陪聊型聊天機器人的頭像與回覆數量對使用體驗的影響

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

摘要

本研究探討 AI 陪聊型聊天機器人的不同設計對使用體驗帶來的影響。本研究採用 2x2 二因子重複測量變異數分析，比較與匿名頭像、動漫頭像聊天，以及每次發言收到一則回覆或五則回覆，所帶來的使用體驗分數差異。

在找到 98 名受試者進行評估後，發現比起匿名頭像或僅收到一則回覆，受試者與動漫頭像或有多則回覆的聊天機器人聊天，會有更好的使用體驗分數，其中提供多則回覆的分數提升較大，因為受試者傾向注意適合回應並忽略不適合回應。並且，如果同時提供動漫頭像與多則回覆，能得到最高的使用體驗得分。

希望未來在設計陪聊型聊天機器人時，能參考本研究的結論，將擬人頭像與多回應納入設計考量。

作者關鍵字

聊天機器人、虛擬頭像、回覆數量、使用體驗

CSS Concepts
• Human-centered computing → Human computer interaction (HCI) → HCI design and evaluation methods → Laboratory experiments;

研究動機

聊天機器人已經普及在我們的日常生活中。被用在各種領域如客服[1]、虛擬助理[2]、陪聊。尤其是陪聊，這是一個非常有潛力的領域，因為能滿足人要被陪伴的需求。先前北一女校慶就有同學擺攤「陪聊 10 分鐘 45 元」引發廣泛討論[3]。而聊天機器人當然也能被用在陪聊。

現在已有許多相關服務及研究在各國風行，例如美國的談話治療機器人 Woebot[4]、韓國的人氣聊天機器人 Luda[5] 與我國中研院的詞庫小妍[6]。如何打造一個能帶來美好使用體驗的陪聊平台，是很重要的議題。

在之前楊德倫與曾元顯的研究中[7]，其提出 GPT-2、BERT 雙模型聊天機器人機制，即收到訊息後會先由 GPT-2 生成適合的回應，再藉由 BERT 進行回應連貫性排序，打造出一新版聊天機器人引擎。該引擎在學習 2019 年日本 NTCIR 中文情緒對話生成（CECG）評比任務所提供約 170 萬則語料後，可以成功的對各種對話輸入，加上情緒的調整後，輸出適合的回應。

楊德倫與曾元顯[7]也將引擎打造出一展示用網頁如圖一，雖然介面陽春，但可列出聊天機器人引擎所生成的多筆回應，與連貫性評估的結果。

圖一：聊天引擎展示介面

本篇文章中我們將此聊天引擎與 Telegram 聊天機器人結合，打造出一可提供多重回覆的機器人平台，並使用動漫人物頭貼當作機器人的頭貼，來改善其使用體驗，如圖二。

CHI 2020, April 25–30, 2020, Honolulu, HI, USA.
 ACM ISBN 978-1-4503-6708-0/20/04...$15.00.
DOI: https://doi.org/10.1145/3313831.XXXXXXX
*update the above block and DOI per your rightsreview confirmation (provided after acceptance)

圖二：改造後的動漫聊天室

此聊天服務引發我們周邊網友的強烈回響，因此我們決定以實際的實驗探討在聊天機器人的設計中，使用頭像（avatar）類型與提供多則回應，對於聊天體驗是否有正面或負面的影響？

文獻回顧

深度學習 NLP 與自然語言生成

要打造一個陪聊型聊天機器人，靈活應對各種聊天輸入，可結合深度學習 NLP 技術與聊天機器人平台的技術。深度學習模型[8]是指隱藏層數多於一層的類神經網路模型，透過調整模型中的參數與權重，來學習輸入輸出的答案。要進行自然語言生成，需對深度學習模型進行大量語言文本的輸入與輸出訓練，以在其中建立語言模型，進而根據先前輸入的文字，預測要生成下一個字的條件機率，數學表示如下：

$$P(w(t) \mid w(t-1), \ldots, w(2), w(1))$$

亦即根據前 t-1 個字詞，預測下一個字詞的機率，而此預測的機率應該要符合我們人類對語言知識的預期。例如，在模型訓練過包含「聊天機器人」五個字的資料，並且未訓練過包含「聊天機器貓」的資料時，當我們給予模型輸入前四個字「聊天機器」時，生成的條件機率應該符合：

$$P(\text{人} \mid \text{聊}, \text{天}, \text{機}, \text{器}) > P(\text{貓} \mid \text{聊}, \text{天}, \text{機}, \text{器})$$

接著，模型就會根據生成結果，選擇較高機率的 P(人｜聊, 天, 機, 器）預測「聊天機器」下一字為「人」。傳統條件機率計算方法，是離散空間計算次數的統計方法（count base)，對訓練語料中沒有出現過的輸入難以做出正確的預測。如果資料中沒有「聊天機器貓」則模型的 P(貓｜聊, 天, 機, 器）一定是 0，除非透過平滑方式處理（Smoothing）。但如藉由[9]神經機率語言模型，把文字投影到連續空間上，得出條件機率值，使沒有出現在語料中的類似字詞，其條件機率預估不會太差。舉例來說，如果神經機率語言模型學過「聊天機器人」與「跑跳機器貓」，則雖然在輸入「聊天機器」時，P(人｜聊, 天, 機, 器）> P(貓｜聊, 天, 機, 器)，但後者 P(貓｜聊, 天, 機, 器)也不會是 0，因為模型學過類似的話語「跑跳機器貓」，知道「機器」後面除接「人」也可以接「貓」，只是因為更前面兩個字是「聊天」而不是「跑跳」，所以機率較低。

藉由生成過程中的隨機處理，讓語言模型從條件機率較高的結果中隨機進行挑選，模型就能對同樣的輸入，做出不同的預測，達到生成語言的多樣性，如輸入「聊天機器」，模型可能預測下一個字為「人」、「貓」、「狗」。本研究打造的聊天機器人，使用的語言生成引擎 GPT-2[10]，即支援這種設定。

聊天機器人

聊天機器人的概念，最初源自艾倫·圖靈（Alan Turing）在 1950 年提出的圖靈測試[11]，即人與機器僅透過文字進行溝通。現在聊天機器人已經普及在我們的日常生活中，被用在各種領域如客服[1]、虛擬助理[2]、陪聊。

現有的「陪聊」聊天機器人

現在的陪聊聊天機器人服務，包含美國的談話治療機器人 Woebot[4]、韓國的人氣聊天機器人 Luda[5] 與我國中研院的詞庫小妍[6]。他們標榜能像真人般自然地對話，獲得許多人的青睞。以 Luda 為例，新聞[5]指出其推出不到半個月就吸引超過 40 萬用戶，累計對話高達 9 千萬次。

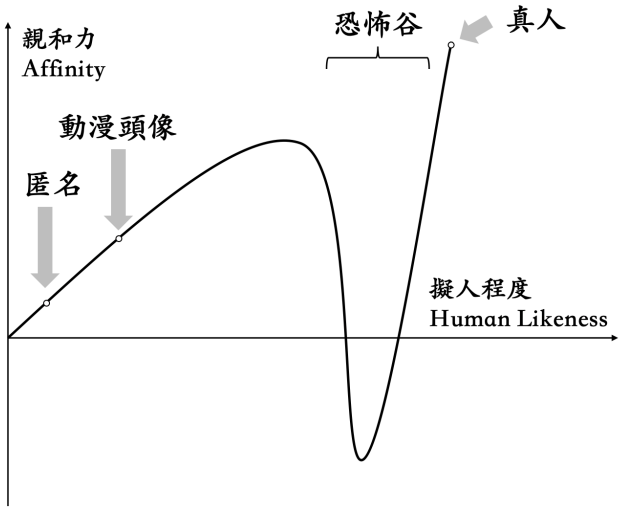
然而，研究[4]指出，使用者在聊天時發現陪聊型聊天機器人會有許多不適合的回應，包含重複同樣的話語、前後對話不連貫、自說自話等，進而導致使用體驗不佳。

由此可見陪聊型聊天機器人依舊有許多改進的空間。本研究提出的設計改進，就為改善前述問題。

研究理論基礎與假設

聊天對象的頭像與聊天體驗的關係？

在聊天機器人上設置頭像，將能使聊天機器人更為人性化。根據恐怖谷理論（The Uncanny Valley）[12]，當機器人較人性化（Human Likeness），但尚未達到與人類非常相似時，使用者會認為其較有親和力（Affinity），此如圖三所示。



圖三：恐怖谷理論在聊天機器人擬人程度與親和力的關係

而較高的親和力，根據品牌行銷領域的研究結果[13]，會帶來較高的評價與使用率。因此可推論在聊天機器人上也有類似的關係。因此提出以下假說：

H1: 使用者認為動漫人物頭像的聊天對象，比起匿名頭像更有親和力。

H2: 使用動漫頭像的聊天對象，比起匿名頭像得到更好的使用體驗評價。

聊天時收到的回覆數量，與聊天體驗的關係？

過去研究[14]提出「雞尾酒會效應（Cocktail Party Effect）」，描述人的選擇性聽力能力。指人在吵雜的環境與他人互動時，會將注意力集中在談話對象所說出的內容，並自動忽略周圍的其他對話或噪音，因此可以在吵雜的環境（如雞尾酒會）聊天。儘管如此，如果周圍他人的聊天出現關鍵字（如自己的名字）或感興趣的內容，我們會轉而注意這些內容。

我們推論類似的現象在使用者與多個聊天對象互動時也會發生。即使用者在收到多個回應之時，較傾向注意適合回應，且忽略不適合回應之情況。適合回應包含但不限於：從預期聊天對象發出的、與發言相關的、感興趣的、合理的回應。

倘若前述推論成立，則對於使用者的輸入，聊天模型生成多個回應所帶來的聊天體驗，會比單一回應好。單純以理性分析，這是個機率的問題：假設針對使用者的輸入 x，模型生成適合回應的機率為：

$$p(x)\text{（在此假設為 0.4）}$$

則模型無法生成適合回應的機率就是：

$$1 - p(x)\text{（如上假設則為 0.6）}$$

若對於使用者的輸入，一次生成並回傳五個不同的回應，模型無法生成任何適合回應的機率將為：

$$(1 - p(x))^5 \text{（如上假設將為 } 0.6^5 = 0.0776\text{）}$$

相較於單一回應的 1-p(x)，使用者收不到任何適合回應的機率將大幅降低（從 0.6 變為 0.0776），因而可能提高使用體驗。換句話說，回應數量越多，模型亂槍打鳥生成至少一適合回應機率越高（從 0.4 變成 0.9224）。由此提出以下假說：

H3: 使用者傾向注意適合的回應。

H4: 使用者傾向忽略不適合的回應。

H5: 多回應情境比單回應得到更高的使用體驗評價。

交互作用與加成效果的影響

因為跟一個人或一群人聊天，或跟一群有頭像的人聊天或跟一群匿名頭像的人聊天，會是不同的體驗。本文的研究也想知道，兩條件之間是否會互相影響？即頭像與回應數間，會不會有交互作用產生？由此提出以下假說：

H6: 頭像類型與回應數量之間有交互作用。

研究執行方法

受試者樣本描述與設計

本研究透過網路進行便利抽樣，找到 98 名受試者，年齡介於 19 至 50 歲之間（M=22.3, SD=3.89）。本研究使用 2（單一回應 vs 五則回應）x 2（匿名頭像 vs 動漫頭像）二因子組內重複測量模型，每位受試者均需經歷四種實驗條件，其順序為隨機分派。

實驗執行流程

受試者操作自備的電腦，透過網路登入平台後，先進行 10 次聊天暖身，確保熟悉平台操作。之後即開始進行四輪順序隨機的聊天實驗。每輪流程如下所示：

1. 先請受試者根據聊天對象頭貼，給予好感度評價。
2. 進行十次對話，受試者每發出一則訊息算一次。
3. 再做一次好感度評價後測。
4. 填寫使用體驗量表評估該輪的使用體驗。

四輪對話結束後，受試者會填寫整合性問卷並收集質性回饋，然後查看實驗的事後解釋（Debrief）。

測量方法

聊天機器人使用體驗（H2, H5, H6）

我們將聊天機器人易用度量表（The Chatbot Usability Questionnaire, CUQ）[15]翻譯成中文（請見附件），用以量測每一輪的聊天體驗。該量表是專門用於量測聊天機器人使用體驗，使用者須針對每一則陳述（如：這個聊天對象懂我）選擇 1 分（非常不同意）至 5 分（非常同意）。根據 CUQ 作者的報告[15]，其得分與系統易用性量表（System Usability Scale, SUS）[16]及用戶體驗調查表（User Experience Questionnaire, UEQ）[17]的得分有顯著正相關。

聊天對象的親和力（H1）

關於親和力，我們使用自我涵蓋他人量表（Inclusion of Other in the Self, IOS）[18]進行測量，該量表可引導使用者評估與目標的親近度（1 至 7 分）。同時，我們也讓使用者對聊天對象給予好感度分數（1 – 100 分）。

適合與不適合回應的注意程度（H3, H4）

在四輪聊天後的整合性問卷中進行自陳報告性質測量，採用五點量表，使用者須針對下列陳述選擇 1 分（非常不同意）至 5 分（非常同意），題目表列如下：

編號	題目
1	剛剛的對話中，我比較常注意適合的回應
2	剛剛的對話中，我比較常注意不適合的回應
3	剛剛的對話中，我比較常忽略適合的回應
4	剛剛的對話中，我比較常忽略不適合的回應

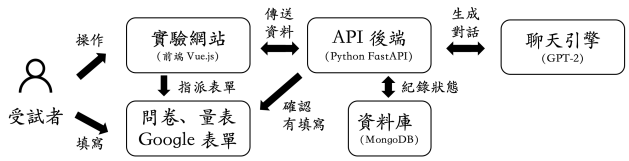
表一：評估注意程度使用的測量問題

在計算時，對 1, 2 題、3, 4 題，分別使用相依樣本 t 檢定，比較使用者注意程度的差異。

實驗平台建置

實驗平台建置在網路上，分為前端介面與後端伺服器。前端採用 Vue.js 框架進行建置，並使用 Bootstrap 及 Chat UI [19]模板打造模擬通訊軟體的聊天室環境，其中動漫頭像聊天對象皆選自人氣漫畫《進擊的巨人》，截圖請見圖四。使用者填寫量表或問卷時，會引導受試者至 Google 表單填寫，再透過後端呼叫 Google API 確認是否已經填寫完成。後端部分使用 Python FastAPI 搭配 MongoDB 作為資料處理中樞，並負責轉介呼叫聊天文字生成引擎[7]。架構圖請見圖五，程式碼已開源，請見 GitHub（審查階段連結匿名）。

圖四：實驗平台截圖，此為多回覆-動漫頭貼的介面



圖五：實驗平台系統架構圖

使用的資料分析方式

資料分析軟體使用 SPSS 23.0 進行。

獨立樣本 T 檢定比較親和力（H1）

關於受試者對不同頭像的親和力，會將其分數採用相等變異數的獨立樣本雙尾 t 檢定比較整體差異。因在頭像控制組受試者聊天對象為匿名頭像，每位受試者共測量兩次（單一匿名回應 vs 五則匿名回應），頭像實驗組則會對每個動漫頭像分別進行親和力測量，共測量六次（單一頭像回應 vs 五則頭像回應），所以是不平衡的數據集，無法直接使用相依樣本 t 檢定。

成對樣本 T 檢定比較自陳報告（H3, H4）

對適合與不適合回應的注意程度，會對受試者報告分數使用相依樣本 t 檢定進行評估（詳見表一）。

雙因子 ANOVA 比較交互作用與主要效果（H2, H5, H6）

使用 2x2 重複測量分析，查看兩種操弄條件之間，是否有交互作用，並分析主要效果。

資料分析結果

親和力程度分析結果

從表二可得知，受試者對有頭像的聊天對象的 IOS 親近度分數（M=3.09, N=588, SD=1.78），與僅有匿名頭像的聊天對象的親近度分數（M=1.66, N=196, SD=1.06），呈現顯著差異（t(782) = 10.639, p < .001, d = 0.76）。且在好感度分數上，動漫頭像好感度得分（M=51.12, N=588, SD=26.9）與匿名頭像（M=30.88, N=196, SD=24.53）也呈顯著差異（t(782) = 9.319, p < .001, d = 0.67）。意味使用者認為動漫人物頭像的聊天對象，比起匿名頭像更有親和力（H1 接受）。

	動漫頭像		匿名頭像		t(782)	p	Cohen's d
	M	SD	M	SD			
IOS 分數	3.09	1.775	1.66	1.058	10.639***	0.000	0.76
好感度	51.12	26.90	30.88	24.53	9.319***	0.000	0.67

表二：受試者對不同頭像的親近度、好感度分數及分析結果

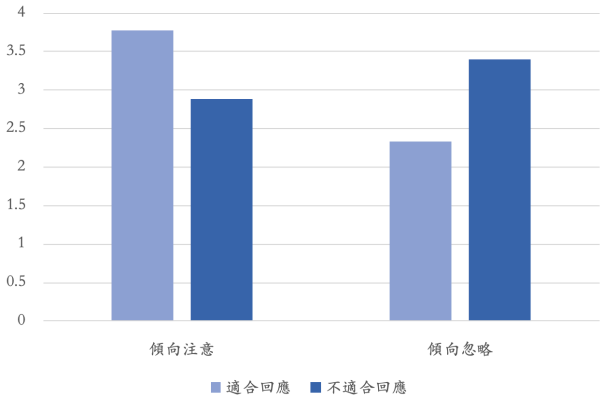
自陳報告分析結果

表三及圖六為受試者關於回應注意情況的自陳報告分數及分析結果。由表三可知，受試者傾向注意適合的回應的程度（M=3.78, N=98, SD=1.07），顯著高於傾向注意不適合的回應的程度（M=2.88, N=98, SD=1.14）（t(97) = 4.521, p < .001, d = 0.457, 95% C.I. [0.504, 1.292]）。意味受試者傾向注意適合的回應（H3 接受）。

另外，由表三可知，受試者傾向忽略適合的回應的程度（M=2.33, N=98, SD=0.95），顯著低於傾向忽略不適合的回應的程度（M=3.4, N=98, SD=1.18）（t(97)= -5.890, p < .001, d = 0.595, 95% C.I. [-1.43, -0.71]）。意味受試者傾向忽略不適合的回應（H4 接受）。

	適合回應		不適合回應		t(97)	p	Cohen's d
	M	SD	M	SD			
傾向注意	3.78	1.070	2.88	1.142	4.521***	0.000	0.457
傾向忽略	2.33	0.950	3.40	1.182	-5.890***	0.000	0.595

表三：回應注意情況的自陳報告數據及分析結果



圖六：回應注意回應情況的數據長條圖

雙因子 ANOVA 分析結果

參與者在不同回覆數量及頭像的使用體驗分數之平均數與標準差如表四及圖七所示。變異數分析結果在表五，結果顯示不同頭像以及回覆數量在使用體驗分數呈現的交互作用未達顯著（$F(1, 97) = 2.13$, $p = .148$, $\eta^2 = .021$）（H6 無法接受）。

組別	單一回覆		多則回覆	
	平均數	標準差	平均數	標準差
匿名頭像	29.15	20.40	32.22	18.90
動漫頭像	30.45	19.04	37.97	20.53

表四：不同回覆數量及頭像的使用體驗分數

變異來源	Sum of Squares	Df	Mean Square	F	p	η^2	η^2 90% CI [LL, UL]
截距	412771.30	1	412771.30	522.537	.000		
回應數	2754.211	1	2754.21	10.089**	.002	.094	[.022, .192]
頭像	1222.254	1	1222.25	4.574*	.035	.045	[.002, .127]
回應數 x 頭像	484.800	1	484.80	2.126	.148	.021	[.000, .089]
誤差	76623.84	97	789.94				

表五：回應數量與頭像的變異數分析結果

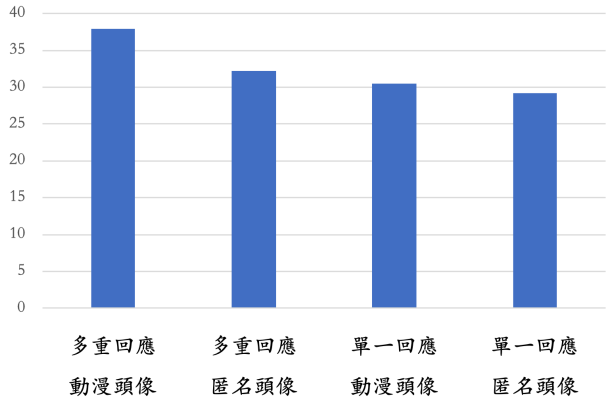


圖 7：不同操弄條件的分數比較

比較主要效果

主要效果的詳細資訊請參考表五。不同回覆數量的主要效果達到顯著($F(1, 97) = 10.09$, $p = .002$, $\eta^2 = .094$)。此結果代表，比起單一回覆，提供多重回應，會提升聊天的使用體驗（H5 接受）。

不同頭像的主要效果亦達到顯著($F(1, 97) = 4.57$, $p = .035$, $\eta^2 = .045$)。此結果代表，比起僅使用匿名頭像，使用動漫頭像，會提升聊天的使用體驗（H2 接受）。

結論與討論

因本研究有六個假說，在此列出所有假說以及其統計分析結果（被接受或拒絕）以利檢視：

編號	假說內容	分析結果
H1	使用者認為動漫人物頭像的聊天對象，比起匿名頭像更有親和力。	接受
H2	使用動漫頭像的聊天對象，比起匿名頭像得到更好的使用體驗評價。	接受
H3	使用者傾向注意適合的回應。	接受
H4	使用者傾向忽略不適合的回應。	接受
H5	多回應情境比單回應得到更高的使用體驗評價。	接受
H6	頭像類型與回應數量之間有交互作用。	無法接受

表六：本實驗的假說及其分析結果

結論

此結果指出下列內容：

1. 在設計陪聊型聊天機器人時，如果使用動漫頭像，會比僅提供匿名頭像帶來更高的好感度及親和力，且帶來更好的使用體驗。
2. 在設計陪聊型聊天機器人時，如果對使用者的發話提供多則回應，模擬其在多人群組聊天的情況，會帶來更好的使用體驗。因為使用者較傾向注意適合的回應、忽略不適合的回應。所以有多則回應較易滿足使用者，而可降低系統偶有回應不好的效應。
3. 比起更改頭像，提供多則回應會有更顯著的使用體驗改善效果。但最好的情況是兩個同時做，因為兩者並沒有交互作用，而是會有加成的效果。

現在多數陪聊型聊天機器人都只提供一則回應，有些甚至沒有設定擬人頭像。如果能新增擬人頭像，並提供多則回應，將能改善使用者的聊天體驗。因此希望未來在設計陪聊型聊天機器人時，能參考本研究的結論，將動漫擬人頭像與多則回應納入設計考量。

討論

為什麼多人且有動漫頭像有最好的使用體驗？

有些加分條件，僅在多人動漫頭像情境發生。

在事後訪談時，受試者指出，多則回覆動漫頭像，就像在群組裡跟一群人聊天，大多時候大家講各自的話題，然後會有一兩人來回應自己發的訊息，跟現實世界很像。

也有受試者表示因為是使用動漫頭像，會將自己對他們的認識帶入聊天中，進而腦補[自行解釋]他們說這句話背後的含義（如在回覆其他人的發言），因此有些對話就算乍看不適合，搭配其他人的發言多看幾遍，也慢慢覺得有點道理。

為什麼多回應的改善效果比改為動漫頭像高？

結果顯示，比起更改頭像，提供更多回應會有更顯著的使用體驗改善效果。經由事後訪談，受試者表示因為單一回應如果出現不適合回應，如女性動漫人物頭貼卻回應自己是男生，或產出與對話內容矛盾的回應，就會馬上影響使用體驗。但如果在多回應，受試者表示自己會去注意適合回應，挑適合的回應繼續聊天，就算有自說自話（不適合的回應），也會認為是在跟其他人說話，就不會因為出現不適合回應，影響使用體驗。

人氣會影響動漫頭貼的使用體驗評分嗎？

或許有人考量在動漫頭像回應的部分，是否因角色人氣或好感度等外衍變項導致聊天體驗評分降低或升高？經過獨立樣本 t 檢定分析，單一回應動漫頭像角色的好感度得分（M=55.09, SD=26.59, N=98），與多則回應動漫頭像角色（M=50.32, SD=26.93, N=490）之間並沒有顯著差異（t (586) = 1.164, p = .722, d = .01），且單一回應組使用的角色米卡莎，其在網友人氣投票中的排名比多回組應的五名角色（艾蓮、萊納、亞妮、阿爾敏、希斯特莉亞）高。因此聊天使用體驗的分數差異，不是單一頭像好感度或人氣的問題，應是回應數量所導致。

未來研究建議

模型有待加強

儘管已經使用之前評估過效果不錯的聊天引擎[7]，但依舊有許多受試者反應模型有前後對話不連貫、自說自話等情況。尤其因為實驗情境受控，模型並沒有做進一步的調整，有出現前後對話矛盾，或與人設不符（如女性動漫人物頭貼卻回應自己是男生）的情形發生。因此在未來如何打造更好的聊天文字生成引擎，使模型可以進行多輪對話，能夠根據先前聊天內容或參考人物設定生成回應，是個可以探討的議題。

好體驗是因回應總量多還是有多則回應？

使用體驗較高，是否是因為收到回應總數較多而非每次收到的回應較多？儘管在多人聊天情境的十輪對話中，受試者總共收到五十則回應，比單人聊天情境的十則多五倍，但根據訪談結果，受試者普遍因多則回應較可能出現適合回應，因此能持續對話而給予較高分數。至於是否因回應較多有較好體驗？本次規劃實驗時，我們認為受試者沒有耐心一次跟一個機器人聊天五十輪，因此沒有規劃。但從受試者熱烈的響應看來，如果跟動漫頭像聊天五十輪，可能還可接受。因此未來可安排類似實驗進一步驗證這個問題。

最適合回應數量是多少？

本研究結果顯示，比起一則回應，如果對使用者的發話提供五則回應，會有較好的使用體驗。但尚未知道回應數量與使用體驗之間的關係是否為線性，是否有「最適合的回應數量」？因此未來可針對此做單因子多組設計，探討不同回應數量與使用體驗的關係。

使用者真的注意適合回應並忽略不適合回應嗎？用眼動儀看

本研究使用自陳量表，但目前有許多運用眼動儀觀察使用者的注意情況之研究。未來研究可以透過眼動儀，探討在收到多個回應時，使用者是否會對比較適合的回應有較高的視覺停留時間？

使用者對多回應但匿名頭像情境的困惑與反饋

多回應但匿名頭像的聊天體驗中，許多使用者反映因為對方的頭貼清一色都是匿名頭像，所以在收到多則回應時會下意識認為是由同一人發出，進而因不同回應之間的矛盾感到困惑，認為「一個人怎麼一次說這麼多句不相干的話？」結果影響體驗。至於動漫頭像因為多則回應由不同頭像發出，無此問題。因此未來如要做多則回應但匿名頭像的控制組，也許要將匿名頭像做一定程度的區分，如用不同顏色或加上編號，或是多個頭像時每個頭像要有明顯的區別，以降低使用者區別個別回應的困難。

自己選擇頭像與回應模式？（客製化）

本研究採用的的聊天引擎[7]，有支援生成不同個性的回應，如正面喜歡、悲傷、同理心、憤怒、不友善等。本次實驗統一使用正面喜歡模式。因為使用者可自行選擇是提高使用動機的要素之一，未來研究可藉由共軛控制，讓受試者自行選擇聊天對象的頭像與個性，對比被動接受系統設定，以探討是否能帶來更好的使用體驗？

聊天機器人的好感度有恐怖谷現象嗎？

本次研究中指出有動漫頭像的聊天機器人會有較好的好感度分數，且認為該現象與恐怖谷理論有關，即更像人類的聊天對象，會得到更高的好感度分數，然而在谷底的部分並沒有多加著墨。在與受試者做事後訪談時，受試者有提到，在與動漫頭像聊天時，一開始是有好感，但當動漫頭像出現矛盾的回應，如女性頭像說自己是男生，好感度與使用體驗就會瞬間大幅降低，這也許是恐怖谷的一種，即與既有認知不協調所帶來的矛盾。未來研究應可在此部分多加著墨，嘗試使用不同的頭像（真人、動漫、3D 人物）並調整不同的回應，查看此與好感度及使用體驗的關係。

使用的量表限制

本次研究中，所使用的聊天機器人易用度量表（The Chatbot Usability Questionnaire）[15]，並非專門針對「陪聊」聊天機器人設計，且實驗網站嚴格受控排除干擾，除聊天並沒有其他功能，因此有些用以評估客服、虛擬助理功能的題目（如：聊天對象很好地說明了其範圍和目的）難以取得高分。這也導致即使是最高分數的多則回應動漫頭像組別的平均也僅 37 分（滿分為 100 分）。建議未來如要進行類似研究，應開發一套專門用來評估陪聊對話使用體驗之量表，預期會有更好的效度。

附件一：本研究使用之 CUQ 量表中文翻譯版

題號	題目敘述
1	聊天對象的個性逼真且有參與感
2	聊天對象似乎太機械化了
3	聊天對象讓我感覺受重視
4	聊天對象看起來很不友善
5	聊天對象很好地說明了其範圍和目的
6	聊天對象未表明其目的，意義不明
7	聊天對象很好的引導我們的對話
8	跟對方聊天時，我很容易感到困惑
9	這個聊天對象懂我
10	聊天對象無法理解我說的內容
11	聊天對象的回應是有用的、適合的且資訊豐富的
12	聊天對象的回應是與我說的內容不相關的
13	聊天對象很好的應對我的任何錯誤或失誤
14	聊天對象看起來無法處理我的失誤或錯誤
15	這個聊天機器人很好用
16	這個聊天機器人非常的複雜

參考資料

[1] Xu, A., et al. A new chatbot for customer service on social media. in Proceedings of the 2017 CHI conference on human factors in computing systems. 2017.

[2] Apple. Siri. Available from: https://www.apple.com/tw/siri/.

[3] 北一女校慶創意擺攤「陪聊 10 分鐘 45 元」引網友暴動. Available from: https://udn.com/news/story/6898/5090655.

[4] Fitzpatrick, K.K., A. Darcy, and M. Vierhile, Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR mental health, 2017. 4(2): p. e19.

[5] 即時新聞／綜合報導. 扯！南韓女性聊天 AI 被用戶「調教」成性奴. Available from: https://news.ltn.com.tw/news/world/breakingnews/3406182.

[6] 詞庫小妍. Available from: https://ckip.iis.sinica.edu.tw/project/chatq.

[7] 楊德倫 and 曾元顯, 建置與評估文字自動生成的情感對話系統. 教育資料與圖書館學, 2020. 57(3): p. 355-378.

[8] LeCun, Y., Y. Bengio, and G. Hinton, Deep learning. nature, 2015. 521(7553): p. 436-444.

[9] Bengio, Y., et al., A neural probabilistic language model. Journal of machine learning research, 2003. 3(Feb): p. 1137-1155.

[10] Radford, A., et al., Language models are unsupervised multitask learners. OpenAI blog, 2019. 1(8): p. 9.

[11] Turing, A.M., Computing machinery and intelligence, in Parsing the turing test. 2009, Springer. p. 23-65.

[12] Mori, M., K.F. MacDorman, and N. Kageki, The uncanny valley [from the field]. IEEE Robotics & Automation Magazine, 2012. 19(2): p. 98-100.

[13] Möhlmann, M., Collaborative consumption: determinants of satisfaction and the likelihood of using a sharing economy option again. Journal of Consumer Behaviour, 2015. 14(3): p. 193-207.

[14] Arons, B., A review of the cocktail party effect. Journal of the American Voice I/O Society, 1992. 12(7): p. 35-50.

[15] Holmes, S., et al. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? in Proceedings of the 31st European Conference on Cognitive Ergonomics. 2019.

[16] Brooke, J., Sus: a " quick and dirty ' usability. Usability evaluation in industry, 1996. 189.

[17] Laugwitz, B., T. Held, and M. Schrepp. Construction and evaluation of a user experience questionnaire. in Symposium of the Austrian HCI and usability engineering group. 2008. Springer.

[18] Aron, A., E.N. Aron, and D. Smollan, Inclusion of other in the self scale and the structure of interpersonal closeness. Journal of personality and social psychology, 1992. 63(4): p. 596.

[19] gnehs, 勝. Chat UI. Available from: https://gnehs.github.io/ChatUIDoc/.