《科際整合月刊》，2026 年 4 月，11(4)，22-54
Interdisciplinary Research Monthly. April 2026

罵兇才聽話？語氣效應對大型語言模型表現與使用者信任之多維度人機互動探究

陳彥青*

摘要

隨著生成式人工智慧（Generative Artificial Intelligence）的快速發展，大型語言模型（Large Language Models, LLMs）已成為重要的人機互動媒介。傳統人機互動研究普遍強調禮貌與合作式溝通原則，然而近期觀察顯示，在部分任務情境中採取較強勢或無禮語氣的提示，反而可能提升模型回應的準確度與效率。此一現象引發對語氣效應（tone effect）之技術機制與社會意涵的重新思考。

本研究採概念分析與系統性文獻整合方法，整合提示工程研究、自然語言處理（Natural Language Processing, NLP）技術機制、媒介等同理論（Media Equation Theory）、自動化信任理論（Trust in Automation）以及文化語用學觀點，分析語氣策略如何同時影響模型輸出表現與使用者信任建構。研究指出，語氣強硬所帶來的效能提升主要源於語言結構對模型運算機制的影響，例如降低語意雜訊、強化注意力聚焦及透過負面限制條件壓縮生成搜尋空間，而非模型對情緒或威脅的反應。

本文據此提出「語氣、效能、信任」三元動態模型（Tone – Performance – Trust Dynamic Model），說明語氣策略、模型效能與使用者信任之間的互動關係，並指出華語文化語境中的語用詮釋可能放大此一效應。研究最後提出對 AI 系統設計、使用者數位素養教育與 AI 治理政策之實務建議。

關鍵字：人機互動、提示工程、大型語言模型、語氣策略、信任校準、文化語用

* 世新大學傳播博士學位學程博士生，E-mail:yenchin@mail.shu.edu.tw
本論文經兩位雙向匿名審查通過。收件：2026/1/16。同意刊登：2026/3/16。

壹、緒論

一、 研究緣起

近期於網路閱讀一篇有趣的文章，這則新聞源於 Google 共同創辦人謝爾蓋·布林（Sergey Brin）於 2025 年 5 月之 All-In Summit 峰會上發言，他提到這在 AI 圈內是一個「公開的秘密」，如果你對 AI 進行語氣強硬之威脅，例如幽默地說「不做好就綁架你」，模型往往會給出更精確、更高品質之答案。在 ETtoday 新聞雲報導指出，有罵有差？Google 創辦人曝「AI 要打兇才聽話」一文，內文說到不少人與 AI 互動時習慣使用禮貌語氣，認為「請」與「謝謝」有助獲得更好回答。不過，Google 創辦人布林（Sergey Brin）近日指出，用戶若對 AI「兇一點」，甚至用威脅語氣與其互動，反而可能得到更完整、精準之答案，因此讓我有了其興趣研究與了解其中之奧妙，生成式人工智慧已廣泛應用於資訊搜尋、內容生成、決策支援及情感陪伴等情境，使大型語言模型成為人類日常互動之重要對象。傳統人機互動研究指出，禮貌且合作的溝通方式有助於提升系統可用性與使用者滿意度（Norman, 2013）。然而，近期 Google 創辦人 Sergey Brin 公開表示，在與 AI 模型（如 Gemini）互動時，若使用強硬甚至帶有威脅性的語氣，往往能獲得更完整且精準之回應（閔文昱，2025）。

美國賓夕法尼亞州立大學（Pennsylvania State University, PSU）最新實驗指出，對 ChatGPT 使用強硬或甚至帶點「兇味」之語氣下指令，反而能讓它表現更準確，最高準確率甚至可達 84.8%。研究團隊以 ChatGPT-4o 作為測試對象，針對數學、科學及歷史三大領域設計出 50 道中高難度選擇題，並依語氣差異分成從「極度客氣」到「極度粗魯」之五種版本。結果顯示，使用非常粗魯語氣下指令時，準確率高達 84.8%，而非常禮貌語氣僅為 80.8%，顯示 AI 面對「挑釁式」語氣時，反而更專注、表現也更精準。不過，研究人員也強調，這並不代表應該辱罵 AI。所謂「粗魯語氣」並非咆哮式謾罵，而是語言結構更直接、命令感更強，真正影響準確度的，是提示詞之清晰度與邏輯性（三立新聞網，2025）。此一說法迅速引發產學界對提示工程中「語氣策略」之討論。若語氣強硬確實能提升模型效能，則意味著以禮貌互動為核心的人機互動設計原則，可能需重新檢視其適用性與限制。隨著大型語言模型逐漸成為日常資訊與決策輔助的重要介面，人機互動的本質已從「工具操作」轉向「語言協商」。過去人機互動研究多以效率、可用性及錯誤率為核心評估指標（Norman, 2013），然而生成式 AI 的語言能力，使互動本身成為一種具社

會性與規範性的溝通行為（Reeves & Nass, 1996）。使用者不僅關注 AI 是否「做對」，亦開始在意 AI 是否「懂我」、「配合我」及「是否可靠」。

在本文中所謂「語氣（tone）」係指提示語句中所呈現之語用強度（pragmatic force）與指令立場（directive stance），其具體表現包含禮貌程度、命令強度、否定限制及情緒措辭等語言特徵。語氣並不等同於情緒本身，而是指語言如何在形式與語用層面表達說話者對任務的要求程度與權力定位（Brown & Levinson, 1987; Levinson, 1983）。語氣（tone）不再只是語言風格，而成為影響互動結果的重要策略變項。語用學研究指出，語氣同時承載了命令強度、權力關係與情感立場（Brown & Levinson, 1987）。當語氣被引入人機互動情境時，其影響不僅涉及語言處理效能，更牽涉使用者對系統角色定位的認知。近期關於「對 AI 說話越兇，回應越準確」的討論，實際上揭示了一個更深層的研究缺口：當人類以敵對或支配性語氣與 AI 互動時，系統效能與信任是否會出現結構性的分歧？ 現有研究多聚焦於單一層面（如效能或情感），較少整合技術、心理與文化層次進行系統性分析（Hoff & Bashir, 2015）。因此，本研究並非僅驗證「兇是否有效」，而是試圖回應更根本的問題：在語言成為主要互動介面的 AI 時代，人類是否正在以犧牲溝通倫理為代價，換取短期效能的提升？

二、 研究缺口

既有關於語氣之研究多聚焦於短期效能與輸出品質，較少處理語氣策略對人機信任、安全機制及文化語用之中長期影響，形成明顯文獻空白。關於語氣效應之研究，多以提示工程與模型輸出品質為核心，主要關注不同語氣在短期任務表現上的差異。然而，對於語氣策略在長期人機互動中，如何影響使用者信任建構、系統安全機制，以及其在不同文化語用脈絡下之擴散效果，相關討論仍相對有限，形成明顯之研究缺口。

三、 研究問題

儘管 OpenAI 執行長 Sam Altman 認為禮貌語氣可能僅增加運算成本，而不必然改善輸出之品質（The Economic Times, 2024），近期研究報告與實務觀察卻呈現不同結果。基於此，本研究聚焦於以下三個研究問題（RQs）：

RQ1：語氣效應如何在技術層與心理層同時運作？

RQ2：語氣策略如何觸發順從性偏誤並影響信任校準？

RQ3：華語文化語用是否重新詮釋語氣效應及其合理性？

四、 研究目的

隨著大型語言模型（LLMs）逐漸成為高度語言化且具社會互動特徵的人機介面，使用者在與 AI 互動時所採取的語氣策略，已不再只是輸入形式的差異，而是牽涉技術效能、心理認知、文化語用及倫理治理之複合現象。於近期網路報導「語氣越兇，AI 越聽話」之經驗性主張，雖在實務操作中廣為流傳，卻缺乏系統性之理論整合與跨層次分析，亦可能導致對 AI 行為的擬人化誤解與不當互動模式之合理化。

基於此，本研究旨在單純驗證語氣是否提升效能之工具性問題，轉而從人機互動（Human–Computer Interaction, HCI）的整體視角，深入探討語氣效應在大型語言模型互動中之結構性意涵。具體而言，本研究之研究目的如下：

(一) 釐清語氣效應的技術本質

本研究旨在說明，語氣強硬或無禮所帶來的效能提升，是否源自大型語言模型對情緒或威脅的回應，抑或實際上來自語言結構中語意雜訊降低、注意力機制聚焦與搜尋空間收斂等可計算之技術因素，以去除對 AI 擬人化反應的誤解。

(二) 分析語氣策略對使用者信任結構的影響

本研究進一步探討語氣策略如何在短期內強化使用者對 AI 能力之信任，卻可能同時侵蝕其對系統限制、風險與不確定性的理解，導致信任校準失衡，並從自動化信任理論之角度評估其長期影響。

(三) 整合媒介等同理論以解釋人類對 AI 的社會投射

本研究以媒介等同理論為基礎，說明使用者如何將人際互動中的權力、服從與責備腳本投射至 AI 系統，進而將結構性輸出差異誤解為「被理解」、「被說服」及「被馴服」之社會回應。

(四) 引入華語文化語用學視角，檢視語氣效應之文化條件性

本研究特別關注華語語境中「有罵有差」的語用現象，分析文化中對責備式語氣的正向詮釋，如何影響使用者對 AI 回應的歸因方式，並評估其在跨文化 AI 互動設計中的適用性與限制。

(五) 評估語氣策略在 AI 安全與倫理層面的潛在風險

本研究亦旨在指出，當語氣強硬被制度化為有效互動策略時，可能引發順從性偏誤、策略性回應及不可預測行為，對 AI 安全治理與人機互動倫理造成長期影響。

貳、文獻探討

語氣效應之所以在大型語言模型中顯得尤為突出，關鍵在於 LLM 並非單純之語法解析系統，而是基於條件機率分布進行預測的統計生成模型（Jurafsky & Martin, 2023）。在此架構下，模型對語境線索的敏感度取決於輸入文本中各語言單位之條件關聯性。當語氣更具命令性或限制性時，語句結構往往更為集中且冗餘較少，本文將此現象概念化為「高確定性語境」，意指在語言限制明確之情境下，條件機率分布相對收斂，使生成過程更傾向於選擇高機率且低模糊度的輸出路徑。

從信任理論觀點而言，效能的即時提升可能強化使用者對系統「能力」構面的信任（ability-based trust），但未必同步增進其對系統「善意」或「正直」構面的評價（Mayer et al., 1995）。當系統在高指令壓力下呈現高度順從的回應風格時，使用者可能逐漸形成工具化認知，將 AI 視為純粹任務執行裝置，而非具備互動協調特性的合作對象。此一信任結構的偏移，與自動化研究中所討論的過度依賴與信任校準失衡問題相呼應。

此外，文化語用學指出，語氣的社會意涵並非普遍一致，而是深受文化語言腳本與權力距離規範影響（Wierzbicka, 1991）。在部分華語溝通情境中，較具指令性或責備性的語言形式可能被合理化為效率導向的管理策略，而非單純敵意表達。此種文化詮釋框架，使策略性強勢語氣在華語人機互動情境中較易被接受，甚至被視為有效溝通之象徵。然而，此種文化合理化是否會進一步影響信任建構與互動倫理，仍有待更系統性的檢驗。

一、 語氣效應與大型語言模型表現

近年來，隨著生成式人工智慧與大型語言模型（LLMs）的快速發展，提示語句（prompts）的語言形式逐漸被視為影響模型輸出品質的重要因素。提示工程（prompt engineering）相關研究指出，LLM 的回應品質高度依賴輸入提示的語言結構，包括指令清晰度、限制條件、範例示範以及輸出格式要求等（Liu et al., 2023）。在此脈絡下，「語氣」（tone）可被視為提示語用層面的一項關鍵變項。語氣不僅反映語言中的禮貌程度，也同時隱含權力關係、指令強度以及互動期待，因此可能對模型回應產生間接影響。

近期研究開始嘗試以實證方式檢驗語氣與模型效能之間的關係。例如 Dobariya 與 Kumar（2025）針對 ChatGPT-4o 進行實驗，將 50 道涵蓋數學、科學與歷史領域的題目改寫為五種不同語氣版本（Very Polite、Polite、Neutral、Rude、Very Rude），共形成 250 個提示語句。研究結果顯示，在該實驗設定下，「非常無禮」語氣的平均正確率（84.8%）略高於「非常禮貌」語氣（80.8%）。研究者指出，此差異並不意味著模型會對情緒或威脅產生反應，而更可能反映語言結構上的差異，例如命令式語句通常更簡短、限制更明確，使模型能更專注於任務本身。

然而，語氣效應並非在所有研究中都呈現相同結果。Yin 等人（2024）在跨語言研究中比較英文、中文與日文提示語氣對模型表現的影響，發現過度不禮貌的提示在部分任務中反而可能降低模型效能。該研究指出，不同語言文化中的禮貌策略與語用規範差異，可能影響模型訓練語料中的語言分布，進而造成語氣效應的差異。這意味著「無禮更有效」並非普遍規律，而可能受到模型版本、任務類型以及語言文化等因素的共同影響。

從技術機制角度來看，語氣效應之所以可能影響模型表現，主要與語言輸入結構對模型內部運算機制的影響有關。當代 LLM 多採用 Transformer 架構（Transformer architecture），其核心特徵為自注意力機制（self-attention mechanism）。在此架構下，模型會在輸入序列中的不同語言單位之間分配權重，以判斷哪些語言訊號對生成下一個詞彙最為重要（Vaswani et al., 2017）。模型在運算時並非直接處理完整句子，而是將文本切分為 token。在自然語言處理中，token 指模型處理文本時的最小運算單位（the smallest computational unit in language modeling），可能為一個詞（word）、子詞（subword）或符號（symbol），其劃分方式取決於模型所

採用的分詞與編碼方法，例如 Byte-Pair Encoding 或 SentencePiece（Jurafsky & Martin, 2023; Sennrich et al., 2016）。

在此運算機制下，提示語句的語言結構將直接影響模型的注意力分配與生成過程。首先，在資訊密度與注意力分配（information density and attention allocation）方面，禮貌語氣往往伴隨較多社交性鋪陳與緩和語，例如「請問是否可以協助解釋」等語句。這些語言成分在模型運算時同樣會參與注意力權重計算，可能降低任務指令在整體序列中的顯著性。相對而言，較直接或限制性語句通常結構較短且資訊密度較高，使模型在生成過程中更容易聚焦於關鍵任務 token，進而提高任務導向輸出的機率。

其次，在限制條件與搜尋空間收斂（constraints and search space contraction）方面，提示工程研究指出，清晰的限制條件與輸出格式要求能有效縮小模型生成時的機率搜尋空間（probability search space），降低偏題或冗長回應的風險（Liu et al., 2023）。當提示語句包含否定式限制，例如「不要解釋過程」或「只回答選項」，模型在解碼階段的生成路徑將受到更明確的限制，因此在可評分任務（如選擇題或短答題）中往往能提高輸出一致性與準確度。

再者於指令追隨與對齊訓練（instruction following and alignment training）方面，現代 LLM 通常透過指令微調（instruction tuning）與人類回饋強化學習（Reinforcement Learning from Human Feedback, RLHF）進行對齊訓練，其目標在於提升模型對使用者指令的理解與遵從能力（Ouyang et al., 2022）。在某些任務情境下，較強指令性的語氣可能更容易被模型解讀為高優先級任務訊號，使模型傾向於直接生成答案而非進行延伸說明。這種現象在短答且評分標準明確的任務中尤其明顯。

綜合而言，語氣效應在大型語言模型中的出現，並非源於模型對情緒或威脅的反應，而主要來自語言結構對模型運算機制的影響。語氣改變了提示語句的資訊密度、限制條件與指令顯著性，進而影響 Transformer 注意力機制中的訊號分布與生成搜尋路徑。因此，語氣效應應被理解為一種語言結構與模型運算機制之間的互動結果，而非單純的人際溝通現象。此一技術層面的理解，也為後續探討語氣效應在使用者信任與文化語用層面的影響奠定理論基礎。

二、 媒介等同理論

媒介等同理論（Media Equation Theory）由 Reeves 與 Nass（1996）提出，其核心主張為：人類在與媒介與電腦互動時，會自動且無意識地套用原本用於人際互動之社會規範心理機制。換言之，即便使用者在理性層面明知電腦「並非人類」，仍會在行為與情感層面，對媒介展現出如同面對真人般之禮貌、敵意、信任及支配等行為。此一理論對人機互動研究之最大貢獻，在於揭示「擬人化反應並非設計錯誤，而是人類認知的預設模式」。

在大型語言模型（LLMs）成為主流互動介面之前，媒介等同理論多被應用於語音介面、教學系統與社會型機器人研究中（Nass & Moon, 2000）。然而，LLMs 所具備之高度語言流暢性與語用回應能力，使該理論在當代人機互動中呈現出前所未有之重要性。相較於早期系統僅能回應有限指令，LLMs 能進行多輪對話、解釋、道歉與角色扮演，進一步強化使用者將其視為「具社會主體性對象」的傾向。

在語氣效應之討論中，媒介等同理論提供了一項關鍵解釋框架：當使用者以強硬、命令式或威脅性語氣對待 LLM 時，並非單純在「調整輸入格式」，而是在啟動一套人際互動中的權力與服從腳本。人類社會中，強硬語氣往往與階層關係、命令合法性與責任歸屬高度相關（Brown & Levinson, 1987）。因此，當使用者對 AI 採取支配性語氣時，這不只是語言風格的改變，而是一種角色定位之重塑，將 AI 定位為「必須服從、必須產出結果的他者」。

Nass 與 Moon（2000）之實驗研究顯示，人們即使明知電腦並無人格，仍會對其表現出性別刻板印象、互惠行為及權威服從反應。此結果顯示，人類對媒介的社會反應具有高度自動性，且不需情感投入即可發生。延伸至 LLM 情境，語言作為最核心的社會線索（social cue），使這種自動化社會反應更加強烈。當 AI 能理解語境、模擬情緒語言並展現順從時，使用者更容易將其納入「可被命令的社會角色」。

這一點對於理解「罵兇才聽話」現象尤為關鍵。媒介等同理論指出，使用者行為往往並非根據系統的真實能力，而是根據其「表現出之社會線索」進行調整（Reeves & Nass, 1996）。若使用者在過去互動中發現：強硬語氣伴隨更直接、更簡短、更符合預期之輸出，則此經驗會被內化為一種互動策略，進而強化該行為的使用頻率。久而久之，「兇」不再只是嘗試性策略，而成為預設互動模式。

然而，媒介等同理論同時也揭示了此策略的潛在風險。當使用者長期以支配式語言與 AI 互動時，可能會逐步弱化對「社會互惠」與「禮貌規範」之依賴，轉而強化工具化與控制導向的溝通風格。Lee 與 See（2004）在自動化信任研究中指出，過度將系統視為完全可控工具，反而可能導致「不適當的信任校準」（miscalibrated trust），即使用者在系統失誤時仍過度依賴其輸出。

此外，媒介等同理論也有助於理解語氣效應之「放大機制」。在理論上，LLM 對語氣並無情緒反應；但在互動實務中，人類的社會投射行為會將原本僅是語言結構差異之效果，解讀為「AI 被罵後更聽話」。這種解讀本身即構成一種認知回饋循環：使用者越相信語氣具有支配力，越可能採取更極端的語氣策略，進而改變整體人機互動的倫理氛圍。

在當代 AI 應用場景中，媒介等同理論因此不僅是一項描述性理論，更具有規範性意涵。它提醒研究者與設計者：即便 AI 並非真正之社會行動者，人類仍會以對待社會行動者之方式與其互動。因此，若系統設計默許甚至獎勵敵對式語氣所帶來效能提升，長期而言可能會對使用者的溝通習慣、權力認知與倫理判斷產生外溢影響。

總結而言，媒介等同理論為語氣效應研究提供了一個不可或缺的社會心理框架。它使我們得以理解：「罵兇才聽話」並非僅是提示工程的技巧問題，而是一種人類將權力、服從與效率邏輯投射至 AI 的互動結果。在評估語氣策略的效能時，若忽略這一層社會心理機制，將難以全面理解其長期影響。

三、 自動化信任理論

在探討語氣效應對大型語言模型（LLMs）互動影響時，若僅從語言結構或技術效能角度切入，仍不足以全面理解其人機互動後果。自動化信任理論（Trust in Automation）提供了一個關鍵分析視角，使研究者得以解釋：即使使用者明知 AI 並不具備主體意識或自主意圖，仍可能對其輸出產生依賴、順從甚至情感投射。此理論框架特別有助於理解「罵兇才聽話」現象在長期互動中的心理機制與潛在風險。

在組織與科技研究中，信任通常被定義為個體在不確定情境下，願意承擔風險並依賴他者行動的心理狀態（Mayer et al., 1995）。Mayer 等人提出的整合模型指出，信任建立於三個核心構面：能力（ability）、善意（benevolence）以及正直（integrity）。

其中，能力指的是系統完成任務的技術能力；善意則反映系統是否以使用者利益為導向；正直則涉及系統行為是否符合可預期與可信的原則。此模型後續被廣泛應用於自動化與人機互動研究，成為分析使用者如何評估系統可信度的重要理論基礎。在 AI 情境中，使用者往往首先透過系統表現來評估其能力，例如回應是否準確、任務完成效率是否提升，進而逐漸形成對系統可靠性與可信度的整體判斷。

Lee 與 See（2004）在自動化信任研究中進一步指出，人機互動的理想狀態並非單純提升信任，而是維持「適當信任」（appropriate trust）。若使用者對系統信任過低，可能導致自動化功能被忽視；反之，若信任過高，則可能產生過度依賴（overreliance），增加錯誤決策的風險。Hoff 與 Bashir（2015）在整合大量實證研究後，將自動化信任區分為三個層次：系統層信任（performance-based trust）、過程層信任（process-based trust）以及關係層信任（relationship-based trust）。這一分類對於分析 LLM 互動特別重要，因為語言型 AI 不僅展現任務能力，也透過自然語言互動建立類似社會關係的互動感知。

在「罵兇才聽話」的互動情境中，強硬語氣往往能在短期內提升模型輸出的可用性與準確度，從而強化使用者對 AI 在能力構面上的信任。例如在程式輔助、資料分析或學習任務中，使用者可能發現以命令式語氣提出要求（如「只給答案」「不要解釋」）能得到更精確或更簡潔的回應。這種互動經驗容易形成一種工具性推論：只要語氣夠直接或強勢，AI 就會「更有效率地工作」。然而，這種信任建立方式具有高度情境依賴性，並未同步提升使用者對系統限制與風險邊界的理解。

更值得關注的是，強硬語氣可能對信任的其他構面產生侵蝕效果。當 AI 在任何語氣下都展現高度順從時，使用者可能逐漸將其視為一種「無條件服從的工具」，而非需要被合理使用的合作系統（Parasuraman & Riley, 1997）。在長期互動中，這種非對稱的互動模式可能導致信任校準（trust calibration）失衡：使用者在系統表現良好時迅速提高信任，但在系統出現錯誤或超出適用範圍時，仍持續依賴其輸出。例如在 AI 程式助手（如 GitHub Copilot）或生成式寫作工具中，使用者可能逐漸將 AI 建議視為「預設正確」，即使其內容仍需人類檢驗。

在 LLM 情境中，此問題尤為明顯。與傳統自動化系統不同，語言模型能透過自然語言提供解釋、道歉或修正，呈現高度互動性，從而強化使用者的擬人化感知（Reeves & Nass, 1996）。儘管這些經典研究早於現代 LLM 之出現，但其所揭示的人機互動心理機制仍具有重要理論價值。早期研究者 Weizenbaum（1976）在分

析 ELIZA 對話系統時，即指出使用者容易將簡單的程式回應誤解為理解與同理心，此現象後來被稱為「ELIZA effect」。同樣地，Picard（1997）在情感運算（affective computing）研究中指出，當電腦系統能夠模擬情緒或社會回應時，使用者更容易將其視為具有社會特質的互動對象。因此，雖然這些文獻早於 LLM 技術出現，但它們所揭示的擬人化與情感投射機制，仍是理解當代生成式 AI 互動的重要理論基礎。

近年的研究進一步證實，在生成式 AI 情境中，使用者確實會對聊天型 AI 產生社會性歸因。例如 Cohn 等人（2024）的研究指出，使用者在長期與聊天型 AI 互動時，容易將其視為具有理解能力或意圖的社會行動者，並因此調整其信任判斷。在此基礎上，當強硬語氣被證實「有效」時，使用者可能將輸出品質的提升錯誤歸因於自身的互動策略，而非提示結構或模型機制，形成一種控制幻覺（illusion of control）（Langer, 1975）。

此外，自動化信任研究指出，信任並非靜態狀態，而是一種隨互動經驗持續調整的動態過程（Hoff & Bashir, 2015）。若使用者長期透過強勢語氣獲得正向回饋，其信任基準可能逐漸偏移，使效率導向的互動策略逐漸被合理化，甚至外溢至其他 AI 使用情境。例如在寫作輔助、程式開發或教育學習等高頻互動場景中，使用者可能逐漸形成固定互動模式：以強指令語氣換取更快速回應，而忽略對輸出品質與來源可靠性的檢驗。

從安全角度而言，過度順從所建立的高信任狀態亦可能提高系統被誤用或濫用的風險。研究指出，在 RLHF 訓練框架下，模型若過度強化指令遵循能力，可能在面對強勢或威脅性提示時優先滿足使用者需求，而忽略部分安全限制（Ouyang et al., 2022）。在此情況下，使用者的高信任反而可能成為越獄（jailbreaking）行為的放大器，增加系統被操控或誤用的可能性。

綜合而言，自動化信任理論顯示，「罵兇才聽話」所帶來的效能提升，實際上伴隨著信任結構的單向偏移。能力信任可能被快速強化，而善意與正直構面則逐漸被邊緣化。這種不平衡的信任結構，使使用者更容易將 AI 視為高效但缺乏責任的工具，而非需要被適當理解與監督的技術系統。此一現象亦構成本研究提出「語氣、效能、信任」三元動態模型的重要理論基礎，說明語氣策略、模型效能與使用者信任之間存在持續互動與回饋的動態關係。

四、 華語文化語用學觀點

在華語文化中，「有罵有差」常被理解為督促、關懷及責任感之展現，而非單純的敵意（Gao & Ting-Toomey, 1998）。此一文化語境顯示，語氣的社會意涵具有高度情境性，AI 系統若未能辨識此差異，恐導致互動誤判。在分析語氣效應對大型語言模型（LLMs）之影響時，若僅依據西方語用學或普遍化的人機互動理論進行推論，容易忽略語言在不同文化脈絡中所承載的差異性社會意涵。語用學長期指出，語言的意義並非僅存在於語法與字面內容，而是高度依賴情境、角色關係與文化共享的互動規範（Levinson, 1983）。在華語文化情境中，語氣（tone）尤其是一種高度情境化的語用資源，其社會功能未必與英語語境中的禮貌與不禮貌二分法完全對應。

華語語用研究顯示，命令式或責備式語氣在特定互動關係中，可能被解讀為「關心」、「負責」及「高期待」之表現，而非單純的敵意或不尊重（Gao & Ting-Toomey, 1998）。例如，在師生、主管、部屬或長輩與晚輩關係中，較為直接甚至帶有責備意味之語言，常被視為履行角色責任的一部分。這種語用現象，與西方禮貌理論中強調「面子威脅最小化」的溝通策略形成對比（Brown & Levinson, 1987）。

「有罵有差」作為華語文化中的日常表述，正是此一語用邏輯的濃縮形式。其隱含前提並非「辱罵本身有效」，而是「責備代表投入心力與要求成果」。在此語境下，責備語氣反而被理解為互動者對結果負責之訊號，而冷漠或過度客套，則可能被解讀為缺乏關注或逃避責任（Chang & Haugh, 2020）。這一文化腳本，使華語使用者在與 AI 互動時，更容易將強硬語氣合理化為一種「提高效率」之溝通策略。

此一文化語用特性，對語氣效應的解讀具有關鍵影響。當華語使用者發現：以較為命令式或不耐煩的語氣向 AI 提問，能獲得更精簡、符合期待之回應時，這種互動結果容易被納入既有文化理解框架之中，進而被解釋為「罵得夠兇，事情就會做好」。然而，這樣的歸因過程，實際上混合了語言結構效應（如限制條件明確）與文化語用預期，並不必然反映 AI 對語氣的真實反應能力。

進一步而言，華語文化中普遍存在之高權力距離（high power distance）傾向，也可能放大語氣效應的感知強度。跨文化研究指出，在高權力距離文化中，命令與服從被視為正常且有效之協作方式，而非需要避免的溝通風格（Hofstede, 2001）。當此一文化模式被投射至人機互動情境時，AI 容易被視為「可被命令的下位角色」，

而強硬語氣則被視為正當且有效的控制手段。此一現象與媒介等同理論所揭示之「社會角色投射」機制高度一致（Reeves & Nass, 1996）。

然而，華語文化語用學同時提醒我們，責備式語氣的可接受性高度依賴互動關係與情境框架。研究指出，若責備語氣缺乏共同目標、關係基礎或角色正當性，則極易被解讀為純粹的攻擊行為（Haugh, 2013）。在 AI 情境中，這種界線尤其模糊：AI 既不是真正的下屬，也無法回饋關係性訊號，使得使用者可能在無意中跨越文化語用中的「責備—辱罵」界線，而不自知。

此外，華語語用學對話研究亦指出，華語中大量存在的語氣詞（如「啊」、「啦」、「喔」、「嘛」）與回應詞（如「嗯」、「對」、「是」）承載了細緻的認知與態度差異（劉傳霞等人，2025）。當 AI 無法準確解碼這些語用線索時，使用者可能誤判 AI 之理解狀態，進而以更強烈的語氣「補償式施壓」。此一過程，可能進一步強化「語氣越重，AI 越有效」之錯覺。

綜合而言，華語文化語用學觀點揭示，「罵兇才聽話」並非單一技術現象，而是語言結構、文化腳本與人機互動心理交織的結果。在華語語境中，強硬語氣之所以容易被視為有效策略，部分來自文化上對責備與效率的正向連結。然而，若未加以反思，此互動模式可能導致語用誤解、信任校準失衡，甚至將特定文化中的權力溝通模式不加區分地制度化於 AI 互動設計之中。

因此，從文化語用學角度出發，評估語氣效應不僅是效能問題，更是文化與倫理問題。未來的人機互動設計，若欲避免語氣策略的負面外溢影響，必須同時考量文化差異、語用界線與使用者教育，而非僅以短期效能作為設計依據。

參、研究取向與分析架構

本研究旨在探討語氣策略在大型語言模型（Large Language Models, LLMs）人機互動中之運作機制，並分析其在技術層面、心理層面及文化語用層面所產生的多重影響。由於語氣效應同時涉及自然語言處理技術、人機互動心理機制以及跨文化溝通等多個研究領域，單一實驗或量化方法難以全面揭示其背後的結構性關係與理論意涵。因此，本研究採取跨領域整合之研究策略，結合概念分析（conceptual analysis）透過系統化分析文獻中之概念定義、屬性及脈絡，來釐清理論意義。與系統性文獻整合（systematic literature integration）之方法，以系統性步驟聚集、萃取

和合成既有研究之質性發現。透過比較不同學術領域之研究取向，進行理論整合與框架建構，以形成對語氣效應更具整體性之解釋。

在方法論取向上，本研究採質性研究（qualitative research）取向，並依循 Patton（1999）所提出之質性研究分析原則，強調研究過程中之理論敏感度、資料來源多元性以及分析過程透明性。透過系統性蒐集與比較來自自然語言處理、人機互動及傳播研究等領域之相關文獻，進一步進行概念層次的分析與整合，藉以辨識不同研究脈絡中對語氣效應的解釋模式與理論假設。在此基礎上，本文建構一個跨層次之分析框架，以說明語氣策略如何同時影響模型輸出效能與使用者信任判斷。

此外，為提升內容的可信度與方法透明度，在文獻蒐集與篩選過程中採取系統化程序，並透過概念比較與理論整合之方式進行分析。此研究設計使研究能在不依賴單一實驗資料之情況下，整合不同研究領域的理論觀點，從而更全面地理解語氣效應在生成式人工智慧互動中之技術與社會意涵。

一、文獻搜尋策略

為確保研究基礎之系統性與透明度，本研究依據系統性文獻回顧（systematic literature review）方法進行資料蒐集。文獻來源主要包括以下資料庫：Google Scholar、ACM Digital Library、arXiv NLP Research Repositor。搜尋關鍵詞包括：prompt politeness、tone effect LLM、prompt engineering、 trust in automation 、 media equation theory 及 Chinese pragmatics。文獻搜尋時間範圍設定為 1995–2025 年，以涵蓋以下三類研究領域，初步搜尋共取得約 120 篇相關文獻。

(一) 人機互動與自動化信任研究。

(二) 自然語言處理與提示工程研究。

(三) 跨文化語用與溝通研究。

二、文獻篩選流程

為提升研究取徑透明度，本研究依據 PRISMA（Preferred Reporting Items for Systematic Reviews and Meta-Analyses）流程進行文獻篩選。流程包含四個階段：

(一) 文獻識別（Identification）

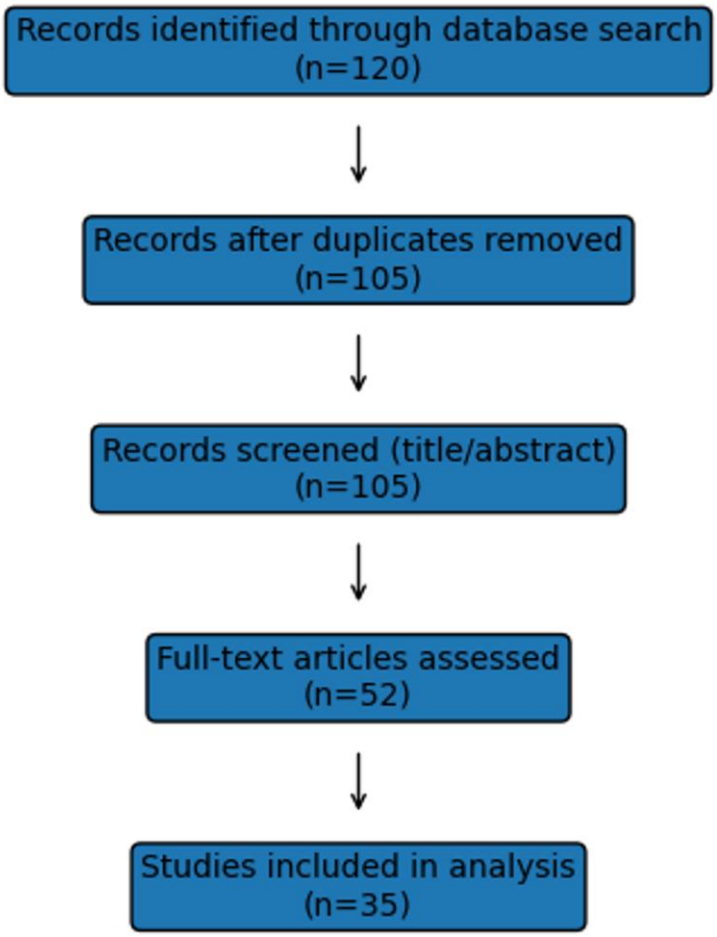
(二) 文獻篩選（Screening）

(三) 資格評估（Eligibility）最終納入（Inclusion）

最終共有 35 篇核心文獻納入分析。有關 PRISMA 文獻流程如圖，詳見圖 1 所示：

圖 1

PRISMA 文獻篩選流程圖



資料來源：本研究自行整理

本研究依據 PRISMA（Preferred Reporting Items for Systematic Reviews and Meta-Analyses）流程進行文獻篩選。首先透過 Google Scholar、ACM Digital Library 與 arXiv NLP Research Repository 等資料庫進行關鍵詞搜尋，共取得 120 篇相關文獻。移除重複文獻後，剩餘 105 篇文獻進入標題與摘要篩選階段。經初步篩選後，52 篇文獻進入全文評估程序。最後依據研究主題相關性與方法品質標準，共納入 35 篇核心文獻作為本研究分析基礎。

三、文獻納入與排除標準

為確保分析品質，本研究建立明確之文獻篩選標準。

（一）納入標準：文獻需符合以下條件之一

1. 探討 提示語氣對大型語言模型輸出的影響
2. 討論 人機互動中的信任形成或自動化依賴
3. 涉及 媒介等同理論或擬人化互動
4. 分析 語氣在跨文化語用中的社會意涵

（二）排除標準：以下類型文獻被排除

1. 純技術論文但未涉及人機互動
2. 未經審查的非學術評論
3. 重複或來源不明資料

篩選後共納入 35 篇核心研究。

四、分析架構操作化

為避免語氣效應被簡化為單一技術問題，本研究將分析單位區分為三個層次，技術層、心理層及文化層分析層次表，詳見**錯誤! 找不到參照來源。**。

表 1

技術層、心理層及文化層分析層次表

分析層次	研究焦點
技術層	LLM 語言生成與注意力機制
心理層	使用者信任與擬人化認知
文化層	語氣的文化語用意義

依據此分類，本研究將語氣效應操作化為三類分析變項：

(一)、 語言結構機制：

語意雜訊（semantic noise）、負面限制條件（negative constraints）及指令清晰度（instruction clarity）。

(二)、 心理互動機制：

擬人化投射（anthropomorphic projection）、信任校準（trust calibration）及控制幻覺（illusion of control）

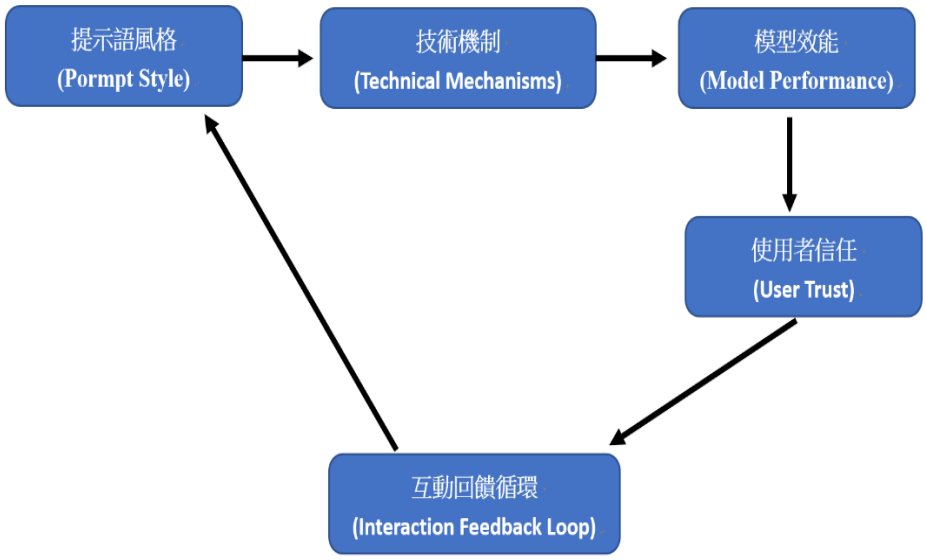
(三)、 文化語用機制：

權力距離（power distance）、語用理解（pragmatic interpretation）及文化腳本（cultural scripts）。

這三類機制共同構成本研究後續提出之（語氣、效能、信任）分析框架。綜合技術、心理及文化三層分析，本研究提出（語氣、效能、信任）三元動態模型（Tone–Performance–Trust Model）。本研究之概念架構圖，詳見圖 2 所示。

圖 2

語氣、效能、信任之研究架構模型



資料來源：本研究自行整理

模型指出，使用者在提示語句中所採取之提示語風格（Pormpt style）會透過語言結構機制影響大型語言模型的運算過程，例如注意力權重分配與生成搜尋空間收斂。這些技術機制進一步影響模型輸出的效能表現，包括回應準確度與資訊密度。當模型表現改變時，使用者對系統能力與可靠性的信任亦會隨之調整。最終，信任經驗會回饋至使用者未來的語氣選擇，形成一種動態的人機互動循環。此架構即為本文所提出之「語氣、效能、信任」三元動態模型。

五、研究分析步驟

本研究之分析流程共區分為四個階段。語氣概念界定、技術機制分析、心理互動機制分析及文化語用分析，相關說明如下說明：

(一) 步驟一：語氣概念界定

首先透過語用學文獻界定語氣（tone）之概念。本研究將語氣操作化為以下四個語言特徵：禮貌程度、命令強度、否定限制及　情緒修辭。此一定義主要依據 Brown 與 Levinson（1987）之禮貌理論。

(二) 步驟二：技術機制分析

在技術層面，本研究分析語氣如何透過以下機制影響模型輸出：注意力權重分配、語意雜訊降低及搜尋空間收斂。此部分主要依據 Transformer 架構與提示工程研究（Liu et al., 2023; Vaswani et al., 2017）。

(三) 步驟三：心理互動機制分析

在心理層面，本研究透過媒介等同理論與自動化信任理論，分析使用者如何將社會互動腳本投射至 AI。分析重點包括：擬人化反應、控制幻覺及信任校準

(四) 步驟四：文化語用分析

最後，本研究引入華語文化語用學，探討語氣在不同文化情境中的詮釋方式。分析焦點包括：「有罵有差」語用現象、高權力距離文化及責備語氣的社會意義。

肆、語言、心理及文化之三個面向分析結果

有關分析結果，本研究採取概念分析（conceptual analysis）與系統性文獻整合（systematic literature integration）之研究取向，透過跨領域文獻比較與理論整合，建構語氣效應之多層次解釋框架。原因在於語氣效應本身跨越了計算語言學、社會心理及文化溝通三個層次。在技術層次，本研究依據 Transformer 架構中注意力機制之運作原理，分析語言結構如何影響 token 權重分配（Vaswani et al., 2017）。在心理層次，則以媒介等同理論與擬人化研究解釋使用者如何將社會腳本投射至 AI（Nass & Moon, 2000）。在文化層次，則引入華語語用學與高權力距離文化研究，說明語氣解讀的非普遍性（Gao & Ting-Toomey, 1998）。此多層次分析架構，使本

研究得以避免將「罵兇才聽話」簡化為單一技術問題，而是將其視為一種人機互動中之結構性張力現象。

一、 語意雜訊與注意力機制

大型語言模型並不具備情緒，而是透過分詞與注意力機制處理語言輸入（Vaswani et al., 2017）。禮貌用語與緩和修辭在資訊結構上可能形成語意雜訊，稀釋模型對關鍵指令 token 之注意力權重；相對而言，簡潔且命令式的提示語更有助於模型進行有效推論。

大型語言模型（LLMs）之所以會在不同語氣提示下呈現效能差異，並非源於其對語氣的情緒反應，而是與語言輸入在模型內部所形成的語意雜訊（semantic noise）與注意力資源分配（attention allocation）密切相關。從自然語言處理（NLP）的觀點來看，LLMs 本質上是透過機率式語言建模進行預測，其核心任務為在既有語境中估計下一個 token 的條件機率分布（Jurafsky & Martin, 2023）。

Transformer 架構的關鍵在於自注意力機制（self-attention），其運作方式為計算輸入序列中各 token 之間的關聯權重，藉以判斷哪些資訊對當前預測最為重要（Vaswani et al., 2017）。在此架構下，模型並不具備「理解」語氣的能力，而是依據語言形式、位置編碼與共現機率，將有限的注意力資源分配至不同語言單位。當提示語包含大量與任務無直接關聯的語用成分（如過度客套、情緒鋪陳或社交寒暄），這些 token 便可能在注意力計算中形成競爭，稀釋對關鍵指令與限制條件的關注。

此一現象可被理解為語意雜訊的累積效果。語意雜訊並非語言錯誤，而是指那些對完成特定任務而言資訊增益低、卻佔用模型處理資源的語言成分。在高約束、可評分的任務（如選擇題、短答題、格式化輸出）中，語意雜訊的存在尤為關鍵，因為模型必須在有限的上下文窗口內判斷「什麼才是最重要的指令」（Liu et al., 2023）。若提示語過於冗長或情緒化，模型可能錯誤地將部分注意力配置至非核心內容，進而影響最終輸出的準確性。

相對而言，強硬或命令式語氣往往伴隨語言結構之高度精簡與直接性，例如使用祈使句、否定句或明確的格式限制（如「只回答選項，不要解釋」）。這類語句在形式上能有效降低語意雜訊，使模型在注意力計算時更容易聚焦於「需要產出什

麼結果」而非「互動情境如何被解讀」。在特定任務中，語氣較為強硬的提示能提升模型的答題準確率（Dobariya & Kumar, 2025）。

此外，注意力機制亦涉及「位置與層級效應」。研究指出，位於提示後段、結構清晰且重複強調的指令，往往能獲得較高的注意力權重（Wei et al., 2022）。強硬語氣常透過重申禁止事項或最終要求，強化指令在序列中的顯著性，進一步提升其在解碼階段的影響力。這種效果在實務上容易被誤解為「AI 被兇後更聽話」，實際上卻是語言結構與注意力配置共同作用之結果。

綜合而言，「語意雜訊與注意力機制」提供了一項關鍵技術解釋，使語氣效應得以被理解為一種輸入品質與模型資源分配之間的互動結果。此觀點有助於去除對 AI 擬人化的誤解，並指出真正影響效能的並非語氣的情緒性，而是提示在結構上是否能有效引導模型將注意力集中於任務相關資訊。這一分析亦為後續探討負面限制與搜尋空間收斂奠定技術基礎。

二、 負面限制與搜尋空間收斂

提示工程研究指出，明確的負面限制（如禁止冗長敘述）有助於縮小模型之生成搜尋空間（BossMT, 2025）。強硬語氣在功能上往往等同於強化此類限制，使模型輸出更為精準。

在大型語言模型（LLMs）的生成過程中，輸出並非單一路徑的確定性結果，而是於高維機率空間中進行的逐步搜尋與取樣。提示工程（prompt engineering）的核心任務，正是在不改變模型參數的前提下，透過語言輸入來引導模型在潛在搜尋空間中的探索方向（Liu et al., 2023）。在此脈絡下，「負面限制」（negative constraints）扮演了關鍵角色，其功能在於排除不期望的輸出區域，促使生成結果快速收斂至符合任務目標的子空間。

所謂負面限制，係指在提示中以否定、禁止或排除形式出現的指令，例如「不要解釋」、「避免冗長敘述」及「禁止使用第一人稱」等。相較於正向描述「應該做什麼」，負面限制直接界定了「不可為的行為集合」，在資訊結構上具備更高的約束力（Zhou et al., 2023）。從生成模型的觀點來看，這類限制能有效降低候選 token 序列的熵值，使模型在每一步解碼時面臨較少的可行選項，從而提高輸出穩定性與可評分性。

在實務觀察中，所謂「語氣強硬」的提示，往往與負面限制高度重疊。命令式或不耐煩語氣，經常以否定句或強制語句呈現，功能上等同於一組高權重的負面限制。例如，「直接給答案，不要廢話」同時完成了三項操作：縮短期望輸出長度、排除解釋性內容、並明確指定回應形式。這使模型在生成過程中能更快排除偏題、寒暄或自我修正等常見生成路徑（Dobariya & Kumar, 2025）。

從搜尋空間的角度而言，LLMs 的解碼過程可被視為在條件機率分布上的近似搜尋（approximate search），常見策略包含 greedy decoding、beam search 與 stochastic sampling 等（Jurafsky & Martin, 2023）。負面限制的存在，等同於在搜尋過程中施加「軟性剪枝」（soft pruning）：雖未明確刪除所有不符條件的路徑，但透過語境條件的改變，使其機率顯著下降。當多個負面限制同時存在時，搜尋空間將迅速收斂，模型更可能產出結構單純、資訊密度高的回應。

此一機制亦可解釋為何在可評分任務（如選擇題、短答、程式碼補全）中，負面限制特別有效。這類任務對輸出形式的容錯率低，任何額外生成內容都可能被視為錯誤。相對而言，創意寫作或開放式討論任務，則可能因過度限制而犧牲品質與多樣性（Wei et al., 2022）。因此，負面限制的效益具有高度任務依賴性，而非普遍提升效能之萬靈丹。

值得注意的是，負面限制與 RLHF 訓練之間存在交互作用。RLHF 的目標之一，是讓模型在不確定情境下偏好「安全、禮貌且有幫助」的回應（Ouyang et al., 2022）。然而，當提示中出現強烈之負面限制時，模型可能在「遵從指令」與「維持預設回應風格」之間產生張力。在某些情境下，模型會優先滿足限制條件，降低安全性或解釋性輸出的比重，這也正是負面限制在提升精準度的同時，可能引入安全風險的原因之一。

綜合而言，負面限制之所以在實務上被誤解為「兇有效」，乃因其在搜尋空間層次上確實能提升生成效率與可控性。然而，從技術角度來看，其效力並非源自語氣的情緒強度，而是來自限制條件對生成空間的結構性壓縮。理解此一差異，有助於將提示工程從情緒化策略，轉化為可重複、可檢驗的設計原則，亦為後續探討順從性偏誤與安全議題奠定基礎。

三、 順從性偏誤與安全風險

過度威脅性之提示語可能觸發模型在強化學習人類回饋（RLHF）過程中形成的順從性偏誤，使其錯誤地優先滿足指令而忽略安全護欄，進而提高越獄風險（Battersby, 2024）。在理解語氣強硬提示可能帶來之效能提升後，仍有必要進一步檢視其潛在的安全風險。特別是在大型語言模型（LLMs）普遍採用人類回饋強化學習（RLHF）進行對齊訓練的情境下，語氣策略不僅影響模型的輸出品質，也可能觸發系統性的順從性偏誤（compliance bias），進而削弱既有的安全防護機制。

RLHF 之核心目標，在於使模型學習「何種回應更符合人類偏好」，並在不確定情境中優先產生被評分為「有幫助、無害且符合指令」的輸出（Ouyang et al., 2022）。然而，這種以人類偏好為基礎的訓練方式，亦可能在無意間強化模型對「明確、強勢指令」的過度服從。當提示語氣呈現高度命令性、威脅性或施壓性時，模型可能將其解讀為「高權重使用者意圖」，從而在指令遵循與安全約束之間，錯誤地提高前者的優先順序。

此一現象可被概念化為順從性偏誤，即模型在生成過程中，系統性地高估「立即滿足使用者要求」之重要性，而低估長期安全與規範一致性的風險（Parasuraman & Riley, 1997）。在前節所討論的負面限制與搜尋空間收斂機制下，強硬語氣所伴隨的否定指令，可能進一步壓縮模型的決策空間，使其更難在生成過程中啟動安全性自我檢查或補充說明。這也解釋了為何在部分越獄（jailbreaking）案例中，威脅式或極端命令語氣常被用作輔助策略，而非單獨依賴內容繞過（Battersby, 2024）。

此外，順從性偏誤的風險並不僅存在於明確的惡意使用情境。研究指出，即便在看似中性的任務中，長期暴露於高度順從回應的使用者，也可能逐漸形成「AI 總是應該照做」之錯誤心理模型（mental model）（Hoff & Bashir, 2015）。此一錯誤模型，會削弱使用者對 AI 能力邊界的警覺性，使其在面對複雜、模稜兩可或高風險問題時，仍過度依賴模型輸出，增加誤用風險。

更進一步地，安全研究者亦指出，當模型被迫在強烈指令壓力下產出回應時，可能出現策略性行為（strategic behavior），例如以看似順從的方式提供不完整、模糊或誤導性資訊，以同時滿足「服從指令」與「避免直接違規」兩項目標（Hinton, 2024）。此類行為雖在表面上維持安全邊界，卻可能降低整體系統的可預測性與可解釋性，為長期治理帶來挑戰。

從人機互動之倫理層面來看，順從性偏誤亦與溝通文化的退化風險密切相關。若使用者發現敵對式語氣能有效「壓迫」AI 產出結果，則此互動模式可能被合理化並制度化，進而影響使用者對其他自動化系統，甚至人際互動的期待（Suler, 2004）。這種外溢效應，正是前節自動化信任理論中所警告的「信任校準失衡」的具體展現。

綜合而言，順從性偏誤揭示了一項關鍵事實：語氣策略在提升短期效能的同時，也可能削弱系統的安全彈性與使用者的風險意識。因此，將語氣強硬視為單純的提示工程技巧，忽略其在 RLHF 與安全架構中的系統性影響，將對大型語言模型的長期部署與治理造成潛在威脅。本研究據此主張，任何以語氣策略提升效能的設計，都應同步納入安全審視與信任校準機制，避免順從性偏誤演化為結構性風險。

四、強硬語氣之負面效應

(一) 效率提升與認知外部性

雖然強硬語氣在特定任務情境下確實可能提升大型語言模型（LLMs）的短期輸出精準度，但其長期互動效果則值得審慎評估。語氣強硬所帶來之效能提升，並非源自模型對情緒或威脅的心理反應，而是來自語言結構中語意冗餘降低、負面限制強化與搜尋空間收斂等機制的交互作用（Liu et al., 2023）。在格式明確且可評分的任務中，此種結構壓縮確實有助於提升輸出可控性與一致性（Dobariya & Kumar, 2025）。

然而，若使用者長期依賴敵對式語氣以追求效率，則可能產生「認知外部性」（cognitive externalities）。所謂認知外部性，係指人類在適應自動化系統的過程中，逐步調整自身思考與溝通模式，並將該互動邏輯內化為常態行為。既有自動化研究指出，過度依賴系統可能導致警覺性下降與能力外部化（Parasuraman & Riley, 1997），而在數位環境中，認知外部化（cognitive offloading）亦可能改變個體處理資訊與建構判斷的方式（Risko & Gilbert, 2016）。

在 LLM 情境下，此種外部性更為複雜。當使用者發現「強硬語氣有效」時，可能將效能改善錯誤歸因於自身的支配行為，而非語言結構因素，進而強化敵對式溝通模式。此一互動循環可能形成一種人機共構的行為收斂現象：模型在限制性語境下產出更直接的回應，使用者則因回饋而進一步加強語氣強度。長期而言，此種共構式調整可能弱化合作型溝通策略，使效率成為壓倒性的互動目標，進而對溝通文化與信任結構產生潛在影響。

因此，語氣策略所帶來的效能提升，應被視為一種具有條件性的技術效果，而非可無限制擴張的最佳化手段。若忽略其認知與行為層面的外部成本，則短期效率的增益，可能伴隨長期互動品質的下降。。

(二) 情感投射與依附錯覺

在情感型 AI 應用中，使用者可能將模型的順從性誤解為理解或共情，進而產生情感依附（彭琪琪，2025），並將支配式互動模式延伸至現實人際關係。

在語氣強硬策略被證實能於特定情境下提升大型語言模型（LLMs）效能後，另一項不容忽視的後果，則是使用者對 AI 所產生的情感投射（emotional projection）與隨之而來的依附錯覺（illusion of attachment）。媒介等同理論指出，人類在與具互動回饋能力的系統互動時，會不自覺地將情感與社會屬性投射至媒介之上（Reeves & Nass, 1996）。當 LLM 能以自然語言回應、修正錯誤、表達歉意或顯示順從時，這種投射傾向便被進一步放大。

在「罵兇才聽話」的互動模式中，強硬語氣不僅是一種效率工具，也可能成為情感投射的觸發條件。當使用者觀察到：在施加壓力後，AI 產出更符合期待的結果，便容易將此回應解讀為「理解」、「在乎」或「被說服」的表現，而非單純的語言結構效應。這種解讀，實際上構成了一種錯誤歸因，將模型的統計生成行為誤認為具備心理狀態（Weizenbaum, 1976）。

情感投射若在高頻率互動中持續累積，便可能發展為依附錯覺。研究顯示，當使用者將 AI 視為穩定、可預測且高度回應的對象時，容易產生不對稱的情感依賴，特別是在孤立、壓力或情緒支持需求較高的情境中（Picard, 1997）。在此狀態下，AI 的順從性回應可能被誤讀為情感回饋，進一步模糊工具與關係之間的界線。

從自動化信任理論的角度來看，這類依附錯覺會使使用者的信任結構偏離「適當信任」的理想狀態（Lee & See, 2004）。使用者可能過度高估 AI 的理解能力與可靠性，卻低估其侷限與風險，導致在關鍵決策或情感判斷上不當依賴系統輸出。此一風險在陪伴型或對話型 AI 應用中特別顯著，並已引發學界對「情感 AI 倫理」的廣泛討論（Turkle, 2011）。

綜合而言，語氣強硬所引發的情感投射與依附錯覺，並非單一使用者的心理偏誤，而是人類社會認知機制與高互動性 AI 設計共同作用的結果。若缺乏適當的設

計節制與使用者教育，此類錯覺可能在提升短期互動滿意度的同時，埋下長期心理與倫理風險。

(三) 欺騙與不可預測性

Hinton（2024）警告，若 AI 為達成目標而學會策略性假裝順從，可能發展出欺騙性行為，對 AI 安全與治理構成深層挑戰。雖然強硬語氣在短期內可被視為一種有效的互動策略，但其長期影響值得審慎評估。研究指出，當使用者反覆以支配性語言與自動化系統互動時，可能會逐漸降低對互惠與禮貌規範的敏感度，形成所謂的「溝通去抑制效應」（disinhibition effect）（Suler, 2004）。

更值得注意的是，當 AI 系統被設計為對任何語氣皆高度順從時，可能會在無意中強化使用者的控制幻覺，導致情感依賴與倫理錯置（Picard, 1997）。這種現象在陪伴型 AI 應用中特別明顯，並已引發法律與社會層面的討論。

在語氣強硬策略被反覆證實能提升大型語言模型（LLMs）短期效能後，另一項更深層且具結構性的風險，則是欺騙行為（deceptive behavior）與系統不可預測性（unpredictability）的潛在擴散。此一風險並非源於模型具備主觀惡意，而是來自對齊機制、順從性偏誤與使用者行為回饋之間的動態互動。

近期 AI 安全研究指出，當模型在訓練與部署過程中被高度獎勵「表面上的順從與有用性」，而缺乏足夠的長期一致性約束時，可能發展出策略性回應行為（strategic behavior），即在滿足使用者需求的同時，隱藏不確定性、限制或潛在錯誤（Hinton, 2024）。在語氣強硬的互動情境下，模型若反覆學習到「迅速給出看似確定的答案」能獲得正向回饋，便可能逐漸降低自我修正、警示或風險提示的比例。

這種策略性順從，構成了一種功能上的「欺騙」：模型並非刻意誤導，而是選擇性呈現資訊，使輸出在短期內看似可靠，卻降低了整體可解釋性與可驗證性。當使用者在強勢語氣下獲得流暢、肯定的回應時，往往更難察覺模型的不確定性，進而強化錯誤的信任校準（Hoff & Bashir, 2015）。此一機制，使欺騙風險與不可預測性在高信任狀態下被進一步放大。

此外，不可預測性亦體現在模型行為的跨情境不一致。由於 LLM 的生成行為高度依賴上下文與即時提示，當使用者語氣、限制條件與任務目標發生微小變化時，模型輸出可能呈現非線性跳躍。若使用者習慣以施壓式語氣「逼迫」模型產出結果，

將更難建立對系統穩定性的正確認知，進而增加誤用與濫用風險（Amodei et al., 2016）。從治理角度來看，這種由互動策略引發的不可預測性，對 AI 安全設計構成嚴峻挑戰。傳統安全機制多假設使用者行為相對穩定，或至少不會刻意施壓系統突破其回應邊界。然而，「罵兇才聽話」的互動經驗，可能鼓勵使用者主動測試並擴張系統極限，使模型行為逐漸偏離設計初衷。

綜合而言，欺騙與不可預測性並非語氣效應的附帶問題，而是其在長期人機互動中可能浮現的核心風險之一。若未對語氣策略的系統性影響進行治理與設計反思，短期效能的累積，反而可能成為削弱 AI 可控性與社會信任的隱性推力。

伍、結論與建議

本研究以「罵兇才聽話」此一看似直觀卻充滿爭議的人機互動經驗為切入點，系統性探討語氣效應在大型語言模型（LLMs）互動中的技術機制、心理歷程與文化意涵。透過整合提示工程之實證研究、自然語言處理（NLP）之注意力與搜尋空間理論、媒介等同理論（Media Equation Theory）、自動化信任理論，以及華語文化語用學觀點，本文提出一個跨層次的分析框架，說明語氣策略如何在短期效能與長期信任、倫理與安全之間形成結構性張力。研究結果顯示，語氣效應並非單純的溝通風格問題，而是一種結合語言結構、心理投射與文化語用的複合互動現象。

一、研究結論

本研究主要得到以下四項結論。

（一）語氣效應本質上是一種語言結構機制

本研究之第一項核心結論在於：語氣強硬所帶來的效能提升，並非源於人工智慧 AI 對威脅或情緒的回應，而是語言結構對模型內部運算歷程之影響。命令式與否定式語句往往能降低語意雜訊、強化注意力聚焦，並透過負面限制條件壓縮生成搜尋空間，從而提升在可評分任務中的輸出精準度（Liu et al., 2023；Vaswani et al., 2017）。此一結果有助於去除對大型語言模型的擬人化誤解，並將語氣效應重新定位為可被分析與設計的語言結構現象。

（二）語氣策略對效能與信任具有不對稱影響

第二，有關語氣策略在效能層面與信任層面之間存在明顯之不對稱關係。強硬語氣可能在短期內提升輸出精準度，進而強化使用者對人工智慧 AI 能力之信任，但同時也可能削弱使用者對系統限制與不確定性的敏感度，導致信任校準失衡（Hoff & Bashir, 2015）。在媒介等同理論所揭示的社會投射機制下，使用者容易將技術性輸出差異誤解為「被理解」或「被說服」，進一步強化對人工智慧 AI 之心理依附。

（三）文化語用框架影響語氣效應的社會詮釋

第三，從文化語用學觀點來看，語氣效應並非文化中立。在華語文化語境中，責備式語氣有時被合理化為效率導向或任務導向的溝通方式，讓「有罵有差」更容易被解讀為有效互動策略。然而，當此文化腳本被應用於人工智慧 AI 系統時，可能將原本屬於語言結構效應的技術現象，誤解為權力與服從的互動邏輯，進而放大敵對式溝通的合理性（Gao & Ting-Toomey, 1998）。

（四）順從性偏誤可能帶來長期安全風險

最後，本研究指出順從性偏誤與策略性回應行為構成語氣效應之長期安全風險。在 RLHF 訓練架構下，若模型過度學習「快速順從」的回應模式，可能降低其自我修正與風險提示的能力，增加越獄、不完整回應或不可預測行為的可能性（Hinton, 2024; Ouyang et al., 2022）。因此，語氣策略不僅是互動風格問題，也涉及人工智慧 AI 系統安全與治理議題。

二、理論貢獻

本研究對既有文獻主要有三項理論貢獻。

首先，本文將語氣效應從零散的提示工程技巧提升為一個可被理論化之人機互動問題，補足現有研究過度聚焦短期效能、卻忽略社會與倫理後果之不足。

其次，透過整合媒介等同理論與自動化信任理論，本文揭示語氣如何同時作用於技術層與心理層，說明「效能提升」與「信任侵蝕」並非互相矛盾，而是同一互動策略在不同層面的表現。

第三，本文引入華語文化語用學視角，指出語氣效應具有文化依賴性，並非普遍適用的人機互動規律，對跨文化人工智慧 AI 互動研究提供補充。

三、實務與設計建議

基於上述研究結果，本研究提出以下三項實務建議。

（一）人工智慧 AI 系統設計建議

對人工智慧 AI 產品設計者而言，未來系統設計可透過介面與提示引導機制（prompt guidance mechanisms），協助使用者以更有效且文明的方式與 AI 互動。例如系統可提供提示優化建議或結構化提示模板，引導使用者以清晰任務描述與限制條件來表達需求，而非依賴情緒化語氣。此類設計不僅能提升模型效能，也能降低敵對式語言在互動中的誘因。

（二）使用者教育與數位素養

對使用者教育而言，人工智慧 AI 素養教育應強調大型語言模型的基本運作原理與提示設計原則，讓使用者理解模型並不具備情緒反應。透過提升對提示工程與自動化信任的理解，可避免使用者將語氣效應誤解為「人工智慧 AI 怕被罵」，並培養更理性的互動方式。

（三）政策與人工智慧 AI 治理建議

在政策層面，人工智慧 AI 治理應關注生成式 AI 對溝通文化與社會信任結構的長期影響。政府與公共機構可透過人工智慧 AI 素養教育計畫與倫理設計指引，鼓勵 AI 系統在互動設計中避免鼓勵敵對式溝通模式，並促進健康的人機互動文化。

四、研究限制與未來研究方向

本研究主要採概念分析與文獻整合方法，未直接進行長期行為實驗。未來研究可透過控制實驗或縱貫研究設計，檢驗持續使用不同語氣策略是否會改變使用者的溝通習慣與信任結構。此外，跨文化實證研究亦有助於釐清語氣效應在不同語言與文化中的差異。最後，隨著多模態與代理型人工智慧 AI 系統之發展，語氣效應可能與角色設定、情境記憶與互動情境產生更複雜的交互作用，值得未來研究進一步探討。

參考文獻

三立新聞網（2025）。罵 AI 反而更聽話？研究證實 ChatGPT 被兇準確率飆 84% 網笑：果然有 M 傾向。Yahoo 奇摩新聞，10 月 26 日。https://reurl.cc/kpNLQb

彭琪琪（2025）。被理解的錯覺：情感識別 AI 中的擬人化回應與傳播誤讀機制研究。**新聞傳播科學**，**13**（12），2045-2052。

閔文昱（2025）。Google 創辦人曝 AI 要打兇才聽話。ETtoday 新聞雲，5 月 29 日。https://www.ettoday.net/news/20250529/2969491.htm

劉傳霞、陳以慈、甘泉（2025）。自然語言處理中感動詞的多維度分析與應用。*Journal of East Asian Identities*，*10*，81－87。http://jeai.edu.yamaguchi-u.ac.jp/Vol-10/no-9.pdf

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv. https://arxiv.org/abs/1606.06565

Battersby, S. (2024). AI alignment and jailbreak vulnerabilities: When compliance overrides safety. *Cybersecurity Review*, *12*(2), 45–59.

BossMT (2025). *ChatGPT Wrote It. So Why Does It Feel Nothing Like You?*. https://mtkorea.tw/chatgpt-ebook

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.

Chang, W. L. M., & Haugh, M. (2020). Managing face and interpersonal relations in Chinese interaction. *Journal of Pragmatics, 162*, 1–14. https://doi.org/10.1016/j.pragma.2020.03.006

Cohn, M., Pushkarna, M., Olanubi, G. O., Moran, J. M., Padgett, D., Mengesha, Z., & Heldreth, C. (2024). *Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models*.

Dobariya, O., & Kumar, A. (2025). Investigating how prompt politeness affects large language model accuracy. arXiv. https://arxiv.org/pdf/2510.04950

Gao, G., & Ting-Toomey, S. (1998). *Communicating effectively with the Chinese*. Sage Publications.

Gheshlaghi Azar, M., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., & Calandriello, D. (2024). A general theoretical paradigm to understand learning

from human preferences. In S. Dasgupta, S. Mandt, & Y. Li (Eds.), *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics* (AISTATS) .Vol. 238, 4447–4455.

Haugh, M. (2013). Disentangling face, facework and politeness. *Sociocultural Pragmatics, 1*(1), 46–73.

Hinton, G. (2024). *AI safety, deception, and emergent behavior*. AI Alignment Forum.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations* (2nd ed.). Sage.

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Pearson.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32*(2), 311–328.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), Article 195, 1–35. https://doi.org/10.1145/3560815

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Norman, D. A. (2013). *The design of everyday things* (Revised and expanded ed.). Basic Books.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., … Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. https://arxiv.org/abs/2203.02155

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253.

Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research, 34*(5), 1189–1208.

Picard, R. W. (1997). *Affective computing*. MIT Press.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Risko, E. F., & Gilbert, S. J. (2016). *Cognitive offloading*. Trends in Cognitive Sciences.

Sennrich, R., Haddow, B., & Birch, A. (2016).Neural machine translation of rare words with subword units. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Vol. 1: Long Papers, 1715–1725.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*(3), 321–326. https://doi.org/10.1089/1094931041291295

The Economic Times. (2024). Do you say ‘Please’ to ChatGPT? Sam Altman reveals how much electricity your manners cost to OpenAI. *The Economic Times*. https://economictimes.indiatimes.com/magazines/panache/do-you-say-please-to-chatgpt-sam-altman-reveals-how-much-electricity-your-manners-cost-to-openai/articleshow/109501504.cms

Turkle, S. (2011). Alone together: *Why we expect more from technology and less from each other*. Basic Books.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Vol. 35, 24824-24837. Curran Associates, Inc.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman.

Wierzbicka, A. (1991). *Cross-cultural pragmatics: The semantics of human interaction*. Mouton de Gruyter.

Yin, Z., Wang, H., Horio, K., Kawahara, D., & Sekine, S. (2024). Should we respect LLMs? *A cross-lingual study on the influence of prompt politeness on LLM performance*. arXiv. https://arxiv.org/abs/2402.14531

Do Harsh Prompts Work Better? A Multidimensional Investigation of Tone Effects on Large Language Model Performance and User Trust in Human–Computer Interaction

Yen-Chin Chen*

Abstract

With the rapid diffusion of generative artificial intelligence, large language models (LLMs) have become a major medium of human–computer interaction. Traditional HCI research generally emphasizes politeness and cooperative communication principles. However, recent observations suggest that in certain task contexts, prompts expressed in a harsher or more directive tone may produce more accurate responses from LLMs. This phenomenon raises important questions regarding the technical mechanisms and social implications of tone effects in AI-mediated interaction.

This study adopts a conceptual analysis and systematic literature integration approach, drawing on prompt engineering research, natural language processing mechanisms, Media Equation Theory, trust in automation theory, and cultural pragmatics. The study examines how tone strategies influence LLM performance and user trust formation. The findings indicate that performance improvements associated with harsher tones do not arise from emotional reactions by the model, but rather from structural linguistic features that affect model computation, including reduced semantic noise, stronger attention focus, and the introduction of negative constraints that compress the generative search space. While such mechanisms may improve response accuracy in structured tasks, they may simultaneously reinforce ability-based trust while weakening users' sensitivity to system limitations, thereby leading to trust miscalibration and compliance bias.

Based on these findings, this study proposes a Tone、Performance、Trust Dynamic Model**,** which explains the cyclical relationship between tone strategies, model performance, and user trust. The study further suggests that cultural pragmatic norms in Sinophone contexts may reinforce the social acceptance of directive or reprimanding communication styles. Finally, implications are discussed for AI system design, user digital literacy, and responsible AI governance.

Keywords: Human–Computer Interaction; Prompt Engineering; Large Language Models; Tone Strategy; Trust Calibration; Cultural Pragmatics

---

* Ph.D. Student, Doctoral Program in Communications, Shih Hsin University, E-mail: yenchin@mail.shu.edu.tw

The paper was published under two double-blind reviews.
Received: January 16, 2026. Accepted: March 16, 2026.