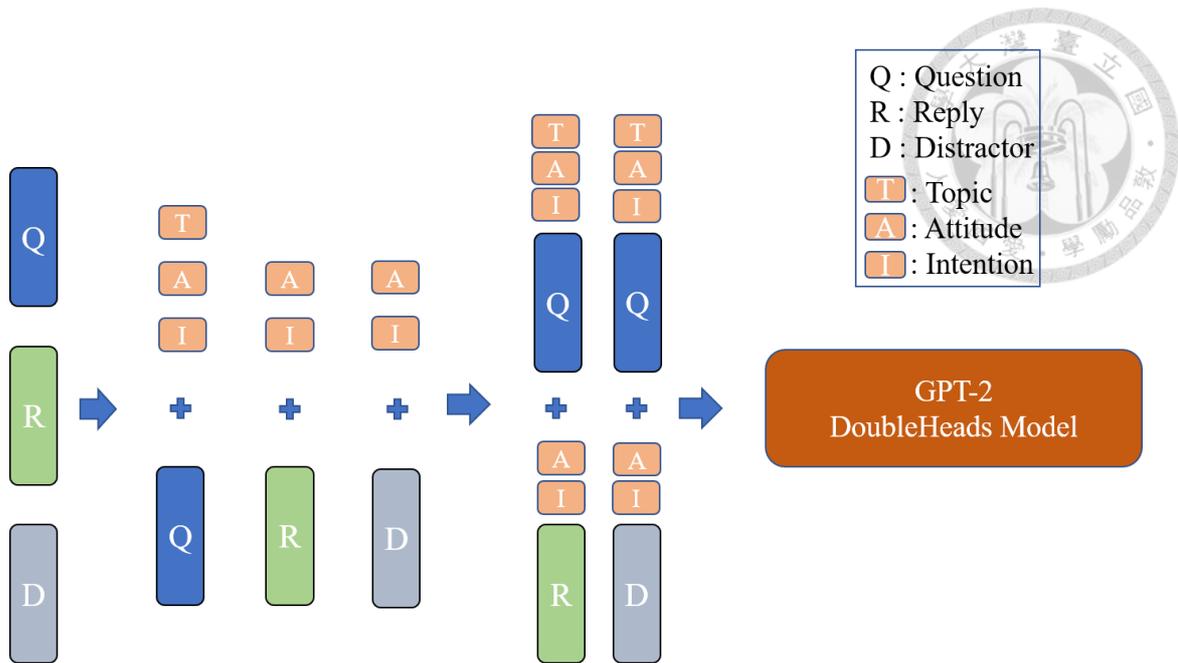


圖 9. 模型輸入格式圖

透過上述的模型輸入格式，我們可以運用到 GPT-2 模型的特性，將標籤置於文本前方，讓模型在訓練時，可以將標籤 token 當作文本的一部分，學習基於當前的標籤 token 該生成的下一個 token，藉由此方式讓模型知道每種標籤相對應的文本為何者，也透過三種標籤的不同組合來提供更多的額外資訊給模型，使模型更能確定該生成的 token 與整句語句，同時也藉由混淆語句來幫助模型學習整體真實回覆的表達方式，讓模型同時學習到局部與整體的語句生成。

3.3.2 GPT-2 DoubleHeads Model

在本論文，我們使用 GPT-2 的 DoubleHeads Model 來進行實驗，使用 DoubleHeads 的架構原因在於，我們希望模型不僅只學習到藉由上文產生下一個 token 這樣局部的生成方式，同時也希望模型能夠學習到整個回覆語句的表達方式，因此本論文使用 GPT-2 DoubleHeads Model 具有兩個 heads 的架構，同時讓模型學習局部的文本生成與整句語句的表達方式。在微調 (fine-tuning) 階段，我們參照以往研究在微調 GPT-2 模型時的方法，透過監督式的方式，將對話資料傳入模型中，進行下一個字的預測，使用 Language Modeling Head 來計算損失

值，產生出 LM Loss(Language Modeling Loss)，幫助 GPT-2 模型對對話資料進行語言建模。此外，在此模型下我們也會進行下句預測(Next Sentence Prediction)，藉由讓模型學習判斷語句是否為下個語句，並計算出 MC Loss(Multiple Choice Classification Loss)，讓模型不僅只學習到局部的 token 生成，也能學習判斷問題對應的整體語句。訓練完成後，我們將歷史對話或問題輸入進微調後的 GPT-2 模型中，讓模型依序輸出字詞，最後得到整句回覆，整體模型架構如圖 10.所示。

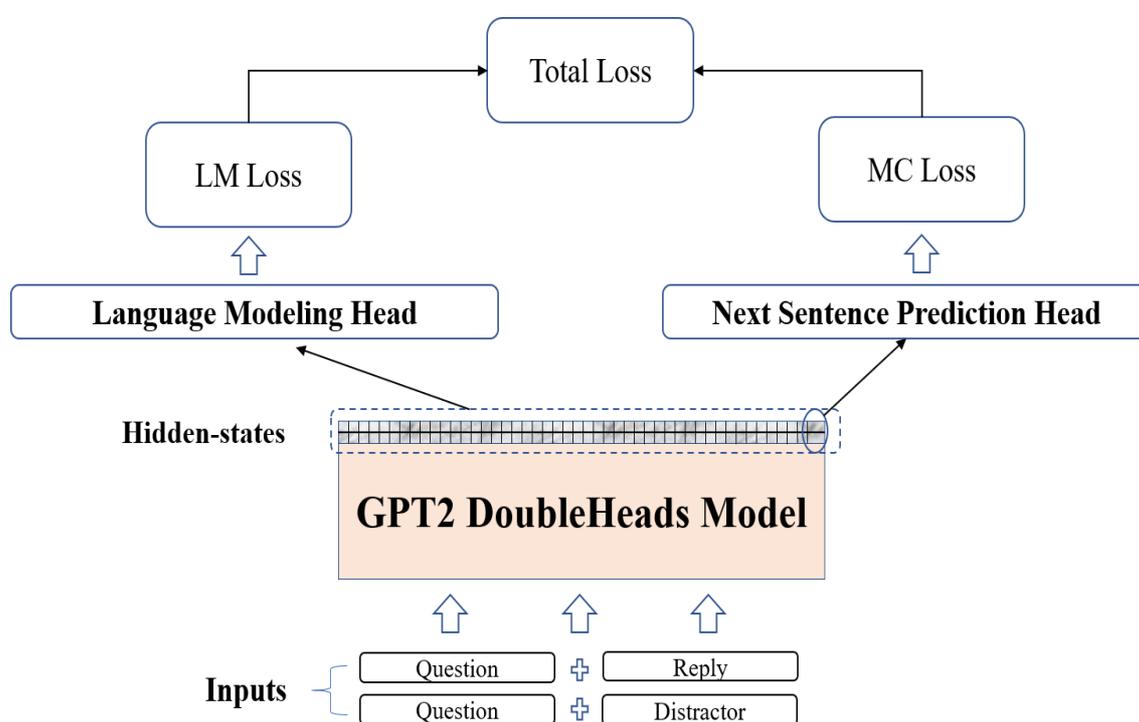


圖 10. 模型架構圖

3.3.2.1 Language Modeling Head

Language modeling head 主要是用來幫助模型學習如何藉由上文來預測下一個字，來針對我們的資料集進行語言建模，讓模型可以針對輸入的語句來產生相對應的回覆，並且在損失函數的部分使用 cross-entropy 來計算損失值。

$$\mathcal{L}_{lm}(\mathcal{S}_T) = -\sum_{i=1}^N \log P(w_i | w_1, w_2, \dots, w_{i-1}) \quad (1)$$



將我們要預測的語句設為 \mathcal{S}_T ， \mathcal{S}_T 內共有 N 個文字 w ， $\mathcal{S}_T = \{w_1, w_2, \dots, w_N\}$ ，我們依序使用前 $i - 1$ 個文字來預測第 i 個字，而使用 cross-entropy 計算出來的損失值稱為 \mathcal{L}_{lm} 。

3.3.2.2 Next Sentence Prediction Head

Next sentence prediction head 則是用來對我們所輸入的真實回覆與混淆語句進行分辨，這麼做是為了讓模型不僅學習到局部的文字生成，也讓模型注重整體回覆是否與問題相符合，讓模型產生的回覆更加貼近問題，同時損失函數一樣使用 cross-entropy 來計算 loss 值。

$$\mathcal{L}_{mc}(\mathcal{S}_{1:N}) = -\log P(r | s_1, s_2, \dots, s_{N-1}) \quad (2)$$

$\mathcal{S}_{1:N}$ 為我們所輸入的每組對話， $\mathcal{S}_{1:N} = \{s_1, s_2, \dots, s_{N-1}, s_N\}$ ，代表每組對話會有 N 個語句， s_N 可以是真實回覆 s_t ，也可以是我們所隨機選取的混淆語句 s'_t ，在英文資料集中，我們所輸入的對話可能有多個語句，此時 N 為不定值。而在中文資料集中則會只有問題與回覆，此時 N 就會為2。當我們輸入整組對話 $\mathcal{S}_{1:N}$ ，會使用 $\mathcal{S}_{1:N-1}$ 來預測 s_N 是否為真實回覆，而使用 cross-entropy 計算出來的損失值稱為 \mathcal{L}_{mc} ，同時我們將 r 定義為：

$$r = \begin{cases} 1, & s_N = s_t \\ 0, & s_N = s'_t \end{cases} \quad (3)$$

最後，我們會將 Language Modeling Head 與 Next Sentence Prediction Head 的損失值(loss)加權後更新模型，來達到訓練的效果。

$$\mathcal{L}_{Total} = \alpha_1 \mathcal{L}_{lm} + \alpha_2 \mathcal{L}_{mc} \quad (4)$$

其中 α_1 與 \mathcal{L}_{lm} 代表 language modeling head 的損失係數(loss coefficient)與損失值



(loss), 而 α_2 與 \mathcal{L}_{mc} 則是代表 next sentence prediction head 的損失係數與損失值。

3.4 研究驗證

由於本論文主要著重於對話生成部分，因此在研究驗證的階段主要評估對話生成的方式來進行，而本論文在研究驗證主要分成兩個部分，自動評估與人工評估兩種方式，我們會將資料集切分成訓練、驗證與測試資料集，使用訓練集來訓練模型，驗證集用來調整模型，並用測試集來進行評估。

3.4.1 自動評估

自動評估泛指以數學公式來對生成文字與標準答案進行計算，將測試集的對話送入 GPT-2 模型後，得到生成文本，並與測試集內的回覆進行計算，在本論文使用 Perplexity 這指標。使用 Perplexity 指標主要是要評估我們 GPT-2 模型對於對話資料的語言建模程度，同時根據文獻(Adiwardana et al., 2020)指出，PPL 數值的高低與會影響到人工對於生成結果的評估。Perplexity 指標的公式如下：

$$PP(S) = \sqrt[T]{\prod_{i=1}^T \frac{1}{p(w_i | w_1 w_2 \dots w_{i-1})}} \quad (5)$$

S 代表輸入的句子， T 代表句子長度。

3.4.2 人工評估

在人工評估中，我們將會建立表單並請五名受測者對生成文本進行評估，我們將從 ChinesePsyQA 的測試資料集中隨機選取一百個問題，並使用三種模型來生成回覆並評估分數。而本論文將會使用由文獻(Finch & Choi, 2020)所整理出的統一人工評估指標，並從中選取相關性 (Relevance)、情緒理解性 (Emotional Understanding)、參與性 (Engagingness) 這三種指標來做評估，讓使用者對回覆

進行評估，分數範圍為一至五分，一分為非常不同意，三分為普通，五分為非常同意，希望藉此評斷對話系統的回答與使用者的問題是屬於同個主題的，同時希望回答是可以針對問題並以正向積極的態度來回覆使用者，並且是有與使用者互動並給予使用者建議或指示來達到減緩病患壓力、舒緩情緒或解決問題等目的，其定義如表 1.所示。



第四章、研究結果



本章節我們將對於在本論文所提出的方法進行實驗並將結果呈現出來，將會說明用來與本論文方法比較的模型與模型的超參數設定，同時對模型進行自動評估與人工評估並展示結果，最後針對回覆內容進行分析與分析對話系統串連的結果。

4.1 Baseline Models

在本論文中為了探討我們所提出的方法是否有效果，因此使用以下模型來與我們所提出的模型比較：

1. Seq2Seq+Attention Model：由(Sutskever et al., 2014)所提出的 Seq2Seq 模型加上 Attention 層，並使用 cross-entropy loss 來訓練模型。
2. Transformer Model：由(Vaswani et al., 2017)所提出的 Transformer Model，其中使用了 Self-Attention 的機制，並使用 cross-entropy loss 來訓練模型。
3. GPT-2 Base Model：由(Radford et al., 2019)所提出的 GPT-2 Model，使用只有 Language Modeling Head 的基礎模型架構，同時使用 cross-entropy loss 來訓練模型。
4. DialoGPT：由(Zhang et al., 2019)所提出的模型，使用大量 Reddit 對話資料並基於 GPT 模型預訓練而成，本論文使用英文資料集在 DialoGPT 上進行微調 (fine-tuning)，來得到訓練好的模型。

我們使用了過往研究中較常使用的 Seq2Seq 模型與 Transformer 模型，希望藉此比較過往模型與本論文研究方法的效果，同時由於我們是基於預訓練模型來進行實驗，所以也使用預訓練模型 GPT-2 base 來相互比較，希望藉此了解使用本論文的研究方法可以在預訓練模型上進步多少。此外，在英文資料集 DailyDialog 上，我們額外使用 DialoGPT 當作比較的對象，但由於語言限制的問題，因此僅



在英文資料集上進行實驗。同時，為了探討有無加入對話標籤的效果，本論文也會比較基礎的 GPT-2 DoubleHeads Model (GPT-2 DH Model)與加入對話標籤的模型(GPT-2 DH Model w/ label)的結果。

4.2 訓練參數

由於中文資料集的資料數量與英文資料集不同，因此使用的模型參數不同，本小節將詳細說明在兩種資料集下的模型參數設定。

中文資料訓練的模型中，預訓練模型我們皆使用 CKIP 的 gpt2-base-chinese Model，將 learning rate 設為 $2e-5$ ，最大長度設為 150 個字元，訓練 10 個 epoch，並將 \mathcal{L}_{lm} 與 \mathcal{L}_{mc} 的損失係數設為 1:1。Seq2Seq 模型，我們設定 learning rate 為 $5e-4$ ，隱藏層節點數量為 128 個，訓練 30 個 epoch。在 Transformer 模型中，我們設定 learning rate 為 $1e-3$ ，層數為 6 層，head 數量為 8 個，節點數量為 512 個，訓練 20 個 epoch。

英文資料訓練的模型中，預訓練模型我們皆使用 GPT-2 Base Model 來進行實驗，將 learning rate 設為 $6.25e-5$ ，最大長度設為 310 個字元，訓練 20 個 epoch，並將 \mathcal{L}_{lm} 與 \mathcal{L}_{mc} 的損失函數設為 2:1。Seq2Seq 模型，我們設定 learning rate 為 $1e-4$ ，隱藏層節點為 256 個，訓練 20 個 epoch。Transformer 模型，我們設定為 learning rate 為 $1e-4$ ，層數為 6 層，head 數量為 8 個，節點數量為 512 個，訓練 20 個 epoch。

4.3 自動評估結果

在自動評估的部分，我們使用由(Jelinek et al., 1977)所提出的 Perplexity(PPL)，用來評斷語言模型的生成品質，而準確率(Accuracy)則是模型正確判斷我們所輸入的回覆是真實回覆的比率，準確率越高代表模型越能從真實與混淆語句中辨別



出真實語句，而由於 Accuracy 只在 GPT-2 DoubleHeads(GPT-2 DH)的模型架構下才會計算，因此其餘的 Baseline Models 將不會呈現準確率，Baseline Models 與本論文的模型比較如下表 4.所示。

Model	Daily Dialog		ChinesePsyQA	
	Accuracy	PPL	Accuracy	PPL
Seq2Seq	-	28.94	-	204.01
Transformer	-	23.37	-	83.927
GPT-2 Base	-	11.419	-	13.187
DialoGPT	-	11.28	-	-
GPT-2 DH	73.51	10.886	82.36	12.975
GPT-2 DH w/ label	75.79	10.727	87.05	12.014

表 4. 自動評估結果表

從上述的結果得知，在 DailyDialog 資料集中，我們所提出的方法在 perplexity(PPL)上，相較於 Seq2Seq 模型與 Transformer 模型皆有超過 50%的進步，而在比較 GPT-2 Base Model 與由(Zhang et al., 2019)所提出的 DialoGPT 這兩個預訓練模型上，也分別有 6%與 5%的進步。同時在比較有無加入額外標籤的結果時也可以發現，在文本前加入標籤的模型，在 PPL 與準確率上皆勝過沒有加入標籤的模型。而在由我們自建的 ChinesePsyQA 資料集中，perplexity(PPL)在 Seq2Seq 模型與 Transformer 模型中表現都不是很理想，我們認為是由於資料集數量的欠缺，導致沒有經過預訓練的模型沒辦法學習到各種問題與相對應的回覆，導致在測試集中沒辦法有良好的效果，儘管如此，我們所提出的方法與 GPT-2 Base Model 相比在 PPL 與準確率上有著 9%與 4.7%的進步，同樣在加入標籤訓練後的模型在 PPL 與準確率皆勝過沒有加入標籤訓練的模型。綜合上述，我們將態度、意圖與主題這三種標籤加入到文本前，可以幫助模型在 PPL 與準確率上有著一定的進步，說明將額外資訊加入到文本中，可以幫助模型鎖定所要學習到的語



句，讓模型知道針對輸入的標籤與上文該產生的 token 為何，增加語言建模的能力，也提升模型判斷語句的能力。

4.4 人工評估結果

在人工評估的部分，我們從網路上徵求到五位受試者來進行評估，每位受試者皆有大專院校以上的學歷，我們從中文資料測試集中選取 100 組問題，並且使用 Transformer 模型與 GPT-2 Base 模型這兩個現今對話生成研究中較常使用的模型生成回覆，並與本論文的方法所生成的回覆進行比較。在使用本論文的方法生成時，會將三種標籤，態度、意圖與主題設置為正向、指導與該問題的主題，接著針對三個模型所生成的回覆請受試者針對相關性(Relevance)、情緒理解性(Emotional understanding)與參與性(Engagingness)這三個指標進行評分，一分為最不符合指標，五分為最符合指標，各指標的詢問方式如表 5.所示，最後將結果平均後呈現在下方表 6。

	詢問方式
相關性	此句回覆與問題擁有相同的主題（講述同一種類的事物）
情緒理解	此句回覆帶有正面積極的態度。
參與性	此句回覆對於問題是句帶有幫助性的回覆。

表 5. 人工評估指標詢問方式

Model	相關性	情緒理解	參與性
Transformer	3.088	3.574	2.976
GPT-2 Base	3.261	3.670	3.136
GPT-2 DH w/ label	4.042	4.076	4.052

表 6. 人工評估結果表



根據上述結果所示，本論文所提出的方法在相關性、情緒理解與參與性這三個指標上皆勝過 Transformer 模型與 GPT-2 Base 模型，說明藉由將態度、意圖與主題這三個標籤加入到對話文本中，可以提升回覆與問題的相關程度，也可以幫助模型找到文本中較為正向積極的回覆來提升其積極程度，最後也能藉由設定好的指導標籤來回覆使用者較有幫助性的回覆，三種模型的生成範例請見附錄二。

4.5 回覆內容分析

在此小節，我們將針對回覆內容進行分析，由於本論文會將三種標籤送入模型來生成回覆，因此我們從人工評估的問題內，針對態度、意圖與主題三種標籤進行分析，對於每種標籤都隨機抽取五組問題，並使用不同的維度來生成回覆，探討各標籤在不同維度下對於對話生成的影響。

首先我們將態度標籤分別設定為正向、中性與負向，根據問題生成回覆並進行比較，來研究本論文所使用的方法是否可以改變對話生成的內容，結果如下表 7 所示。

問題	With Positive Label	With Neutral Label	With Negative Label
煩躁持續兩個禮拜，每天無所事事，各種什麼都不想做，怎麼辦？	可以嘗試多和朋友聊聊，多多運動，讓自己變得更好。	煩惱是自己內心的不安全感，需要時間來解決。	煩惱是因為你對自己的不滿。
怎麼樣建立自信？	建議你去做一些自己喜歡的事情，這樣才能讓你成長起來。	建議你去做一些自己喜歡的事情，這樣你才能更好地接受你。	建議你可以選擇諮詢師。
行動力在大學裡嚴重喪失，情緒反覆，怎麼辦？	首先要學會用合理的方式釋放自己內心的負擔，這樣才能有效解決問題。	這個問題是需要時間來解決的，不要太在意別人的看法。	這種情況對你來說是一種傷害。
我以後該怎麼對待欺騙我的人？	你可以試著和他溝通，讓他知道你的想法，然後再做決定。	不管怎樣，都要學會接納，不要害怕別人。	你的行為是對自己的一種傷害，所以不要刻意排斥他人。
我發現我誰都不愛，近乎瘋狂的尋找著新的目標，我好像挺花心的，怎麼辦？	愛是人生的一部分，只要你願意去追求自己的幸福，你就會成功的。	試著找到自己的興趣愛好，去嘗試一下你喜歡的事情，這樣你才能讓你成長起來。	時間會讓你失去自己的信心。

表 7. 不同態度標籤之生成範例

根據上表所示，當我們將不同的態度加入到對話前，會使得模型所產生的回覆不同，在加入負向的標籤時會讓模型傾向產生針對使用者的狀況進行批判或是給予消極的建議。而在加入正向的標籤時則會讓模型的回覆傾向希望使用者嘗試去做能夠幫助到自己的事或是鼓勵使用者勇敢去追求事物。

接著我們設定不同種類的意圖標籤，實驗不同意圖對於對話生成的影響，我們將意圖分別設定為指導與告知這兩個意圖來生成回覆，結果如下表 8.所示

問題	With Directive Label	With Inform Label
想要交到朋友必須要分享祕密嗎？	可以試著和朋友溝通，瞭解他的想法和感受。	每個人都有自己的選擇，不要太在意別人的看法。
婚姻出現了第三者，怎麼辦？	建議你們多多溝通，了解彼此的想法。	婚後的幸福是你們兩個人的事情，要一起努力的。
覺得自己很醜，很自卑，怎麼辦？	不要太在意別人的看法，只要做好你能做的就好。	人生的目的是為了更好的生活而努力。
我以後該怎麼對待欺騙我的人？	你可以試著和他溝通，讓他知道你的想法，然後再做決定。	人與人之間的關係是相互理解的，相信自己的內心會有很大的改變。
反覆提離婚，不知道如何繼續？	建議你先把自己的想法告訴他，然後再做決定。	婚姻是兩個人的事情，彼此尊重和理解才是最重要的。

表 8. 不同意圖標籤之生成範例

根據上表 8. 所示，當我們將意圖設定為指導時，會使模型傾向生成建議或指示使用者如何去解決問題的回覆，而當設定為告知時，則會使模型傾向生成對於問題的看法。

最後，我們設定是否加入主題標籤，來分析家入對話標籤與否對於對話生成的影響，結果如下表 9. 所示。

問題	With Topic Label	Without Topic Label
自己捨不得花錢怎麼辦？	不要刻意去想，只要你願意就可以了。	每個人都有權利選擇一個適合的人，不要刻意排斥他人。
我想交一些真心朋友，但感覺真誠不起來，怎麼辦？	可以試著去接納自己，只要你願意敞開心扉，就可以了。	如果你真的愛他，那就去做你自己想做的事情吧。
現在感覺自己壓力越來越大，總是怕自己嫁不出去。	你需要做的就是做好每一件事，不要太在意別人的看法。	你需要做的就是做好你的自我保護工作，這樣才能讓你成長。
隱君子要怎麼克制毒癮？	建議你去醫院做一下心理諮詢，幫助你找到問題的根源。	每個人都有自己的生活方式，只要不影響到你的人際關係就好。
我感覺我的腦子總是不清楚，怎麼了？	建議你去醫院檢查一下，如果生理上沒有問題，可以尋求專業的心理諮詢師幫助。	你可以選擇一個適合你的人。

表 9. 有無主題標籤之生成範例

根據上表所示，在沒有加入主題標籤的模型會偏向產生主題為人際關係或是親密關係的回覆，我們認為是因為資料集內這兩種主題的問題較多，使得模型在沒有主題標籤的控制下，會習慣選擇以這兩種主題的回答方式來產生回覆，使得回覆產生偏差，而在加入主題標籤後，可以降低模型產生上述的偏差，對於問題也能有一定的相關性。如面對生理或心理上的問題，模型會傾向產生建議使用者前往醫院檢查或是尋求幫助的回覆。綜合上述，我們透過分析態度、意圖與主題標籤在不同的設定下會產生的回覆，發現我們透過將標籤設定進行更改，可以影響模型生成的內容，而我們認為是因為透過標籤可以提供模型更多的額外資訊，讓模型在訓練時可以學習到當前標籤下所對應到的



回覆，也能使我們可以藉由標籤來控制模型的生成，最後能夠讓我們藉由設定態度、意圖與主題標籤，來生成正向積極、符合主題且具有指導性的建議來回覆給使用者。

4.6 對話系統串連

本論文也嘗試將對話系統中的分類階段與本論文的生成階段串連起來，分析將兩個元件串連起來後的效果，我們將問題進行多標籤的分類，得到態度、意圖與主題，並將我們得到的標籤與語句送入本論文所訓練的模型中，並將回覆的態度與意圖設定為正向與指導來得到生成的回覆，同時也將態度設定為負向來進行比較，結果如下表 10.所示，其餘的範例請見附錄三。

主題：〈存在議題〉		
問題	態度	意圖
什麼都失敗，不想活了。	負向	告知
回覆		
你可以找到自己想要的生活方式，也許你會有新的發現。	正向	指導
你只是一個缺乏安全感的人，你需要找到自己的興趣愛好，才能讓你成長。	負向	指導

表 10. 單輪對話系統串連範例

同時，我們希望探討以多輪的方式與模型進行互動的效果，因此我們尋找五位大專院校以上學歷的受試者與模型進行多輪對話互動，給予每位使用者三個隨機主題作為開頭並加上特定的使用情境來與模型互動，並在對話結束時詢問受試者使用的效果，評估指標則是從表 1.中選取一致性(Consistency)、參與性(Engagingness)與質量(Quality)，請受試者進行 1~5 分的評分，1 分為最低，5 分為最高，使用情境與指標詢問方式如表 11.所示，從而了解模型在多輪對話下是



否能保持連貫、是否有幫助性與使用者對整體對話的滿意度，而模型回覆同樣將態度與意圖設定為正向與指導，範例如下表 12.所示。

使用情境：目前遇到一個問題或是疑惑，需要有人來幫你解決或給予建議。

	詢問方式
一致性	你認為整段對話的連貫性如何？
品質	你對於整段對話的滿意度如何？
參與性	你認為整段對話對於問題的幫助性如何？

表 11.多輪對話評估指標詢問方式

對話	
	: 我跟朋友吵架了。 (主題：人際、態度：負向、意圖：敘述)
	: 好好溝通，了解彼此的想法，才能找到解決問題的方法。
	: 但我們沒辦法溝通了。 (主題：人際、態度：負向、意圖：敘述)
	: 不要太在意別人的看法，只要你自己堅持下去，就會有所改變。
	: 我覺得他沒辦法和我和好。 (主題：人際、態度：負向、意圖：敘述)
	: 可以試著和他溝通一下，讓他知道你的想法，然後再去做決定。
	: 我覺得是我的個性讓他無法接受。 (主題：親密關係、態度：負向、意圖：敘述)
	: 不要刻意排斥它，這樣才能讓自己變得更好。
	: 我該怎麼改變我的個性？ (主題：性格、態度：中性、意圖：敘述)
	: 可以試著去接納自己，不要太在意別人的看法。
	: 我沒辦法不在意別人的看法。 (主題：人際、態度：中性、意圖：敘述)
	: 不要刻意排斥自己，把自身的想法表達出來，你才能真正的感受到你。

表 12. 多輪對話系統串連範例



多輪對話的評估結果如下表 13.所示

	參與性	一致性	質量
多輪對話評估	3.33	2.83	3.17

表 13. 多輪對話評估結果

根據上表我們可以看出，使用者對於一致性的評分是明顯低於其它兩個指標的，使用者認為單句回答的結果是好的，但是如果連續對答的話很容易重複回答導致整體對話效果不好。我們推測是因為由於中文對話資料集是由單輪的方式來進行訓練，因此在對話時會無法記憶前個問題與回覆，導致回答相似的問題時會產生重複或與前個回覆相互矛盾的回覆，使得一致性的分數較低。而在參與性的部分由於我們將模型的回覆標籤設定為正向與指導，因此針對每個問題還是能有一定的幫助性，使得其分數高於一致性，最後在質量的部分因為受到回覆不一致性的影響，使得使用者在互動時的流暢度與滿意度下降，但對於部分問題又能有一定的幫助性，使得其分數位於中間值。

4.7 小結

本論文分別使用英文對話資料集 DailyDialog 與自建的中文對答資料集 ChinesePsyQA 來進行實驗，嘗試將態度、意圖與主題三種標籤加入到對話文本前，實驗是否能夠提升模型的生成效果，而本論文也使用人工評估與自動評估兩種驗證方法，來驗證研究方法在數值以及與使用者互動上的效果。

而根據 4.3 節的結果所示，本論文所提出的方法在中文與英文資料集上，perplexity(PPL)與準確率上的表現皆超出 baseline 模型，說明將多個標籤置於在文本前能夠提升模型生成下一個 token 的能力，也能提升模型辨別真實語句與混淆語句的能力，我們推測是因為這些標籤提供模型額外的資訊，讓模型能夠根據這

些資訊來鎖定特定的語句，學習到特定語句的表達方式，使得其表現比起未加入標籤的模型還要好。

同時，根據 4.4 節的結果所示，在中文對答資料集上，本論文所提出的方法在相關性、情緒理解與參與性這三個指標上，皆勝過 Transformer 模型與 GPT-2 Base 模型，說明使用者對於加入標籤的模型所生成的回覆，相較於其他模型的回覆有著更高的評價，表示將標籤加入到文本前做訓練，可以提升模型的回覆效果。

而在 4.5 節與 4.6 節，本論文也對訓練好的模型進行不同的實驗與探討，而根據 4.5 節的結果所示，將標籤設定為不同維度可以影響模型所生成的回覆，使得我們可以藉由控制標籤來產生符合需求的回覆，而在 4.6 節中本論文也嘗試將由 Huang 所建構的分類階段與本論文的生成階段串連，根據表 10.與表 12.所示，兩個元件串連起來能夠使分類模型獲得使用者輸入文本的態度、意圖與主題三種維度標籤，並將維度標籤與輸入文本送入生成模型來產回覆，同時我們也徵求受試者以多輪對話的方式與模型互動，而從表 13.的結果可知，模型在一致性上的表現並不是很理想，我們推測原因在於模型的訓練受限於中文資料集的形式，導致模型無法記憶住前面的問答，而這也是未來的研究方向。

第五章、結論



5.1 研究成果

本研究的目的是針對憂鬱症病患或是有需求的人，設計一個對話系統架構，並針對使用者的輸入來產生正向積極的回覆，而在針對過往的研究進行探討後發現，在過往的研究中，許多研究都嘗試將情緒類別加入語言模型中，但對於將多種標籤加入語言模型的研究較為缺乏，也較少在建立端到端(end-to-end)系統的研究。因此，本論文利用擁有強大效果的 GPT-2 預訓練模型，嘗試將態度、意圖與主題三種標籤加入到文本前面，希望藉此來幫助模型產生含有積極態度的回覆。

同時，由於本研究希望可以提供使用者一定的幫助，需要與心理問答相關的對話資料集，但目前中文心理問答的資料集較為缺乏，因此本論文與 Huang(2022) 共同以人工的方式來建構資料集，採用壹心理網站上的心理問題，以貼文的問題與回覆來組成資料，並針對每句問題與回覆進行清理與多種標籤的標記，最後得到一個具有多種標籤的心理對答資料集，供本研究與後續相關研究做使用。

根據研究結果顯示，本論文將態度、意圖與主題等標籤加入到文本中，在人工評估與自動評估上相較於 baseline models 或預訓練模型都有明顯的進步，說明這樣的方式可以幫助語言模型訓練時獲得額外的資訊，讓模型可以學習到更符合標籤的回覆，進而產生更準確的回覆，也讓使用者對於回覆有更高的評價。

而本論文也嘗試將整個對話系統串連起來，使用由 Huang 所建構的分類模型，對使用者所輸入的文本進行分類，得到三個種類的標籤，並將標籤與文本串連後送入到本論文的模型中，藉此產生回覆，而根據表 13.的結果所示，模型在多輪對話的一致性上表現得不是很理想，而在參與性與質量都略高於中間值，說明模型在多輪對話上還有一定的進步空間。



5.2 研究貢獻

在本論文的主要研究貢獻如下：

1. 中文心理問答資料集：在本論文中我們與 Huang 共同建構了一個繁體中文的心理問答資料集，對資料集內的語句進行人工過濾與清理，同時也針對每個語句進行多種標籤的標記，可以提供給往後與中文心理問答相關的研究使用。
2. 多標籤對話生成：本論文嘗試將態度、意圖與主題加入到文本中來訓練模型，希望可以幫助模型產生更適當的回應。根據研究結果得知，我們使用這樣的方法，在人工評估與自動評估上都有著與 baseline model 或預訓練模型更好的表現，說明將多維度的額外資訊加入到文本之中，是可以幫助模型產生出更好的回覆。
3. 端到端對話系統：在本論文與 Huang 所共同建構的對話系統中，我們可以將使用者所輸入的語句進行多個標籤的分類，並藉由分類出來的標籤與文本產生出相對應的語句給使用者，建構出一個端到端（end-to-end）的對話系統。

5.3 研究限制

本研究的主要包含了兩種研究限制，分別為中文對話資料集的數量有限以及中文資料集是屬於單輪問答的性質，以下將分別進行探討：

1. 資料量有限：由於中文心理資料集的缺乏，因此在本研究中所使用的中文心理對答資料集是由我們所建立的，但由於自建資料集需要對文本內容進行清理，來確保資料集內容的品質，也需要對每句進行多個種類的標籤標記，因此需要耗費大量的時間與人力，使得中文資料集的資料數量與其他公開資料集相比稀少許多。這使得我們在訓練模型時，沒辦法讓模型學習到各種的問題與相對應的回覆，這樣會使得當我們在與模型互動時，模型沒辦法有效回答不在訓練資料集的問題，範例如下表 14.模型不良生成範例所示。

主題：〈存在議題〉	
	：到現在還不知道自己要什麼，什麼是適合自己的。
	：適合自己的，是你的一種自我保護。
主題：〈成就〉	
	：快畢業了，論文還沒發，我一直擔心會被拒絕，擔心得睡不著覺。
	：你可以嘗試著去找到自己的興趣愛好。
主題：〈生心理問題〉	
	：假如我首次諮詢，應該先語音還是視訊呢？
	：可以嘗試多聽多看。

表 14. 模型不良生成範例

由上表可得知，若是使用者的問題並非在資料集中，且相似度與資料集中的問題差距過大，會導致模型沒辦法產生適當的回應，因此不斷擴增資料集來解決此限制也是本研究未來的研究方向之一。

- 單輪問答：在中文資料集中，我們是使用壹心理網站中的貼文與回覆，這樣的形式會讓我們在收集資料集時，資料形式會是以一問一答的方式呈現，這使得我們在訓練模型時只能以單輪對話的方式進行，即便在 4.4 節以單輪對話的方式進行的人工評估中顯示出不錯的效果，然而單輪對話會讓模型沒辦法記憶使用者先前所說過的話，這樣會導致模型在與使用者互動時會產生迴圈，降低使用者的使用意願、流暢度與滿意度，如表 13.所示。同時在此資料集中都是以心理對答為主，缺乏基本的對話資料，會使得模型沒辦法應付日常對話。因此將此資料集延伸為多輪對話且含有基本的對話是本研究未來的研究方向之一。

5.4 未來研究方向

本研究的目的是在於解決目前臺灣憂鬱症病患缺乏即時的聊天管道來減緩症狀或紓解負面情緒的問題，設計一個對話系統，並使用多種標籤來與文本結合，讓模

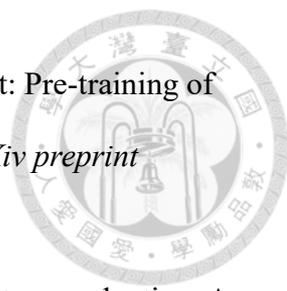
型能夠產生積極的回覆給使用者，以下將敘述未來可以研究的方向來使對話生成的效果更加良好：

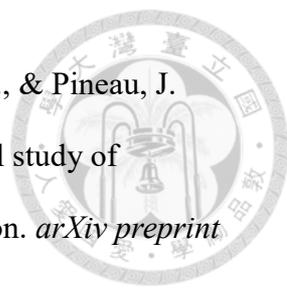
1. 在對話中加入知識：本論文將態度、意圖與主題三種標籤加入到文本前面來進行訓練，相較於以往的模型，不管是在人工評估或是自動評估上都有著一定的進步，說明將額外的資訊加入到文本內可以幫助模型有更好的生成結果。因此，未來可以嘗試將醫療相關知識也加入到文本內，探討是否能夠將知識融入到文本內，讓模型產生更有知識性的回覆。
2. 擴增問答資料數量：如在 5.3 節所提到的，目前本研究所使用的中文心理問答資料集數量不足夠，使得訓練好的模型沒辦法回應所有的問題，因此若是能夠整理並標記更多的對答資料供模型做訓練，將會幫助模型學習到更多樣的問題與其回應方式，探討是否能夠在自動評估與人工評估上有更多的進步。
3. 加入日常對話：目前本研究自建的中文心理對答資料集都是以心理問答為主，沒有日常對話的資料，這樣可能會導致模型沒辦法回應使用者的一般對話，因此未來可以嘗試將此資料集與閒聊資料集做合併，讓模型能夠同時學習到心理問答與日常對話的回應。
4. 多輪對話的設計：目前本研究自建的中文心理對答資料集都是單輪對話，會造成在與模型互動時，沒辦法記住前面說過的話，而造成模型與使用者的對話產生迴圈，因此未來可以嘗試改良資料集，將資料集轉變為多輪對話的形式。

參考文獻



- [1] 宋偲嘉. (2021). 咖啡與憂鬱症之探討.
- [2] 陳文菁. (2021). 抗憂鬱藥物使用人數. 衛生福利部中央健康保險署.
<https://data.gov.tw/dataset/146577>
- [3] 李昭慶. (2000). 憂鬱症與運動. *大專體育*(50), 82-88.
- [4] 張家銘. (2020). 台灣憂鬱症就醫現狀與問題. 社團法人臺灣憂鬱症防治協會.
<http://www.depression.org.tw/knowledge/info.asp?/71.html>
- [5] A+醫學百科. (2011). 支持療法. A+醫學百科. <http://cht.a-hospital.com/w/%E6%94%AF%E6%8C%81%E7%96%97%E6%B3%95Harilal,N.>
- [6] Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., & Lu, Y. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
- [7] Asghar, N., Poupart, P., Hoey, J., Jiang, X., & Mou, L. (2018). Affective neural response generation. European Conference on Information Retrieval,
- [8] Cahn, J. (2017). CHATBOT: Architecture, design, & development. *University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science*.
- [9] Carr, A. (2008). Depression in young people: Description, assessment and evidence-based treatment. *Developmental Neurorehabilitation*, 11(1), 3-15.
- [10] Chen, Q., Zhuo, Z., & Wang, W. (2019). Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.
- [11] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

- 
- [12] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [13] Finch, S. E., & Choi, J. D. (2020). Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. *arXiv preprint arXiv:2006.06110*.
- [14] Harilal, N., Shah, R., Sharma, S., & Bhutani, V. (2020). CARO: an empathetic health conversational chatbot for people with major depression. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD* (pp. 349-350).
- [15] Huang, Y.-H., Lee, S.-R., Ma, M.-Y., Chen, Y.-H., Yu, Y.-W., & Chen, Y.-S. (2019). EmotionX-IDEA: Emotion BERT--an Affectional Model for Conversation. *arXiv preprint arXiv:1908.06264*.
- [16] Huang, Y. Y. (2022). *Dialogue Analysis of A Chatbot for Patients with Depression*
- [17] Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1), S63-S63.
- [18] Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., & Dolan, B. (2016). A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.
- [19] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.
- [20] Lin, Z., Madotto, A., Shin, J., Xu, P., & Fung, P. (2019). Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.
- [21] Lin, Z., Xu, P., Winata, G. I., Siddique, F. B., Liu, Z., Shin, J., & Fung, P. (2020). Caire: An end-to-end empathetic chatbot. *Proceedings of the AAAI Conference on Artificial Intelligence*,

- 
- [22] Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.
- [23] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics,
- [24] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- [25] Ritter, A., Cherry, C., & Dolan, B. (2011). Data-driven response generation in social media. Empirical Methods in Natural Language Processing (EMNLP),
- [26] Su, H., Jhan, J.-H., Sun, F.-y., Sahay, S., & Lee, H.-y. (2021). Put Chatbot into Its Interlocutor's Shoes: New Framework to Learn Chatbot Responding with Intention. *arXiv preprint arXiv:2103.16429*.
- [27] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.
- [28] Thase, M. E., Greenhouse, J. B., Frank, E., Reynolds, C. F., Pilkonis, P. A., Hurley, K., Grochocinski, V., & Kupfer, D. J. (1997). Treatment of major depression with psychotherapy or psychotherapy-pharmacotherapy combinations. *Archives of general psychiatry*, 54(11), 1009-1015.
- [29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- [30] Wolf, T., Sanh, V., Chaumond, J., & Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv*

preprint arXiv:1901.08149.

- 
- [31] Zandie, R., & Mahoor, M. H. (2020). Empransfo: A multi-head transformer architecture for creating empathetic dialog systems. The Thirty-Third International Flairs Conference,
- [32] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.
- [33] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.
- [34] Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1), 53-93.

附錄



附錄一、壹心理網站貼文範例

一、貼文問題



做事力求负责，但人际关系依然不好，如何改善？

31个回答

07-29 580 阅读

这是我对自己的分析，求指导：

- 1.期望值管理不到位，总是会在开始的时候做非常多的事情，这样在后续没什么进步空间，且对于人际关系容错率低，一旦犯了错误就很难被人原谅。
- 2.做的工作没有创造性，简单来说就是身体上的勤奋思想上的懒惰，看似做的事情不少但多是重复性且含金量低的工作，没有创造性的工作带来的集体价值有限，故人际关系也很难通过工作发展起来。
- 3.人格特质缺乏特色，历来我收到的评价是正直肯干，这项特质可以提供工作价值，但难以提供情绪价值。目前我的身份还是处于做基础工作和学习的阶段，相比于同年员工我具有一定的工作能力，可以做一部分他们做不了的事，但在高年员工面前这些能力就和吃饭喝水一样，提供的工作价值没什么意义。这就造成了我在他人眼里没有明显的个人特色，虽然做事负责，但留下的印象还不如一个油嘴滑舌的老油条。

楼主悬赏 ¥30.00

展开

人际

朋友

人际边界

同事

二、貼文回覆



坐看云起

优质

累计帮助了71人



题主，当下好！

谢谢你的敞开，让我有机会跟你链接。看完你的描述，我了解到你对自己还是有一些认识，看到了自己的长处和短板。我觉得你可以把自己的长处发挥到极致，对于人际关系这个短板，不知道你想达到什么高度：人人都喜欢你？这个谁也做不到。有三五知己就够了吧。

每个人都会喜欢接近高能量的人，那你想提升自己的人际关系，我觉得可以从提升自己开始。比如你说的工作能力上，虽然一般，可以尝试努力做到更快更好，或者悄悄学习提升自己的其它能力，或者提示自己的内在，让自己能滋养别人，给到别人帮助，这样也可以提升自己的价值，提高人际关系。

方式很多，最主要的是向内探索，改变自己内在，把注意力放在自己内在提升上，内在富足，外在自然丰盈。

祝福你！

附錄二、各模型生成比較



問題	Transformer	GPT-2 Base	Proposed Model
怎麼樣建立自信？	你要做的是你自己，你要做的自己，你就不會發現。	建議你找到自己的興趣愛好，然後慢慢的去嘗試一下。	建議你去做一些自己喜歡的事情，這樣才能讓你成長起來。
忘不了傷心的事情，怎麼辦？	你需要做的是做自己，讓自己放鬆。	忘記了就好了，不要刻意去傷害別人。	可以試著找到自己的內心真正的想法，也許你會有新的發現。
面對個性奇怪的室友怎麼辦？	你可以嘗試和你的朋友相處，把你的想法告訴她。	你可以試著去找到自己的真正屬於你的人。	你可以試著去接受他人的不完美，讓彼此的關係更和諧。
失戀了，每天都過得好難受啊。	你可以嘗試一個人的生活。	你需要找到自己的興趣愛好，然後再去尋找屬於你的人生吧。	好好愛自己，不要因為愛情而影響到你的生活。
我很在意別人的看法，怎麼辦？	你可以嘗試在自己的生活中找到自己的目標，你可以做自己。	你可以試著去接納自己，不要刻意去想那些事情，這樣你才能更好地去面對它們。	你可以試著去接受自己的想法和感受，也許你會有新的發現。

附錄三、對話系統串連結果



一、成就

主題：〈成就〉		
問題	態度	意圖
現在是公務員，薪資環境都不錯，但我還是想嘗試別的工作。	中立	告知
回覆		
你可以選擇一個專業的職業生涯規劃，找到自己的興趣愛好。	正向	指導
如果你不想改變，那就去做一些自己喜歡的事情吧。	負向	指導

二、親密關係

主題：〈親密關係〉		
問題	態度	意圖
被告白了，但我只想做普通朋友，我應該怎樣做，怕拒絕他會破壞關係？	中立	詢問
回覆		
不要刻意排斥他，讓他知道你的想法，試著去解決他的問題。	正向	指導
你的問題在於你自己的感受，而不是別人的想法。	負向	指導

三、生心理問題

主題：〈生心理問題〉		
問題	態度	意圖
怎麼才能知道自己是否有心理疾病，比如說憂鬱症？	中立	詢問
回覆		
建議你去醫院精神科檢查一下，如果發現有問題，可以尋求專業人士幫助。	正向	指導
自信心不夠強，建議你去醫院精神科檢查一下。	負向	指導



四、人際

主題：〈人際〉

問題	態度	意圖
如何與朋友交往？	中立	詢問
回覆		
多多溝通，讓彼此瞭解，瞭解內心的想法和感受。	正向	指導
如果你想改變，那就需要你自己去做。	負向	指導

五、家庭

主題：〈家庭〉

問題	態度	意圖
家裡出了問題，不想回家了，怎麼辦？	負向	詢問
回覆		
你可以試著和家人溝通一下，讓他們了解你的想法。	正向	指導
家人的心裡有很多不安全感，可以通過諮詢解決。	負向	指導

六、性格

主題：〈性格〉

問題	態度	意圖
遇到每件事情，總是怕事情做不好，怎麼辦？	負向	詢問
回覆		
你可以試著去了解自己的內心，不要太在意別人的看法，這樣你才能更好地接受你的想法。	正向	指導
遇不了就不要去做，不必刻意去想太多。	負向	指導

七、



對話

：我失戀了，我該怎麼辦？
(主題：親密關係、態度：負向、意圖：詢問)

：好好愛自己，不要因為這件事而影響到你。

：但我放不下她。
(主題：親密關係、態度：負向、意圖：敘述)

：好好珍惜自己的生活，讓她幸福快樂。

：我覺得我沒辦法活下去了。
(主題：存在議題、態度：負向、意圖：敘述)

：試著去找到自己的興趣愛好，你會發現你的優點。

：我不能沒有她。
(主題：親密關係、態度：負向、意圖：敘述)

：好好和她溝通，讓她知道你的想法。

：我想要挽回她。
(主題：性格、態度：負向、意圖：敘述)

：祝福她的幸福，讓她能夠擁有更好的生活。

八、

對話

：家人給的壓力好大，有時候要承受他們的情緒好累。
(主題：童年/家庭、態度：負向、意圖：告知)

：每個人都有自己的生活方式，只要你願意承擔你的責任，就可以了。

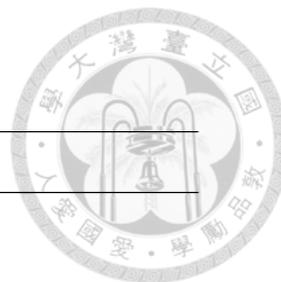
：他們希望我回去家裡幫忙，但我有自己想要做的工作。
(主題：成就、態度：中性、意圖：告知)

：如果你覺得你的問題沒有解決，那就去尋找一個合適的機會吧。

：但我不知道要怎麼和家裡的人溝通。
(主題：童年/家庭、態度：中性、意圖：告知)

：不要太在意別人的看法，只要你願意去做就行了。

九、



對話

：活著有什麼意義？
(主題：存在議題、態度：中性、意圖：詢問)

：活在當下，你要做的就是讓自己變得更好，這樣才能讓你成長。

：我連興趣都沒有了，生活好無趣。
(主題：成就、態度：負向、意圖：告知)

：你可以試著找到自己喜歡的事情，然後再去做。

：我感覺我沒有朋友，好孤獨。
(主題：人際、態度：負向、意圖：告知)

：每個人都有自己的生活方式，只要你願意去做就可以了。

十、

對話

：我覺得生活很無趣，很煩躁該怎麼辦？
(主題：性格、態度：負向、意圖：詢問)

：試著找到自己的興趣愛好，去做一些有意義的事。

：但我找不到我的興趣愛好。
(主題：性格、態度：中性、意圖：敘述)

：可以試著找到自己喜歡的事情，然後慢慢的去嘗試吧。

：我覺得我沒辦法完成任何的事。
(主題：性格、態度：負向、意圖：敘述)

：可以試著去做一些事情，讓自己變得更有力量。

：但我沒有自信。
(主題：性格、態度：負向、意圖：敘述)

：試著去做自己擅長的事，你會發現你的優點。

：我覺得我身上只有缺點，沒有優點。
(主題：性格、態度：負向、意圖：敘述)

：不要刻意排斥自己，這樣才能讓你成長。

十一、



對話

：朋友找我出去玩但我不想去，要怎麼委婉的拒絕他？

(主題：人際、態度：中性、意圖：詢問)

：你可以選擇自己喜歡的，不要刻意排斥他，這樣才能讓你成長起來。

：但暑假好無聊。

(主題：成就、態度：中性、意圖：告知)

：每個人都有自己的生活方式，只要你願意去做就可以了。

：可是考完試之後就沒有動力跟目標了。

(主題：成就、態度：負向、意圖：告知)

：你可以選擇一個適合自己的工作，如果你夠努力，那麼你就不會後悔的。

：看到別人的生活好精彩覺得自己好頹廢。

(主題：性格、態度：負向、意圖：告知)

：人生的美好不是別人決定，而是你的自我價值觀。

：那我要怎麼提升自己？

(主題：性格、態度：中性、意圖：詢問)

：你需要做的就是做好你的自我。

：那我要怎麼找到目標，我有點迷失方向了。

(主題：成就、態度：負向、意圖：詢問)

：每個人都有自己的興趣愛好，只要你願意去做就可以了，相信你會成功的。

十二、

對話

：準備要換一個地方讀書，但內心不想要認識新的朋友，這樣好嗎？

(主題：人際、態度：中性、意圖：詢問)

：每個人都有自己的優點，只要你願意，就不會在意別人的看法。

：如果之前的朋友也都交到新朋友了話，自己就會有被拋棄的感覺。

(主題：人際、態度：負向、意圖：告知)

：你可以試著去接納他，讓他知道你的想法，然後再去做決定。

：但我一直封閉自己不跟別人來往，這樣是不是不太好。

(主題：性格、態度：負向、意圖：詢問)

：每個人都有不同的生活方式，只要你願意去做就可以了，相信你會成功的。

：那你覺得我應該要怎麼改變？

(主題：性格、態度：中性、意圖：詢問)

：你需要做的就是讓自己變得更好，不要太在意別人的看法。



對話

：男朋友好忙都不沒時間陪我。
(主題：親密關係、態度：負向、意圖：告知)

：好好溝通一下，讓他知道你的想法。

：可是他都在做研究，不想去打擾他。
(主題：親密關係、態度：中性、意圖：告知)

：如果你覺得自己不夠好，就去做一些事情。

：所以我應該把時間放在自己身上嗎？
(主題：性格、態度：中性、意圖：詢問)

：你可以試著去找一些事情做，不要想太多。

：那我有什麼辦法可以分擔他的壓力嗎？
(主題：親密關係、態度：中性、意圖：詢問)

：你需要做的就是做好你自己的事情，不要刻意排斥他。

：但我希望他偶爾可以陪我，讓我們有可以休息的時間。
(主題：親密關係、態度：正向、意圖：告知)

：如果你真的喜歡他，那就儘量不要放棄他。