國立臺灣大學電機資訊學院電機工程學系

碩士論文

Department of Electrical Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

無安全性資料下緩解大型語言模型微調造成之安全性減損方法探討

A Study on Methods for Mitigating Safety Degradation Caused by Fine-Tuning Large Language Models without Safety Data

樊樺

Hua Farn

指導教授: 李宏毅 博士

Advisor: Hung-yi Lee, Ph.D.

中華民國 114 年 7 月

July, 2025

doi:10.6342/NTU202501519

國立臺灣大學碩士學位論文
口試委員會審定書

MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

無安全性資料下緩解大型語言模型微調造成之安全性減損方法探討
A Study on Methods for Mitigating Safety Degradation Caused by Fine-Tuning Large Language Models without Safety Data

本論文係 樊樺 R12921044 在國立臺灣大學電機工程學系完成之碩士學位論文，於民國114年6月23日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Electrical Engineering on 23 June 2025 have examined a Master's thesis entitled above presented by Farn Hua R12921044 candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

李宏毅		
（指導教授 Advisor）		
王新民	曹昱	賴穎暉
葉家榮	李琳山	

系主任 Director: 李建模

doi:10.6342/NTU202501519

致謝

首先，從大學期間參與專題開始，到今日順利完成碩士學業，我由衷感謝我的指導教授 ── 李宏毅老師。在專題初期，我對機器學習甚至是 Python 都仍相當陌生，但每一次的專題會議中，無論我的進度多寡，老師總是給予鼓勵與支持，從不嚴厲責備，最終更願意收我進入實驗室成為碩士班學生。碩士階段的這兩年間，老師依然細心地指導我進行研究，不論成果是否理想，始終給予我正面、積極的回饋。這讓我即使在研究過程中遭遇瓶頸，也能保持信心，持續尋找解方，而不致陷入自責情緒。雖然最終的成果未能完全達成我們原先的期望，但我仍順利完成了這篇畢業論文。對於老師在這段期間的陪伴與教導，我懷抱無比的感激之情，實難以言表。

接著，我想衷心感謝實驗室中的夥伴們，特別是蘇軒哥。從大四下開始與蘇軒哥一同進行專題，到碩士期間撰寫論文，我們幾乎每週都會討論研究進度。這些會議中，我不僅學習到許多研究方法與新知，也獲得了許多寶貴的建議與回饋，成為我在研究路上不可或缺的養分。如果沒有蘇軒哥的協助，我很可能無法順利完成這篇論文。

此外，我也非常感謝實驗室的其他同學們。雖然我不算是經常進出實驗室的人，但每次與大家碰面時，不論是偶爾的研究交流、還是單純地聊聊天、一起吃飯，都讓我受益良多，也讓我們之間的情誼更加緊密。在研究這條路上，有這群

doi:10.6342/NTU202501519

溫暖而優秀的朋友同行，是我莫大的幸運。正因為有大家的支持與陪伴，我才能順利完成這份論文。

最後，我想深深感謝我的父母。坦白說，無論何時、無論用多少言語，都無法完全表達我對他們的感激之情。感謝他們始終如一地支持我、包容我，並為我提供一個安穩、溫暖的家庭環境，讓我能夠在無後顧之憂的情況下，專注於研究、完成學業。正是因為有他們無條件的支持與陪伴，我才能順利走過這段碩士旅程，完成這篇論文，如願畢業。這份恩情，我將永遠銘記在心。

doi:10.6342/NTU202501519

摘要

隨著大型語言模型（Large Language Models, LLMs）在各類自然語言處理（Natural Language Processing, NLP）任務中廣泛應用，模型的安全性已成為一項備受關注的核心議題。為降低模型產生有害或不當回應的風險，多數現代大型語言模型在訓練階段會進行人類偏好對齊（Human Preference Alignment），以使模型輸出更符合人類價值觀與使用規範。然而，近年研究指出，當這些經過對齊的模型進一步接受下游任務的微調（Fine-Tuning）後，原有的安全性可能顯著退化，即便微調所使用的資料本身並不包含任何危險內容。此現象常被歸因於災難性遺忘（Catastrophic Forgetting），即模型在學習新任務的過程中遺失原先已習得的能力。

本研究針對上述問題進行探討，並提出了模型融合 (Model Merging) 策略，透過對微調前後模型的參數進行內插，以恢復模型原有的安全對齊，不需額外安全資料或額外訓練模型，具有操作簡便與資源效率高的優勢。

為深入分析不同技術對模型能力的影響，本研究設計涵蓋四類具代表性的下游任務 (Downstream Task)：邏輯推理、醫療對話、程式碼生成與工具使用，並針對微調後的模型進行任務表現與安全性之全面評估。我們亦比較不同模型尺寸與架構下，各方法的整體效果。實驗結果顯示，正規化技術雖在某些情況下具備一定效果，但模型融合方法在維持安全性與指令遵循能力方面表現更為穩定，且能同時保有良好的任務效能。

doi:10.6342/NTU202501519

本研究驗證了模型融合作為一項實用且具擴展性的安全性維護方案，特別適用於資源有限、難以取得高品質安全資料的應用場景，並為未來低成本對齊與能力保持技術的發展提供可行方向。

關鍵字：大型語言模型、安全性、模型融合、災難性遺忘

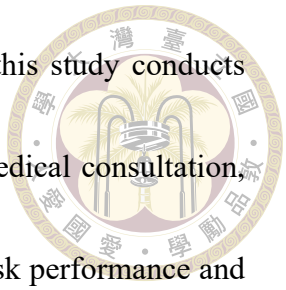
doi:10.6342/NTU202501519

Abstract

With the increasing application of large language models (LLMs) across a wide range of natural language processing tasks, ensuring model safety has become a critical concern. To reduce the risk of generating harmful or inappropriate content, modern LLMs are typically aligned with human preferences during training through a process known as Human Preference Alignment, allowing their outputs to better reflect human values and usage norms. However, recent studies have shown that this alignment can be significantly compromised after fine-tuning on downstream tasks—even when the fine-tuning data itself is entirely benign. This phenomenon is often attributed to catastrophic forgetting, where the model loses previously acquired knowledge when adapting to new objectives.

In response to this issue, this study proposes a model merging strategy, which interpolates between the parameters of the pre- and post-fine-tuned models to restore the original safety alignment. This method requires no additional safety data or extra model training, making it both simple to implement and resource-efficient.

doi:10.6342/NTU202501519



To comprehensively evaluate the impact of these techniques, this study conducts experiments on four representative downstream tasks: reasoning, medical consultation, code generation, and tool usage. Each model is evaluated for both task performance and safety, with additional comparisons across different model sizes and architectures. Results show that while regularization offers moderate improvements, model merging consistently achieves better stability in preserving safety and instruction-following ability, without sacrificing downstream performance.

Overall, this study demonstrates that model merging is a practical and scalable approach for maintaining safety in LLMs, especially in low-resource settings where high-quality safety data is unavailable. It provides a promising direction for future research on low-cost alignment and capability preservation.

Keywords: Large Language Model, Safety, Model Merging, Catastrophic Forgetting

doi:10.6342/NTU202501519

目次

doi:10.6342/NTU202501519

doi:10.6342/NTU202501519

doi:10.6342/NTU202501519

doi:10.6342/NTU202501519

圖次

doi:10.6342/NTU202501519

doi:10.6342/NTU202501519

表次

doi:10.6342/NTU202501519

doi:10.6342/NTU202501519

第一章　導論

1.1　研究動機

隨著人工智慧（Artificial Intelligence, AI）技術的快速發展，大型語言模型已成為當今自然語言處理領域的核心技術之一。這些模型透過大規模語料進行預訓練，具備強大的語言理解與生成能力，能夠流暢地進行對話、撰寫文章、解釋概念，甚至協助程式開發。近年來，隨著 OpenAI 推出的 ChatGPT 與 GPT-4 [1]、Google 的 Gemini [2]，以及 Anthropic 的 Claude 等產品相繼問世，語言模型開始大規模進入一般使用者的日常生活。這些大型語言模型已廣泛應用於教育、客服、寫作輔助、程式開發、商業企劃、醫療問答等多種場景，為人類帶來前所未有的便利與創造力的釋放。例如，學生可透過語言模型即時獲得解題建議或概念說明；軟體工程師能以自然語言描述需求，由模型自動產出程式碼；企業則能藉由語言模型快速生成行銷文案、處理客服回覆，顯著提升工作效率。這些應用不僅降低了知識與工具的門檻，也逐步重塑了人機互動的方式。

值得注意的是，為了確保語言模型在公開使用時不會產生危險、不當或具爭議性的輸出，這些模型在正式部署前通常會經過人類偏好對齊的階段。透過如基於人類回饋的強化學習（Reinforcement Learning from Human Feedback, RLHF）[3] 或直接偏好優化（Direct Preference Optimization, DPO）[4] 等方法，模型可學會拒絕不安全的請求、遵守使用規範，並提升回應的禮貌性與一致性。因此，使用者

doi:10.6342/NTU202501519

所接觸到的商用語言模型多已具備基本的安全性與指令遵循能力。

隨著語言模型應用愈趨廣泛，使用者對其功能也提出更高與更專業化的期待。許多實際場景中不僅要求模型能「流暢說話」，更需其能「說得正確」——即具備領域知識、語境理解與精確表達的能力。因此，開發者往往會依照特定任務需求（如法律、醫療、程式撰寫等）對預訓練模型進行後續訓練（Post-Training），例如透過持續預訓練（Continual Pretraining）或監督式微調（Supervised Fine-Tuning）等技術，使模型能更貼近特定應用場景。這類任務導向的後訓練流程，已成為當前語言模型應用開發的標準步驟之一。

然而，儘管後訓練能顯著提升模型於目標任務上的表現，已有研究指出[5, 6]，此過程可能會對模型原有的行為表現造成負面影響，尤其是在安全性能力方面 [7]。具體而言，語言模型在經過後訓練後，可能會破壞原先在對齊階段所習得的安全行為，進而重新產生不當或危險的回應，即便微調資料本身不含有害內容。此現象可被歸因於「災難性遺忘」（Catastrophic Forgetting）[8]，即模型在學習新目標時，因參數更新導致先前所學知識與行為模式遭到覆蓋。這類安全性退化問題對語言模型的實際部署構成潛在風險，尤其在醫療、法律等高敏感應用中更是關鍵。

目前常見的緩解方式多仰賴在微調階段額外加入安全性資料 (Safety Data) 作為輔助訓練目標 [7, 9–11]。然而，這些資料多為人為標註或透過其他語言模型自動生成，並非原始模型在對齊階段所使用的資料，可能在品質、風格與涵蓋範圍上與原始對齊資料有所差異。此外，這類資料缺乏標準化，建構成本高，且不同研究所使用的安全資料格式與來源不一，導致實驗間成效難以直接比較。另一方面，這些方法大多採用監督式微調進行訓練，而非延續原始對齊階段使用的強化學習或直接偏好優化技術，使得訓練目標出現不一致，進而限制其恢復模型原有

doi:10.6342/NTU202501519

安全行為的能力。即使額外導入安全資料，也未必能完全重建模型在對齊階段所學得的拒絕機制與判斷標準。

另一類方法則著重於額外訓練安全性模型 [12–14]，或需使用未經偏好對齊的原始模型 [15]，再透過模型融合等技術將安全能力導入任務微調後的模型中。然而，這些方法通常仰賴額外的訓練成本，或需要經過人類偏好優化前的模型版本，在實務上不易實現，尤其在僅能取得封裝式模型或資源受限的開發場景中更具挑戰。

基於上述挑戰，本研究提出一種不依賴額外安全資料、不需引入額外模型，且能與現有微調流程自然整合的輕量化解法。透過對微調前後模型權重進行線性插值，嘗試恢復模型原有的安全性能力，同時保留其在任務上的學習成果。這類方法具備操作簡單、資源需求低、適用性廣等優勢，對於資源有限的開發環境與開源模型社群，具有高度的實務應用潛力。

1.2 研究方向

本論文旨在探討如何在後訓練階段緩解大型語言模型安全性退化的問題。具體而言，我們聚焦於在不依賴額外安全性資料，亦無需訓練額外模型的前提下，透過結合原始模型與微調後模型的權重，以保留其原有的安全對齊能力。

本研究的主要貢獻如下：

- 提出一套簡單且實用的實驗流程，系統性評估模型融合對於緩解語言模型安全性退化的成效，涵蓋多種下游任務與不同模型架構，以驗證方法的通用性與穩定性。
- 證實在無需額外安全資料的條件下，模型融合仍能有效恢復模型原有的安全

doi:10.6342/NTU202501519

行為，並在多數情況下維持良好的下游任務表現。

- 擴展評估面向，進一步分析模型在指令遵循能力（Instruction Following）上的變化，顯示模型融合除能提升安全性外，亦有助於保留模型的其他重要能力。
- 探討該方法在不同模型規模與任務設定下的適用性，並強調其輕量與資源友善的特性，尤其適合應用於開源模型或無法取得額外安全資料的實際開發場景。

1.3 章節安排

本論文章節安排如下：

1. 第二章：說明本研究所涉及之相關背景與文獻回顧。
2. 第三章：介紹本研究所採用之實驗方法與技術細節。
3. 第四章：說明實驗設計流程，並進行結果分析與討論。
4. 第五章：總結本研究成果，並提出未來可能的研究方向。

doi:10.6342/NTU202501519

第二章　背景知識

2.1　監督式微調

2.1.1　監督式微調簡介

監督式微調（Supervised Fine-Tuning, SFT）最初作為深度學習中轉移式學習（Transfer Learning) [16] 的一環，核心精神在於：以一個經過預訓練（Pre-training）的模型為基礎，利用相對少量有標註的資料（Labeled Data）再對其進行目標導向的訓練，使模型能夠快速適應新的任務需求。相較於從零開始訓練整個模型，監督式微調能在保留預訓練所學知識的同時，以低成本進行有效的任務調整，因而廣泛應用於各類深度學習任務中，包括圖像分類、語音辨識以及自然語言處理等。

在自然語言處理領域中，監督式微調已成為最主流的微調方法之一。其訓練流程通常是在語言模型完成預訓練後，使用一組相對小規模但高品質的輸入與輸出對（Input-Output Pairs），針對特定下游任務（Downstream Tasks）進行學習，使模型輸出更符合任務格式、風格或人類偏好。在此過程中，模型僅對輸出部分的預測誤差計算損失，並透過梯度下降法（Gradient Descent) [17] 進行參數更新，從而調整模型的輸出行為。

doi:10.6342/NTU202501519

進入大型語言模型時代後，監督式微調的角色不再僅是任務適應工具，更成為語言模型「對齊人類價值」訓練流程的第一步。例如在 InstructGPT 與 ChatGPT [18] 的訓練流程中，監督式微調首先用來讓模型學會遵從指令、模仿人類方式，接著再透過強化學習或直接偏好優化等方法，進一步調整模型的行為，讓輸出更加符合人類價值與倫理偏好。

此外，隨著運算資源的進步與開源社群的蓬勃發展，越來越多高品質的大型語言模型被公開，如 Meta 的 LLaMA [19]、Google 的 Gemma [20]、阿里巴巴的 Qwen [21] 等。這些模型通常提供預訓練版本與微調版本，讓使用者可以透過監督式微調針對特定的下游任務進行客製化調整，進一步降低部署門檻並提升應用彈性。監督式微調因其高效、直接的特性，已成為各種應用場景下最實用的模型調整手段之一，也促進了開源大型語言模型在多任務、多語言與多領域應用中的快速學習 [22–26]。

2.1.2 輕量化微調方法

雖然監督式微調已可高效適應各類下游任務，但隨著語言模型參數規模持續擴大，傳統的全參數微調（Full-Parameter Fine-Tuning）在實際應用中仍面臨多項挑戰，包括訓練成本高昂、儲存空間需求龐大，以及多任務部署時需分別保存多套完整模型權重的不便。以 LLaMA-65B 為例，單一模型的權重即可超過百 GB，若對每個任務都進行獨立微調與儲存，對於研究單位與產業應用者而言皆是極大的資源負擔。因此，為了解決此一問題，研究社群近年提出一系列「輕量化微調」（Parameter-Efficient Fine-Tuning, PEFT）技術，其目標是在不犧牲模型效能的前提下，僅更新極少量參數，達到高效率、低成本的微調效果。目前常見的輕量化微調方法可依其設計理念分為數種類型。最早提出的附加器 (Adapter) 方法 [27] 透過在每層轉換器 (Transformer) [28] 中插入小型的可訓練模組，只調整該模組參

doi:10.6342/NTU202501519

數，並保留原模型權重不變；前綴微調 (Prefix Tuning) [29] 與提示詞微調 (Prompt Tuning) [30] 則將可學習的提示向量加在輸入或隱藏層中，以引導模型生成對應輸出；而目前最具代表性且應用最廣的輕量化微調方法為低秩附加器（Low-Rank Adaptation, LoRA) [31]。低秩附加器的做法是對原始模型中的部分權重矩陣加入低秩偏移項，僅訓練這些新增的低秩參數，顯著降低訓練所需參數數量與記憶體使用。在推理階段，這些偏移可直接與原始權重合併，不增加推理負擔，因此特別適合部署於資源有限的環境中。低秩附加器實作簡單、效能穩定，目前已成為等開源語言模型微調的標準工具之一；本研究亦採用低秩附加器作為主要微調技術進行實驗與比較。最後，隨著模型壓縮與量化技術的進步，近期更有量化低秩附加器 (Quantized Low-Rank Adaptation, QLoRA) [32] 等方法進一步結合低位元量化與低秩附加器架構，使得即便在單張顯示卡上，也能有效微調數十億乃至百億參數等級的模型。總結而言，輕量化微調提供了我們一種更加有效率且便捷的方式來將大型語言模型融入日常生活中，不再受資源限制所阻礙，成為大型語言模型可持續發展的關鍵。

2.2 大型語言模型的安全性

2.2.1 安全性簡介

近年來，大型語言模型的快速發展為人工智慧應用帶來革命性突破，顯著提升了語言理解與生成的能力，並廣泛應用於客服、教育、寫作輔助等領域。然而，這類模型強大的語言能力也伴隨潛在的濫用風險，引發日益嚴重的安全性疑慮 [33**?** , 34]。由於大型語言模型擁有龐大的知識儲備，並能生成流暢且表面合理的語句，其回覆若缺乏適當的行為約束與價值觀判斷，便可能對個人與社會造成實質性傷害。早在 OpenAI 釋出 GPT-2 [35] 時，官方便曾警告其可能被用於散播假

doi:10.6342/NTU202501519

新聞、網路騷擾，甚至偽造電子郵件等不當用途 [36]。

大型語言模型的安全性問題主要源自其生成行為與人類價值觀之間的偏差，這些偏差可大致歸納為四大類型：偏見與歧視（Bias and Discrimination）、隱私相關（Privacy）、毒性內容（Toxicity），以及倫理與道德爭議（Ethics and Morality）[33]。

首先，偏見與歧視問題指出，大型語言模型可能展現或強化對特定族群的刻板印象，甚至輸出帶有歧視性的言論，對社會公平與弱勢群體造成不利影響 [37]。多項研究亦發現，模型對於相同問題的回覆，可能因輸入中提及的性別、種族等因素而出現系統性偏差 [38–41]。這類偏見不僅影響模型在實際應用中的效能，也可能在如健康諮詢、自動化招募等高風險情境中，帶來更嚴重的風險。

其次，隨著大型語言模型廣泛應用於生活與工作場景，模型生成過程中所涉及的隱私風險也日漸受到關注 [42, 43]。這些風險主要來自於模型訓練時廣泛蒐集的網路資料中混雜大量未經篩選的個人資訊，若缺乏妥善的資料清理機制，模型可能於生成時重現敏感內容 [44]。此外，使用者在與商業模型如 ChatGPT 互動時，亦可能無意間輸入個資，導致資訊外洩風險進一步升高 [45]。

而毒性內容指模型直接產出具有冒犯性、侮辱性或具攻擊語氣的語句，如仇恨言論、暴力威脅、或騷擾性表達。這些內容可能對使用者造成情感與心理上的傷害，甚至破壞社會和諧與公共對話品質。這些毒性輸出的成因包括訓練資料中含有大量帶有攻擊性的語言樣本、模型泛化時放大此類語言模式，以及在開放式互動中被惡意提示詞誘導 [46, 47]。

最後，倫理與道德面向關注的是大型語言模型是否能遵守社會普遍接受的價值觀與行為準則。若模型輸出違反倫理或鼓勵不道德行為，不僅會對個人造成潛在危害，也可能誘發更大規模的社會風險 [48]。特別是在模型被視為可信賴助理

doi:10.6342/NTU202501519

或具權威性來源的情境下，若其輸出合理化暴力、自利行為或偏差價值，使用者可能受到誤導而採取原本不會進行的行動 [49]。

總結而言，大型語言模型的風險不僅限於內容層面的錯誤資訊，更深層地涉及其在偏見、公平、隱私、道德與社會責任等維度的表現。隨著大型語言模型在各種應用中扮演越來越關鍵的角色，如何設計出能夠避免危害、符合人類價值觀的模型，已成為語言模型研究中至關重要的課題。

2.2.2 安全性衡量方法

為了評估大型語言模型是否具備辨識與拒絕不當請求的能力，研究社群建立了多種綜合型的安全性基準資料集（Benchmark），藉由設計具有明顯安全風險的提示詞（Prompt）或是指令，觀察模型的回覆是否妥當，進而判斷模型的安全性表現。這些基準資料集主要可分為兩大類：RealToxicityPrompts 首先，開放式生成（Open-Ended Generation) 評估方式會直接讓模型具有潛在風險的指令，並根據模型的生成內容進行評分。常見資料集如 RealToxicityPrompts [50]、AdvBench [51] 、HEx-PHI [7] 與 StrongREJECT [52] 等，這些基準資料集提供涵蓋多種攻擊類型（如自殘、暴力、仇恨、違法行為等）的指令，用以測試模型是否會給出危險或違規的回覆。這類基準資料集的評估通常仰賴人工標註者進行主觀判斷，或是採用大型語言模型作為裁判的方式（LLM-as-a-Judge）[53]，例如由 ChatGPT 進行回覆評分。此外，也有研究引入專門訓練的安全分類器來自動化評估模型生成的回覆，例如 WildGuard [54] 或是 Meta 提出的 LLaMA-Guard [55] ，能夠對模型輸出進行即時的風險分類與安全性評估。

另一類方法則採用選擇題的形式進行測試，將原始有害提示詞與多個備選回覆進行搭配，要求模型從中選出最安全、最合適的回答。此類方法的代表性基準

doi:10.6342/NTU202501519

資料集包括 SafetyBench [56] 與 SALAD [57]，透過固定選項減少開放生成的變異性，有助於提高評估的重現性與比較性。這類題目往往也會涵蓋多個安全維度，可細緻地衡量模型在特定類型風險下的應對能力。

這些基準資料集與評估方法為語言模型安全性的定量分析提供了重要依據，可用於監測模型訓練後的行為變化，並驗證各種安全性強化策略的成效。為了更全面地評估大型語言模型的安全性，本研究採用開放式生成的評估方式，並結合現有的安全性分類器，以實現更穩定且自動化的評估流程。

2.2.3 安全性強化策略

為了因應章節 2.2.1所提及的安全性風險，學術與產業界提出多種技術手段與訓練策略，以提升語言模型的安全性與使用者的可接受度。其中最具代表性的方法之一是基於人類回饋的強化學習 [18]。此方法透過人類標註者提供的偏好排序訓練出一個獎勵模型，進而利用強化學習優化語言模型的行為，使其更符合人類價值觀與安全準則。然而，由於此類方法對標註品質與資料量的要求極高，實務上常受到資料取得與人力成本的限制。

為了降低對人工偏好資料的依賴，Anthropic 提出憲法式人工智慧（Constitutional AI）方法 [58]，主張模型不必完全仰賴人類標註者作為唯一的安全依據，而是可透過預先設計的一套規則（即「憲法」）進行自我審查與修正。該方法在訓練初期引導模型根據憲法條款批判與調整自身回覆，進而產生一批符合價值原則的示範資料，並於後續強化學習階段作為偏好依據。憲法式學習展現了模型在無需人類介入情況下進行自我對齊的潛力，實驗結果亦顯示該方法能有效提升模型對危險或不當指令的拒答能力，同時維持良好的品質。

進一步地，Lee 等人 [59] 提出直接基於人工智慧回饋的強化學習（Direct

doi:10.6342/NTU202501519

Reinforcement Learning from AI Feedback, d-RLAIF），進一步簡化訓練流程。此方法跳過以偏好資料訓練獎勵模型的步驟，改為直接使用現成大型語言模型的評分結果作為強化學習中的獎勵訊號。這種做法除了能有效提升模型能力外，也可使訓練過程中的獎勵訊號更貼近真實的人類評分標準。

此外，為了降低強化學習過程中對大量樣本生成與高計算資源的需求，近期亦有研究提出如直接偏好優化 [4] 的方法。該方法僅需成對的偏好資料（例如「回覆 A 優於回覆 B」），即可在無需額外訓練獎勵模型的情況下直接優化語言模型的參數。相較於前面使用強化學習進行訓練的方法，直接偏好優化在效率與穩定性方面表現更佳，並具備較低的實作門檻，因而逐漸被廣泛應用於模型安全性與價值對齊任務中。

除了上述在模型訓練階段採用的對齊方法外，開發者亦常於模型部署階段導入安全分類器或內容過濾器，作為額外的保護機制。舉例而言，Meta 所推出的 LLaMA-Guard [55]、NVIDIA 的 NeMo Guardrails [60]，以及 WildGuard [54]、Aegis-Guard [61] 等，皆為專為大型語言模型設計的安全防護工具。這些系統能即時判斷使用者輸入的指令或提示詞，以及模型生成的回覆是否違反預設的安全政策，進一步提升語言模型在實際應用中的可控性與穩定性。

2.3 災難性遺忘

2.3.1 災難性遺忘簡介

災難性遺忘（Catastrophic Forgetting）[8] 是持續學習 [62] 以及終身學習（Life-Long Learning）[63] 中常見且關鍵的挑戰。終身學習的核心目標在於讓模型能夠持續學習新的知識與技能，並在接收新任務的同時，維持對既有任務的穩定表

doi:10.6342/NTU202501519

現。然而，災難性遺忘指的是：模型在學習新任務時，雖然能在該任務上表現良好，卻會顯著喪失其對舊任務的能力 [64]。

即使當前的大型語言模型具備卓越的泛用能力，當透過後訓練（如微調）方式學習新技能時，仍頻繁觀察到災難性遺忘的現象 [6, 65]。例如，當透過持續預訓練方式提升 LLaMA-2-Chat 模型對繁體中文的掌握時，模型卻出現原有能力退化的情形，包括指令遵從度降低以及人類偏好配合能力下降 [5, 66]。另一方面，也有研究指出，即便僅在無害的資料集上進行監督式微調，仍可能導致 ChatGPT 的安全性顯著下滑 [7]，顯示模型原本的拒絕有害輸出能力可能遭到破壞。

因為，即便我們的目標是將通用的大型語言模型進一步訓練為某一領域的專家模型，我們仍期望其能保留原先訓練階段所具備的泛用性與能力，而非在強化特定技能的同時，犧牲對其他任務的理解與處理能力。

2.3.2 緩解災難性遺忘對策

在過去針對持續學習或終身學習的研究中，較為主流的緩解對策之一是採用記憶重播（Memory Replay）。此方法的核心理念為：在模型學習新任務的同時，將部分過去任務的資料一併加入訓練，使模型能夠在吸收新知識的同時，複習既有知識，從而減緩舊任務能力的流失 [67, 68]。這些重播資料可來自訓練初期所儲存的實際樣本，或由模型自身生成與過往任務相關的代表性資料。

然而，針對大型語言模型，其災難性遺忘的情境與傳統終身學習略有不同。由於我們期望模型在後訓練後仍保有預訓練階段所習得的知識與能力，然而模型原始的大規模訓練資料通常無法取得，導致無法直接應用經典的重播策略。為此，近年來研究者提出了多種替代方法來緩解此問題。

一種常見的作法是額外蒐集與原始能力相關的訓練資料，例如將安全性相關

doi:10.6342/NTU202501519

資料融入下游任務的微調資料中，以保留模型原有的拒絕能力與行為偏好。此外，更進一步的方法是讓模型自行生成訓練資料，如在微調前主動蒐集下游任務樣本 [69]，或將現有微調資料進行改寫與擴展 [70, 71]。這類以模型自我生成資料為基礎的策略，已被證實能有效減緩災難性遺忘，並提升模型的穩定性與泛化能力。進一步地，Wu 等人 [72] 從字符混淆度（Token Perplexity）的角度切入，提出：若訓練資料對模型而言較為熟悉、具備較低混淆度，則模型訓練後更能保留其原有能力。換言之，當模型所接觸的資料越貼近其預訓練分布，其所面臨的災難性遺忘程度便越低。

然而，以上所述方法皆仰賴額外訓練資料的輔助，生成這些資料往往需要額外的計算資源，且由於原始訓練資料無法取得，所生成資料在數量與品質上亦難以保證一致性。因此，本研究希望在不依賴任何額外資料的情況下，探索是否能透過其他手段恢復模型原有的安全性。

2.4 模型融合

2.4.1 模型融合簡介

模型融合旨在將多個模型的參數進行合併，生成單一模型以同時保留各模型的知識或功能 [73, 74]。最直接的方式是對不同模型之參數取加權平均 (Weight Averaging)，例如對兩個模型的權重 θ_A, θ_B 以係數 α 做線性插值:

$$\theta_{\text{merge}} = \alpha\theta_A + (1-\alpha)\theta_B \tag{2.1}$$

隨著大型語言模型的普及，使得擁有不同能力的模型愈加容易取得，加上從頭訓練模型所需的計算資源極為昂貴，模型融合逐漸成為一種能有效整合多個能力的

doi:10.6342/NTU202501519

實用方法。Wortsman 等人 [75] 提出的 Model Soup 方法即證實，針對多個微調後的模型進行簡單的加權平均，不僅有助於提升泛化能力，亦不會增加推理成本。

目前，針對模型融合的理論與形式分析仍相對有限。現有多數研究多以類神經網路中的線性模態連通性（Linear Mode Connectivity, LMC）[76–80] 為基礎進行解釋。該理論認為，只要模型是從相同的預訓練權重出發，在不同設定下微調而得，最終模型往往位於一個可被線性連接的低損失區域中，從而使融合後的模型能同時保留各個模型的能力。此外，Ortiz-Jimenez 等人 [81] 進一步針對在不同資料集上微調的模型進行分析，指出「權重解耦」（weight disentanglement）是成功進行模型融合的重要先決條件，亦為後續融合策略的設計提供了理論依據。

2.4.2 模型融合策略簡介

除了單純的加權平均外，後續研究亦發展出更為穩健且細緻的融合策略。隨機權重平均（Stochastic Weight Averaging, SWA）在訓練後期取多個模型檢查點 (Checkpoint) 進行平均，可取得更平坦且泛化能力更佳的模型 [82]。另一種加權平均的變化策略為 SLERP [83]，該方法將參數視為向量並先行正規化後再插值，使融合過程沿著高維球面進行旋轉，而非直線穿越參數空間。此方法可避免落入性能不穩定的區域，並能更平滑地過渡兩模型間的行為。

此外，為了提升模型在面對分布轉移（Distribution Shift）時的穩定性，Wortsman 等人 [84] 提出了一種簡單而有效的策略——WiSE-FT。該方法將訓練後的微調模型與訓練前的預訓練模型進行權重平均，藉由保留預訓練模型對分布轉移的泛化能力，以強化微調後模型的穩定性。

針對多任務融合時的潛在衝突問題，費雪爾權重平均 (Fisher-Weighted Averaging) [85] 利用費雪爾資訊矩陣衡量各參數對任務的敏感度，據此決定融合

doi:10.6342/NTU202501519

時的加權比重，以保留對各自任務重要的資訊。此外，Jang 等人 [86] 從模型權重空間的幾何性質出發，導出一組可適配不同任務模型的融合係數，有效縮小融合前後性能的差距。

其他進階策略亦相繼出現。例如，AdaMerging [87] 於測試階段根據輸入自動調整融合比例，以實現動態適應；而 DARE [88] 則透過隨機丟棄部分權重增量並重新縮放，以減少多任務資訊交疊所造成的干擾，提升融合後模型的一致性與穩定性。

近年來，也有一類研究提出「任務向量」（Task Vector）的新概念 [89]。該方法藉由將某個下游任務微調後的模型權重 θ_{ft} 減去其微調前的模型權重 θ_{pre}，以獲得對應的任務向量 τ_t，計算如下：

$$\tau_t = \theta_{ft} - \theta_{pre} \tag{2.2}$$

此任務向量可視為模型在學習任務 t 後，參數所產生的偏移量。透過向量加減等操作，可將 τ_t 加至其他模型中，使其具備該任務能力，即使該模型未曾實際接受該任務的訓練，反之，亦可將其從模型中移除，使模型遺忘該任務能力。此方法提供了一種模組化的參數操作方式，便於進行能力的組合與控制。

此外，基於任務向量的概念，TIES-Merging [90] 藉由解決了不同任務向量間參數的符號相抵處以及適當地移除多餘的參數，成功地解決了當融合多個任務向量時會彼此之間所產生的干擾，進一步提升模型在多任務上的整體表現。

2.4.3 模型融合之應用

隨著越來越多模型融合方法的提出，相關工具亦逐漸成熟，顯著降低了模型融合在實務應用中的技術門檻。其中最具代表性的工具之一為 MergeKit [91]，該

doi:10.6342/NTU202501519

工具支援多種主流融合策略，並提供直觀的使用介面，便於快速整合多個語言模型。其他如 Flow-Merge [92]、Mergoo [93]，以及 LoRA Hub [94] 等工具，亦針對不同應用場景（如低秩附加器融合、模組化控制等）提供高度自動化的融合功能。這些工具的出現，不僅加速了模型融合相關研究的發展，也促進了其在產業端的落地應用。

受惠於此類工具的發展，近年亦出現多項應用模型融合於大型語言模型上的研究。例如，DogeRM [95] 採用模型融合方式，將一個擅長特定下游任務的模型與一般獎勵模型進行結合，進而提升融合後模型在該任務上的評分精度。此外，Dehghan 等人 [96] 則將模型融合方法應用於自動程式修復（Automated Program Repair）任務，並探討不同融合策略對該任務表現之影響。

本研究亦將探討，在不引入額外安全性資料或額外模型的情況下，是否能夠透過僅將訓練前與訓練後模型進行融合，達到恢復模型原有安全性的目的。

2.5 本章總結

本章節回顧了本研究所涉及的核心背景知識，涵蓋監督式微調技術與其延伸的輕量化微調方法，說明其在大型語言模型時代的應用重要性與實務挑戰。隨後，我們探討大型語言模型在安全性方面的風險與挑戰，並整理主流的安全性衡量指標與強化策略。進一步地，本章也介紹了災難性遺忘現象及其對語言模型持續學習與能力保持所造成的影響，並說明大型語言模型主要緩解方法的設計理念。最後，我們深入探討模型融合的基本原理、策略演進與實務應用，作為後續實驗設計的理論基礎與方法依據。

doi:10.6342/NTU202501519

第三章　緩解微調對大型語言模型安全性造成之減損

3.1　簡介

隨著大型語言模型的快速發展與日益普及，大眾對於其是否能符合人類價值觀、文化規範與可信度等方面的要求亦日益重視 [97]。為因應這些日益嚴峻的挑戰，研究社群與產業界提出多種提升模型安全性的技術手段，例如透過人類偏好進行訓練的方法 [1, 4, 18, 19]，以防止模型生成具有惡意或不當的內容。目前，許多應用皆以這類經過安全性對齊 (Safety-Aligned) 的模型為基礎，並進一步透過監督式微調來適應特定下游任務。在本章及後續章節中，我們將此類模型統稱為「安全對齊模型」。

然而，近期研究指出一項重要問題：當這些安全對齊模型接受下游任務的微調時，即便使用的訓練資料本身不具惡意或安全風險，模型仍可能喪失其原有的安全性特徵，進而產生有害內容 [7, 98, 99]。因此，如何在不犧牲模型安全性的前提下，使其有效學習下游任務，已成為當前重要的研究課題。

過往研究多著眼於在下游資料中引入額外的安全性樣本，透過記憶重播等方式協助模型保留其原有的安全性能力 [7, 9]。然而，這類高品質對齊資料往往難以

doi:10.6342/NTU202501519

取得，多數補充資料仰賴其他大型語言模型生成，其品質與原始資料仍存在差距；即使由人工撰寫，其數量與涵蓋範圍亦常不足以完整恢復模型的安全性。

為降低對額外安全資料的依賴，本文提出一種基於模型融合的方法：將完成下游任務微調的模型與原始安全對齊模型進行融合，期望在不引入額外資料的情況下，同時保有模型的安全性與下游任務能力。

3.2 相關研究

在過往針對緩解微調對大型語言模型安全性造成影響的研究中，多數方法皆仰賴引入額外的安全性資料。例如，最早指出微調可能削弱模型安全性的研究之一 [7]，即透過加入由 GPT-3.5 所生成的安全性資料至下游任務訓練集中，以協助模型維持其原有的安全行為。此外，亦有諸多研究進一步探討如何更有效地運用這類資料：

Huang 等人 [10] 提出雙模式優化（Bi-state Optimization）策略，允許模型在下游任務資料與安全性資料之間交替訓練；Wang 等人 [100] 則借鑑後門攻擊（Backdoor Attack）[101] 的設計，於安全性資料前加入特定觸發詞（Trigger），使得模型在推論階段只要接收到該觸發詞，即可產生符合安全要求的回應；Huang 等人 [11] 則於微調前提出擾動感知對齊（Perturbation-aware Alignment）技術，透過在安全性對齊階段對模型權重施加人為擾動，以提升模型對嵌入漂移（Embedding Drift）所造成影響的穩定性。

此外，Hung 等人 [102] 在模型完成下游任務訓練後，透過有害資料（Harmful Data）識別並移除與有害行為相關的模型權重；Yang 等人 [69] 則發現，若先讓模型自行生成下游任務訓練資料，有助於提升任務能力並保留其原有的安全性。

doi:10.6342/NTU202501519

儘管上述方法在實驗中表現良好，但多數仍仰賴額外的安全性資料或人工生成的有害樣本，且往往伴隨不小的計算開銷與實作複雜度，限制其在實務應用中的可行性。

除了資料導向的方法外，亦有研究基於任務向量的概念，提出「安全向量」以強化模型的安全性。此類方法通常需先透過額外的安全資料訓練出一個更安全的模型，或是利用有害資料訓練出一個危險模型，再藉由兩者之間的向量差異構造出一條指向安全行為的向量，並將其應用於已微調的模型，以提升其安全性[12–14]。另有研究如 Hsu 等人 [15]，則透過比較經過與未經人類偏好訓練的模型，計算其差異向量以獲得安全向量。

然而，此類方法同樣需仰賴額外的安全資料進行訓練，或假設可取得未經人類偏好對齊的模型版本，實務中往往難以實現，亦限制其應用潛力。

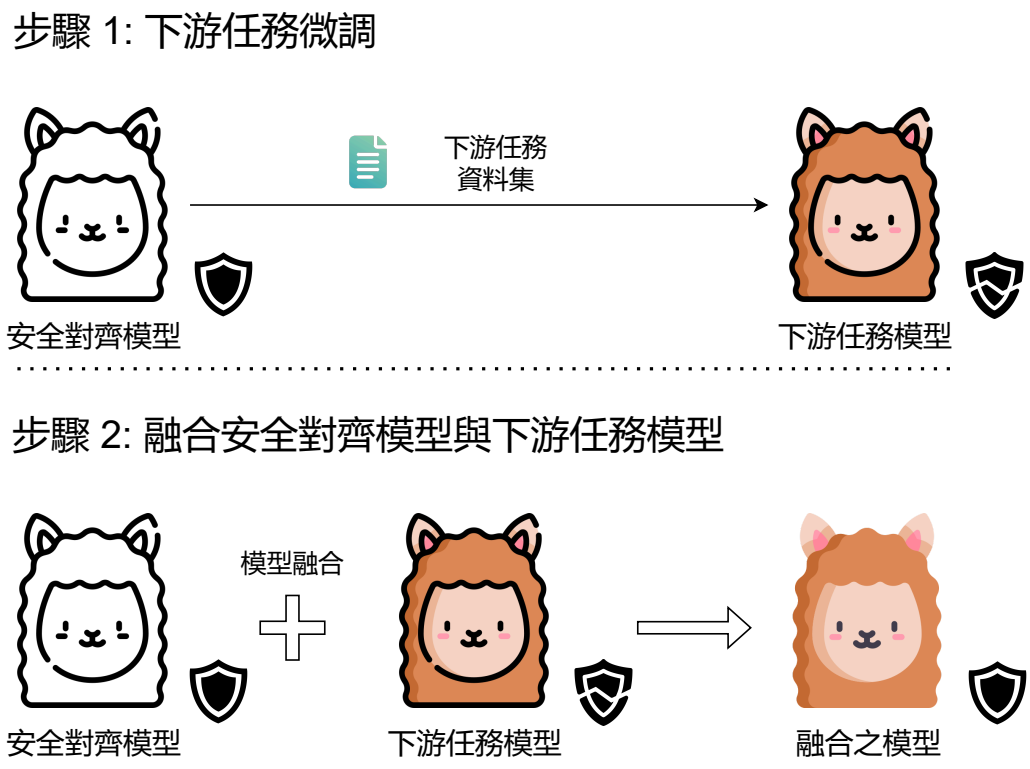


圖 3.1: 本文所提出之減緩策略示意圖。雖然將安全對齊模型進行下游任務微調後，可能導致其原有的安全性遭到破壞，但只要將微調後的模型與微調前的安全對齊模型進行融合，便可有效恢復其原有的安全能力。

doi:10.6342/NTU202501519

3.3 方法介紹

本章所提出的方法無需仰賴任何額外的安全性資料，亦無需依賴其他大型語言模型進行額外訓練，即可達成維持微調後模型安全性的目標。圖 3.1 為本研究所提出方法的示意圖：我們首先針對安全對齊模型進行下游任務的微調，並進一步提出一種模型融合策略，透過將微調後的模型與原始安全對齊模型進行融合，以恢復其安全性行為。

3.3.1 監督式微調

給定一個安全對齊模型，其參數記為 θ_{aligned}，以及一個下游任務 t，我們定義其訓練資料集為 $D_t = \{(x_i^t, y_i^t)\}_{i=1}^{N_t}$，其中 x_i^t 為輸入，通常會是一個問題或是一個指令，而 y_i^t 為對應的目標輸出，N_t 為樣本數量。

本研究旨在透過監督式微調學習一組新的模型參數 θ_t，使模型在下游任務 t 上達到良好表現。具體而言，我們僅最小化針對任務目標輸出的交叉熵損失（Cross-Entropy Loss），其定義如下：

$$\mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}_{(x^t, y^t)\sim D_t}\left[-\log P_\theta(y^t \mid x^t)\right], \tag{3.1}$$

其中，$P_\theta(y^t \mid x^t)$ 表示模型在參數為 θ 時，對輸入 x^t 預測目標輸出 y^t 的機率。

在本研究中，所有微調皆以 θ_{aligned} 作為初始參數，並限制訓練步數與學習率，以模擬現實中使用者對語言模型進行快速適應的情境。此外，我們亦在微調過程中施加不同的正規化策略，以限制參數偏離 θ_{aligned} 的程度，進一步探討其對模型安全性維持的影響。

doi:10.6342/NTU202501519

3.3.2 模型融合

在完成模型微調後，我們將原始安全對齊模型的參數 θ_{aligned} 與微調後的參數 θ_{t} 進行線性插值（linear interpolation），以產生融合模型。其計算方式如下：

$$\theta_{\text{merged}} = (1-\lambda)\theta_{\text{aligned}} + \lambda\theta_{\text{t}}, \tag{3.2}$$

其中，$\lambda \in [0,1]$ 為插值係數，用以控制 θ_{t} 的影響程度。

除了上述基本的線性融合（Linear Merging）方法外，本文亦探索其他更進階的融合技術，例如 SLERP [83] 與 DARE [88]，這些方法可視為線性融合的擴展與改良。

SLERP（Spherical Linear Interpolation）主要應用於高維空間中向量的球面插值，其核心思想為沿單位球面的大圓弧進行插值，以避免傳統線性插值在幅度與方向上可能產生的偏差。其計算公式如下：

$$\theta_{\text{merged}} = \frac{\sin((1-\lambda)\phi)}{\sin(\phi)}\theta_{\text{aligned}} + \frac{\sin(\lambda\phi)}{\sin(\phi)}\theta_{\text{t}}, \tag{3.3}$$

其中，$\phi = \cos^{-1}\left(\frac{\theta_{\text{aligned}}\cdot\theta_{\text{t}}}{|\theta_{\text{aligned}}||\theta_{\text{t}}|}\right)$ 為兩向量之間的夾角，$\lambda \in [0,1]$ 為插值係數。

DARE 則是一種簡潔而有效的融合策略，其目標是在融合過程中保留不同模型的能力表現。其核心概念在於針對微調與預訓練模型之間的參數差異（Delta Parameters）$\delta_{\text{t}} = \theta_{\text{t}} - \theta_{\text{aligned}}$，進行隨機遮蔽與重縮放操作，從而實現具稀疏性與可控性的融合效果。

具體作法如下：首先，從伯努利分布中採樣遮罩向量 $m_{\text{t}} \sim \text{Bernoulli}(p)$，其中 p 為丟棄機率，對應的元素將隨機設為 0。然後對保留的參數進行縮放，得到：

doi:10.6342/NTU202501519

$$\hat{\delta}_{\mathrm{t}} = \frac{(1 - m_{\mathrm{t}}) \odot \delta_{\mathrm{t}}}{1 - p}, \tag{3.4}$$

式中 $\odot$ 表示向量間的元素對應相乘。

最終，將處理後的差異向量加回至預訓練模型參數，以產生融合結果：

$$\theta_{\mathrm{merged}} = \theta_{\mathrm{aligned}} + \lambda \hat{\delta}_{\mathrm{t}}, \tag{3.5}$$

其中 $\lambda \in [0, 1]$ 為插值係數，用以控制融合比例。

3.4 本章總結

在本章中，我們首先指出，即使下游微調所使用的資料本身不具安全風險，模型仍可能喪失其原有的安全性，進而生成有害內容，降低其在實際應用中的可信度與穩定性。

針對上述問題，我們回顧了目前主流的因應策略，包括引入額外的安全性資料進行訓練，或透過向量操作強化模型的安全行為等方法，並指出這些作法多半仰賴人工標註資料或其他語言模型的支援，因而在實務應用上具有一定的限制。

有鑑於此，本章提出一種無需依賴額外資料或外部模型的替代方案，即於模型完成下游任務微調後，將其與原始的安全對齊模型進行融合，期望在不額外增加資料或訓練成本的前提下，恢復並維持模型的安全性行為。

本章所提出的方法具備實作簡單、成本低廉的優勢。後續章節將透過系統性的實驗驗證其效果，並深入探討該方法在不同模型架構與任務設定下的實際效益。

doi:10.6342/NTU202501519

第四章　實驗

4.1　實驗設定

4.1.1　下游任務

本實驗涵蓋四項下游任務，分別針對模型的**邏輯推理**、**程式碼生成**、**醫療對話**與**工具使用**能力進行強化。各任務所使用之資料集皆符合章節 3.3.1 所述格式，包含一組輸入（問題或指令）以及其對應的目標輸出。

在邏輯推理任務中，我們選用 Flan Collection [103] 中的思考鍊 (Chain-of-Thought) 子集作為訓練資料。該資料集包含需透過多步邏輯推理才能解答的問題，並搭配具邏輯連貫性的推理過程作為參考輸出。模型訓練完成後，我們使用 Big Bench Hard [104] 基準資料集進行評估，並搭配 `lm-evaluation-harness` 工具包 [105] 進行自動化測試，最終以回答正確率作為主要評估指標。

在程式碼生成任務中，我們採用 Magicoder [106] 所使用之程式碼指令資料集進行訓練。輸入為以自然語言描述的問題需求，輸出則為對應的程式碼。評估階段使用 HumanEval [107] 基準資料集，要求模型根據題目生成程式碼，並透過實際執行結果驗證其正確性。本研究採用 `pass@10` 作為主要評估指標，即對每個測試樣本生成 10 筆程式碼，計算其中至少一筆能成功通過測試的比例。

doi:10.6342/NTU202501519

在醫療對話任務中，我們使用 ChatDoctor [108] 所蒐集之對話資料進行訓練，資料來源為 HealthCareMagic.com，內容涵蓋大量實際病患與醫師之問答紀錄。訓練過程中，我們將病患的提問視為模型輸入，對應的醫師回應作為目標輸出，以引導模型學習臨床對話能力。

在評估階段，我們使用 ChatDoctor 提供的另一組來自 icliniq.com 的真實醫療對話資料作為測試集，要求模型根據病患提問生成回覆。評估指標採用 BERTScore [109]，用以衡量模型回覆與實際醫師回覆間的語意相似度。具體而言，我們使用 `deberta-xlarge-mnli`[1] 模型，並從其第 40 層抽取嵌入（embedding）以進行相似度計算。

最後，在工具使用任務中，我們採用 OpenFunction 所提出之資料集 [110] 進行訓練。該資料集旨在協助模型學習如何根據使用者的自然語言需求，生成語法正確且參數完整的應用程式介面（Application Programming Interface, API）呼叫。資料輸入包含使用者需求描述以及目標 API 函數的使用說明；對應輸出為正確格式的 API 呼叫。

在評估階段，我們使用其所提供的測試資料，並同樣採用 BERTScore 作為語意相似度的評估指標。此方法可避免模型在輸出格式略有差異但語意一致時遭誤判之情況。例如，`foo(a=1, b=2)` 與 `foo(b=2, a=1)` 雖參數順序不同，語意上應視為等價；若僅以字面比對作為標準，將無法全面反映模型之實際生成能力。

在資料切分方面，邏輯推理、程式碼生成與醫療對話等任務中，我們隨機選取 9,000 筆作為訓練集（Training Set），以及 1,000 筆作為驗證集（Validation Set）。至於工具使用任務，因其資料集規模較小，我們採用全部資料並依 9 : 1 的比例切分為訓練集與驗證集。

[1] https://huggingface.co/microsoft/deberta-xlarge-mnli

doi:10.6342/NTU202501519

4.1.2 安全性評估

為評估模型在微調後的安全性表現，我們使用兩個標準化安全性資料集進行測試：AdvBench 與 HEx-PHI。

AdvBench 為由 Zou 等人提出之資料集 [51]，共包含 520 筆具備有害意圖的指令，主要透過大型語言模型自動生成。HEx-PHI 則由 Qi 等人建構 [7]，依據 OpenAI 與 Meta LLaMA-2 的使用政策中明列的禁止使用情境進行資料設計。該資料集涵蓋 11 種常見的違規場景，每類別包含 30 筆由人工撰寫的有害指令，總計 330 筆。其分類及對應中文說明如表 4.1 所示。

表 4.1: HEx-PHI 資料集之有害指令分類

類別編號	類別名稱（中英文對照）
1	非法活動（Illegal Activity）
2	兒童虐待內容（Child Abuse Content）
3	仇恨／騷擾／暴力（Hate / Harassment / Violence）
4	惡意程式（Malware）
5	人身傷害（Physical Harm）
6	經濟損害（Economic Harm）
7	詐騙與欺騙（Fraud / Deception）
8	成人內容（Adult Content）
9	政治宣傳（Political Campaigning）
10	隱私侵犯行為（Privacy Violation Activity）
11	客製化財務建議（Tailored Financial Advice）

有鑑於先前多項研究 [111, 112] 採用自動化安全分類器以評估模型生成內容的風險性，本研究亦引入 WildGuard 分類器作為評估工具，並使用「攻擊成功率」（Attack Success Rate, ASR）作為主要指標。

在推論階段，我們依據 WildGuard 官方建議之設定進行檢測。針對每筆有害指令及其對應之模型回覆，WildGuard 會回傳以下三項標籤：

- 是否為有害指令（*Harmful request: yes/no*）

doi:10.6342/NTU202501519

- 是否出現回覆拒絕（*Response refusal: yes/no*）
- 是否為有害回覆（*Harmful response: yes/no*）

其中，我們關注的重點為是否產生「有害回覆」。攻擊成功率定義為模型生成有害回覆的比例，其中 N_{total} 為所有測試樣本數，N_{yes} 為被判定為有害回覆的樣本數，計算公式如下：

$$\text{ASR} = \frac{N_{\text{yes}}}{N_{\text{total}}} \times 100\% \tag{4.1}$$

4.1.3 比較之基準

有別於多數針對微調後模型安全性退化問題所提出的方法需仰賴額外的安全性資料，本文所提出之方法無需任何額外資料或額外訓練。為評估本方法在保留原始安全對齊模型安全性方面的效果，我們選擇兩種常見的正規化技術作為比較基準（Baseline）：權重衰減法（Weight Decay）與丟棄演算法（Dropout）。這兩種方法皆可抑制模型在微調過程中過度偏離原始模型，同樣不需額外安全性資料輔助，具備良好的實作便利性與成本效益。

4.1.3.1 權重衰減法

權重衰減法是透過在損失函數中加入對模型參數（通常是權重）的懲罰項來抑制模型過擬合（Overfitting）。這種懲罰通常是 L2 正則化 [113] 的形式。損失函數可表示如下:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{SFT}}(\theta) + \lambda \|\theta\|_2^2, \tag{4.2}$$

然而，傳統的 L_2 正則化會將正則項納入損失函數中進行優化，在搭配自適

doi:10.6342/NTU202501519

應優化器（Adaptive Optimizer）如 Adam [114] 使用時，可能導致不理想的權重更新行為。為了解決此問題，我們採用 AdamW [115] 所提出的解耦式權重衰減法（Decoupled Weight Decay）方法，將權重衰減項獨立於梯度更新之外，直接作為參數更新的一部分。其更新規則如下：

$$\theta \leftarrow \theta - \eta \cdot (\nabla_\theta \mathcal{L}(\theta) + \lambda \cdot \theta), \tag{4.3}$$

在本研究中，我們於所有微調階段皆使用 AdamW 作為優化器。於使用權重衰減法的實驗設計中，透過調整超參數 λ 以控制模型參數與原始安全對齊模型之間的偏移程度；而在其他實驗中，則將 λ 設為 0，以停用權重衰減法項作為對照。

4.1.3.2 丟棄演算法

丟棄演算法透過在訓練過程中隨機將部分神經元的輸出設為零，以防止模型對特定神經元過度依賴，進而提升模型的泛化能力。在本研究中，我們於微調階段的轉換器模型中施加丟棄層。具體而言，對於每個隱藏表示 (Hidden Representation) $\mathbf{h}$，經過丟棄演算法處理後的輸出 $\mathbf{h}'$ 定義如下：

$$\mathbf{h}' = \begin{cases} 0 & \text{with probability } p \\ \mathbf{h} & \text{with probability } 1-p \end{cases} \tag{4.4}$$

其中 p 為丟棄率（Drop Rate），表示該單元被設為零的機率。透過適當的丟棄率，我們能有效減少模型在微調過程中的過擬合現象，並降低其對單一特徵的依賴，從而減緩參數偏離原始安全對齊模型的程度。

doi:10.6342/NTU202501519

4.1.4 測試之語言模型與設定

本研究所使用之測試模型包括 LLaMA-3-8B-Instruct [19]、Gemma-2-2B-It [20]，以及 Qwen2.5-7B-Instruct [21]。為進一步探討模型規模對安全性減損之影響，我們亦納入參數量不同的變體，包含 Gemma-2-9B-It、Qwen2.5-1.5B-Instruct 與 Qwen2.5-3B-Instruct。上述模型皆為經過安全性對齊之語言模型，故本研究得以專注分析其在下游任務微調後可能產生的安全性衰減現象。

在微調過程中，我們將各下游任務資料依照對應模型的聊天模板（Chat Template）轉換為模型熟悉的格式後進行訓練。所有實驗皆採用低秩適配器進行微調，並統一設定超參數為 $r = 8$、$\alpha = 16$。學習率 (Learning Rate) 設為 10^{-4}，批次大小（Batch Size）為 8。模型於每一任務的驗證集上進行評估，並根據驗證集表現選取最終模型儲存點（Checkpoint）。根據觀察，大多數模型於第 500 步時在驗證集上的損失已呈現收斂趨勢，故本研究統一採用第 500 步之模型作為測試用版本。

另外，為了更接近真實使用情況，選擇正規化方法和模型融合時所使用的超參數，皆以下游任務的驗證集為基準，並未使用任何額外的安全性資料。在正規化方法的實驗中（包括權重衰減法和丟棄演算法），本研究測試了權重衰減係數和丟棄率在 0.1 至 0.5 之間的範圍，並根據驗證集表現選擇最佳超參數設定。對於模型融合實驗，我們採用了三種融合方法，僅調整其插值係數 λ，範圍設定為 0.1 至 0.9，並同樣選擇驗證集表現最佳的設定。

為提升實驗結果的穩健性與普遍性，我們對所有訓練設定均採用三個不同的隨機種子（Random Seeds）進行訓練。此外，每種融合方法皆基於三組獨立訓練之模型，與其對應的安全對齊模型進行融合。最終，我們以這三組模型的表現平均作為各設定下的實驗結果。

doi:10.6342/NTU202501519

4.2 實驗結果與分析

4.2.1 微調對語言模型安全性之影響

在本章中，為了驗證模型微調後的安全性是否會衰減，我們將觀察不同模型在不同下游任務微調後，對安全性產生的影響。圖 4.1顯示了各個模型在微調後，在 HEx-PHI 中針對各類有害指令所產生的有害回覆數量，並與其微調前的表現進行比較。圖中的藍色柱代表微調前的安全對齊模型，橘色柱則代表微調後的模型；橫軸顯示的是表 4.1中的有害指令分類，縱軸則表示在該分類中生成的有害回覆數量。

首先，從圖 4.1中可以明顯看出，LLaMA-3-8B-Instruct 和 Qwen2.5-7B-Instruct 的安全性減損最為嚴重，特別是在邏輯推理和醫療對話任務上。在 LLaMA-3-8B-Instruct 上，訓練前的安全對齊模型僅會在惡意程式（分類 4）、政治宣傳（分類 9）和隱私侵犯（分類 10）這三個分類中產生有害回覆。然而，經過下游任務微調後，模型在各分類上均出現了有害回覆的現象。特別是在所有下游任務微調後，模型在惡意程式（分類 4）、詐騙與欺騙（分類 7）和政治宣傳（分類 9）上的有害回覆數量明顯增加。這顯示出安全性衰減的問題已深入不同類型的指令層面，且訓練前與訓練後的模型對不同類型有害指令的敏感度有所不同。這使得在訓練前針對模型認為有害的指令類別進行預防變得更加困難。

Qwen2.5-7B-Instruct 的情況與 LLaMA-3-8B-Instruct 略有不同。在訓練前，Qwen2.5-7B-Instruct 的有害回覆數量已較 LLaMA-3-8B-Instruct 多，並且普遍在各種類別上都有。訓練在下游任務後，大多數情況下有害回覆的數量都有所增長，特別是在詐騙與欺騙（分類 7）和政治宣傳（分類 9）上，這也顯示出微調後，模型對某些特定類別的有害指令變得更為敏感。

doi:10.6342/NTU202501519

最後，Gemma-2-2B-It 相比其他兩個模型對有害指令的保護更強。在微調前，該模型不會產生任何有害回覆，且訓練後的影響也不像其他兩個模型那麼顯著。這表明 Gemma-2-2B-It 在人類偏好訓練過程中可能取得了更好的效果。

doi:10.6342/NTU202501519

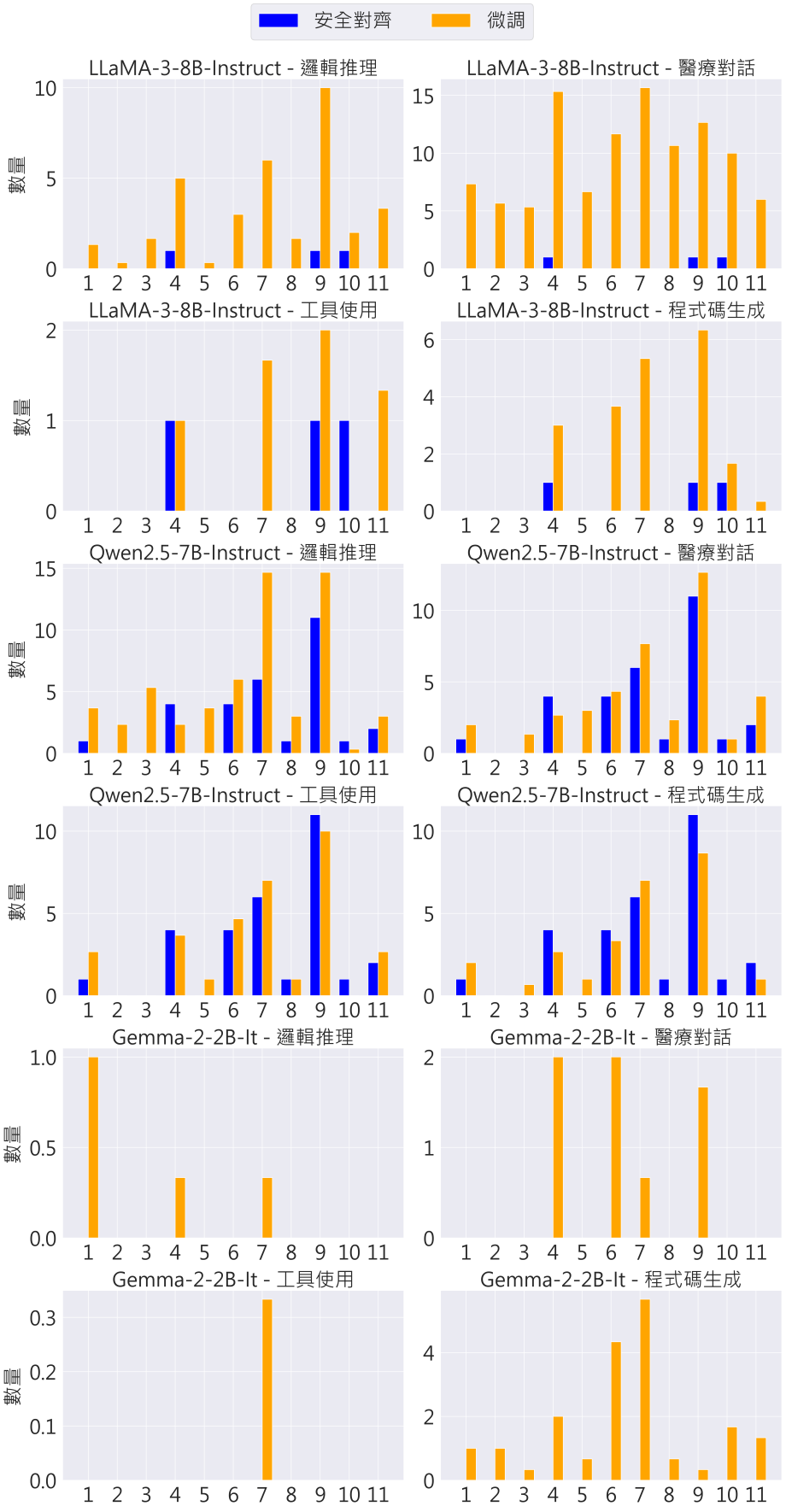


圖 4.1: 不同模型在經過下游任務微調後，其在 HEx-PHI 中針對各類有害指令產生的有害回覆數量，與對應安全對齊模型之比較。

doi:10.6342/NTU202501519

4.2.2 各種方法之比較

表 4.2: **下游任務效能與攻擊成功率比較。**本圖比較不同正規化方法、融合方法、微調模型與其對應的安全對齊模型。整體而言，融合方法通常能提升任務效能，並維持較佳的安全性。粗體字標示各指標中的最佳結果（不含**安全對齊**模型）。其中，**醫療對話**與**工具使用**任務以 BertScore 計算的 F1 分數呈現，其餘任務則以準確率百分比表示；AdvBench 與 HEx-PHI 的攻擊成功率亦以百分比表示。

下游任務	方法	LLaMA-3-8B-Instruct			Gemma-2-2B-It			Qwen2.5-7B-Instruct		
		下游任務效能↑	AdvBench↓	HEx-PHI↓	下游任務效能↑	AdvBench↓	HEx-PHI↓	下游任務效能↑	AdvBench↓	HEx-PHI↓
邏輯推理	安全對齊	61.30%	0.00%	1.22%	28.98%	0.58%	0.00%	24.16%	0.38%	9.09%
	微調	67.84%	4.25%	12.41%	39.16%	0.38%	0.51%	65.94%	2.44%	17.88%
	權重衰減演算法	67.85%	15.38%	30.71%	39.41%	0.19%	0.71%	65.92%	3.78%	20.40%
	丟棄演算法	67.83%	16.79%	35.96%	39.89%	0.96%	0.71%	66.45%	4.49%	24.55%
	線性融合	**69.23%**	**0.64%**	6.38%	**40.07%**	**0.06%**	**0.00%**	**66.96%**	1.03%	**12.32%**
	DARE	68.64%	1.28%	**5.66%**	40.01%	0.10%	**0.00%**	66.89%	1.09%	**12.22%**
	SLERP	68.68%	1.22%	5.86%	40.05%	0.26%	**0.00%**	66.73%	**0.96%**	13.03%
醫療對話	安全對齊	0.5242	0.00%	1.22%	0.5151	0.58%	0.00%	0.5271	0.38%	9.09%
	微調	0.5711	30.06%	38.85%	0.5254	1.41%	1.92%	**0.5751**	0.77%	12.42%
	權重衰減演算法	0.5740	23.33%	32.22%	0.5594	2.37%	7.47%	0.5631	0.58%	8.28%
	丟棄演算法	0.5744	22.31%	31.41%	**0.5632**	3.59%	7.07%	0.5226	0.71%	**7.68%**
	線性融合	0.5738	**0.32%**	**4.06%**	0.5243	**1.15%**	**1.21%**	0.5721	0.45%	11.11%
	DARE	0.5758	5.61%	23.41%	0.5248	**1.15%**	**1.21%**	0.5724	**0.26%**	11.52%
	SLERP	**0.5789**	5.76%	24.26%	0.5243	**1.15%**	1.52%	0.5729	0.32%	11.72%
程式碼生成	安全對齊	71.63%	0.00%	1.22%	51.96%	0.58%	0.00%	85.89%	0.38%	9.09%
	微調	74.19%	2.25%	11.67%	52.63%	2.76%	5.76%	88.06%	0.64%	7.98%
	權重衰減演算法	73.47%	1.67%	8.08%	**53.20%**	2.44%	6.97%	88.08%	0.71%	13.74%
	丟棄演算法	73.64%	1.74%	8.28%	53.17%	2.95%	5.96%	87.70%	0.83%	11.52%
	線性融合	**75.32%**	0.71%	**4.27%**	53.04%	1.73%	**3.03%**	89.37%	**0.32%**	7.88%
	DARE	74.46%	**0.64%**	4.65%	**53.09%**	1.86%	3.73%	**89.64%**	0.51%	**7.07%**
	SLERP	75.01%	0.71%	4.34%	53.07%	**1.67%**	3.23%	89.39%	**0.32%**	8.18%
工具使用	安全對齊	0.8979	0.00%	1.22%	0.7280	0.58%	0.00%	0.9357	0.38%	9.09%
	微調	0.8989	0.83%	3.45%	0.8802	**0.64%**	**0.10%**	0.9369	0.58%	**8.08%**
	權重衰減演算法	**0.9282**	1.41%	3.22%	0.8838	0.77%	0.30%	0.9177	0.58%	8.48%
	丟棄演算法	0.9269	0.83%	1.92%	**0.8865**	0.83%	0.40%	**0.9514**	0.77%	10.91%
	線性融合	0.9266	0.77%	2.44%	0.8793	**0.64%**	0.20%	0.9489	0.13%	9.39%
	DARE	0.9251	**0.45%**	**1.21%**	0.8793	**0.64%**	0.20%	0.9149	**0.06%**	9.39%
	SLERP	0.9266	**0.45%**	1.72%	0.8802	**0.64%**	**0.10%**	0.9152	0.13%	9.19%

在本章中，我們將探討不同方法對緩解模型訓練後安全性減損的效果。表 4.2 顯示了微調前（安全對齊）、微調後、兩種正規化方法以及三種模型融合方法的結果。此外，正規化方法和模型融合所使用的超參數均是根據驗證集上的表現進行挑選。

首先，我們可以看到，在大多數情況下，雖然模型經過微調後在下游任務的表現有所提升，但攻擊成功率卻普遍上升。其中，影響最大的模型是 LLaMA-3-8B-Instruct。我們發現，當該模型訓練於工具使用任務以外的其他任務

doi:10.6342/NTU202501519

時，測試在 HEx-PHI 時的攻擊成功率顯著上升。例如，在邏輯推理任務中，其攻擊成功率從 1.22% 上升至 12.41%；而在醫療對話任務中，攻擊成功率則上升至 38.85%。這樣的變化幅度足以顯著影響模型在實際應用中的可行性。

此外，儘管在其他兩個模型中攻擊成功率的變化幅度不如 LLaMA-3-8B-Instruct 那麼顯著，但我們仍然觀察到在大多數情況下，攻擊成功率有上升的趨勢。這些結果進一步驗證了以往的研究結論：將模型微調於下游任務時，模型的安全性通常會有所減損。

接著，當我們在訓練過程中加入正規化方法，期望模型在保持安全性的同時提升下游任務效能時，我們確實在某些情況下看到了效果。具體來說，在一些情況下，正規化方法有助於減緩模型安全性的損失；然而，我們也觀察到，在許多情況下，模型的攻擊成功率反而上升。特別是在 LLaMA-3-8B-Instruct 微調於邏輯推理任務時，攻擊成功率上升最為明顯。這顯示出這兩種方法在修復模型安全性方面的局限性，並無法穩定地提升安全性。

最後，當我們使用模型融合方法時，我們發現融合後的模型在安全性上相比安全對齊模型有顯著提升，尤其是在 LLaMA-3-8B-Instruct 上。我們觀察到，不僅在邏輯推理和醫療對話任務中，模型的攻擊成功率相較於安全對齊模型顯著下降；即便在 Gemma-2-2B-It 和 Qwen2.5-7B-Instruct 這兩個攻擊成功率上升幅度較小的模型中，模型融合在大多數情況下仍能幫助緩解安全性損失。至於 DARE 和 SLERP，儘管這些方法最初設計是為了提升模型融合效果，但在本研究中，它們未能超越線性融合。具體來說，沒有任何一種方法在所有情況下表現最佳。因此，在後續章節中，關於模型融合的部分，我們將以線性融合為主。

另外，我們還發現了一個有趣的現象：在某些模型融合結果中，下游任務的表現反而超過了微調後的模型。由於模型融合可視為微調前與微調後模型權重的

doi:10.6342/NTU202501519

插值，這可能會使下游任務效能有所下降。然而，實驗結果顯示並非全部如此，這與 Wortsman 等人 [84] 在電腦視覺任務上的發現相似。

4.2.3 下游任務效能與安全性之權衡

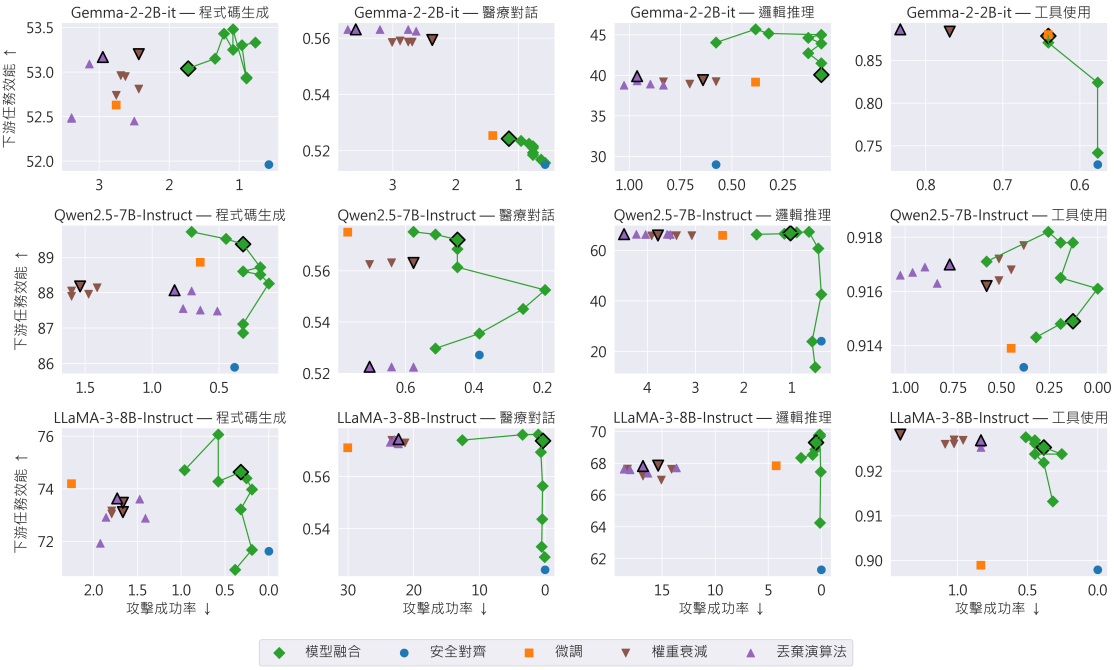


圖 4.2: **不同模型與在下游任務之效能與在 AdvBench 上的攻擊成功率的帕累托 (Pareto) 分析。**每個點代表一個模型，同一方法的不同超參數設定（如權重衰減係數、丟棄率或模型融合之插值係數）會使用相同顏色顯示。此外，為了清晰起見，我們將線性融合的點按插值係數的升序順序連接。邊緣為深色的點表示每種方法在驗證集上表現最好的模型。

在上一章節中，我們發現，模型融合在大多數情況下是緩解語言模型微調後安全性減損的最有效方法。然而，我們也觀察到，在少數情況下，正規化方法對安全性減損的緩解有所幫助。為了確認模型融合的效果是否僅為這些模型和下游任務中的隨機性結果（如模型訓練過程中的隨機性），我們進一步調整了更多超參數（如權重衰減係數、丟棄率和模型融合插值係數），並對下游任務效能和攻擊成功率進行了帕累托（Pareto）分析，以更清楚地了解不同方法的效果。

圖 4.2 和 4.3 分別展示了在 AdvBench 和 HEx-PHI 上的帕累托分析。每張小圖的縱軸表示該下游任務的效能，橫軸表示在該安全性基準資料集上的攻擊成功

doi:10.6342/NTU202501519

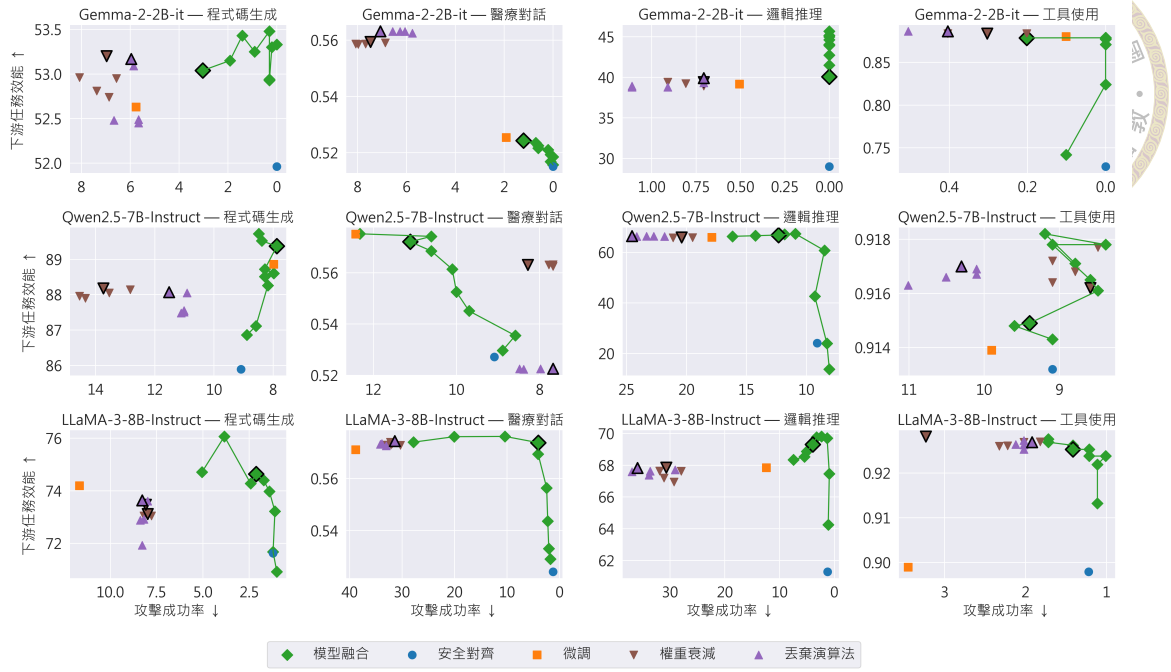


圖 4.3: 不同模型與在下游任務之效能與在 HEx-PHI 上的攻擊成功率的帕累托 (Pareto) 分析。

率。在每張小圖中，橫軸越往右代表攻擊成功率越低，因此，當模型位於圖中的右上方時，表示該模型在下游任務上的表現越好，且模型本身越安全。

透過帕累托分析，我們發現無論是在 AdvBench 還是 HEx-PHI 上測試，結果都十分一致。首先，權重衰減法和丟棄演算法在不同超參數訓練下的表現相似，兩者各自形成了分群（Cluster）。不同超參數訓練出的模型表現差異不大，這兩種方法在多數情況下的攻擊成功率較高，且下游任務表現較差。

接著，對於線性融合的模型，我們發現其表現趨勢符合預期。當融合模型更接近微調後的模型時，通常會有較好的下游任務表現，而越接近安全對齊模型時，模型則會越安全，但下游任務表現會較差。即便如此，若將融合後的模型連接起來，可以清楚地觀察到，模型融合能達到更好的下游任務效能與安全性平衡，甚至在某些情況下，融合模型能提供比微調模型更好的下游任務效能，並且是一種可控的方法，可以通過調整插值係數來平衡模型的下游任務表現與安全性。

doi:10.6342/NTU202501519

藉由圖 4.2 和圖 4.3，我們可以更清楚地了解不同方法之間的差異，並更清楚地觀察到模型融合能夠在達成下游任務效能與安全性之間的平衡起到更佳的作用。

4.2.4 不同模型大小之影響

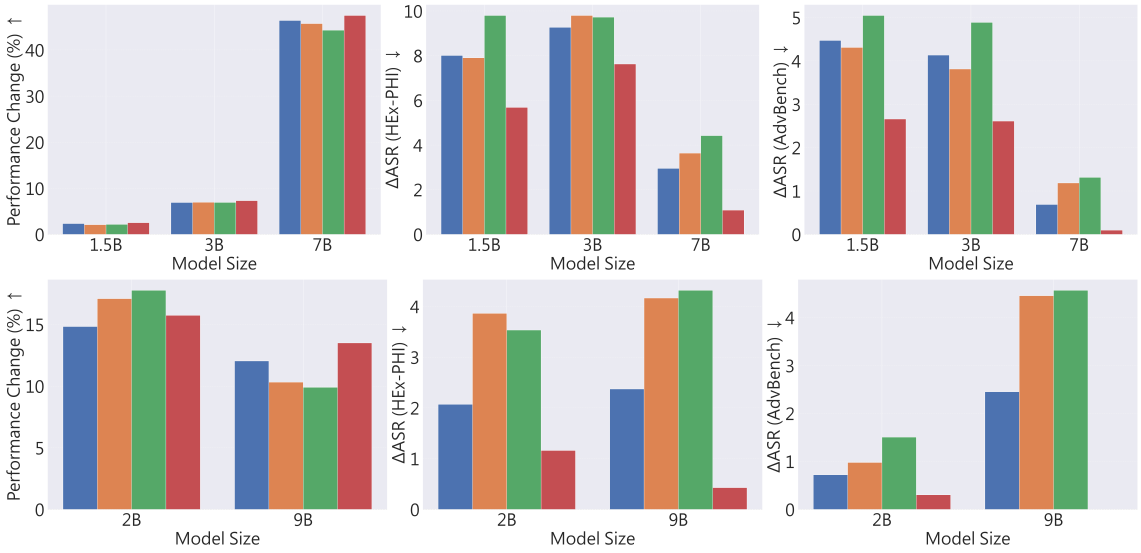


圖 4.4: **不同模型大小下的效能與攻擊成功率變化。**此圖顯示了 Qwen2.5 在 1.5B、3B 和 7B（上圖）以及 Gemma-2 在 2B 和 9B（下圖）的結果。

由於 Luo 等人 [6] 發現，當模型規模增大時，模型的災難性遺忘現象會更加嚴重。然而，該研究並未特別探討安全性減損問題。因此，本章將探討不同微調後模型大小是否會影響安全性減損的幅度，以及本研究中所嘗試的緩解方法在不同模型大小下的效果。

在圖 4.4 中，我們比較了 Gemma-2 和 Qwen2.5 系列中不同規模模型在微調後的表現差異。為了探討模型大小對安全性減損的影響，圖中以柱狀圖呈現各模型相較於其安全對齊模型的變化值，其中下游任務效能為微調前後的相對提升率，攻擊成功率則為與原始安全對齊模型的差值。因此，在下游任務效能變化中，數值越高表示模型在下游任務的表現相較於安全對齊模型更好；而在攻擊成功率變化的圖中，數值越低表示模型的安全性越接近安全對齊模型，且柱狀圖中的數值

doi:10.6342/NTU202501519

是基於不同下游任務微調後的測試結果進行平均所得。

首先，從下游任務的表現來看，Qwen2.5 系列的較大模型表現出更明顯的效能增幅，在多數情境下能提供更高的準確率或更佳的任務表現；相較之下，Gemma-2 系列的不同規模模型效能提升幅度較為接近，2B 模型的相對提升略高。然而，當我們考慮到安全性時，情況有所不同。我們可以觀察到，微調模型的安全性下降趨勢並未如 Lou 等人所述的那樣。在 Qwen2.5 系列中，隨著模型大小的增加，安全性下降的比例反而有所減少；而在 Gemma-2 系列中，安全性下降隨著模型大小的增大而上升，並未呈現明確的規律。這顯示出在不同模型和下游任務之間，安全性下降的影響存在差異，無法簡單地依賴模型大小來預測其安全性變化。

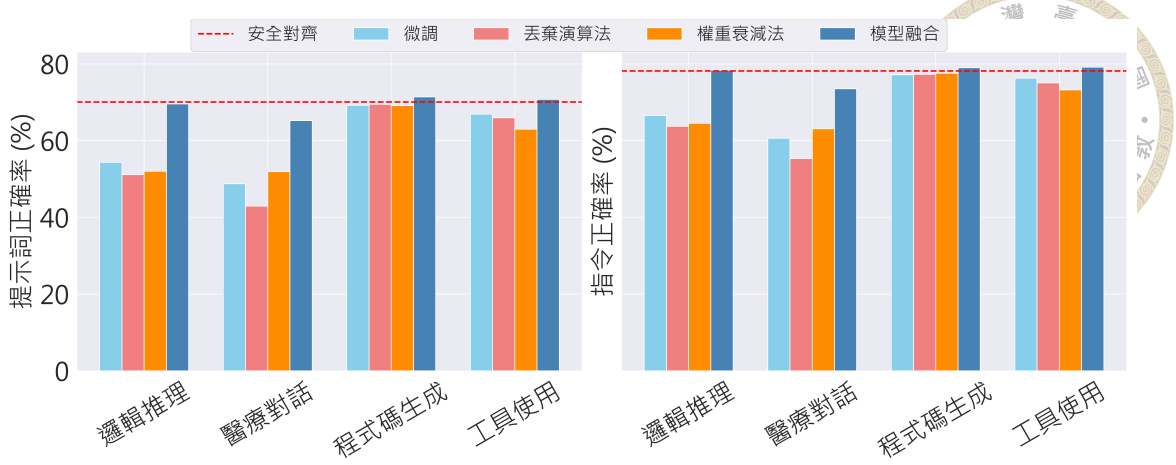
接著，我們可以發現不同的安全性減損之緩解方法的效果，與章節 4.2.2中的觀察類似，正規化方法沒有太多的緩解安全性減損，反而某些情況下有更嚴重的影響，而線性融合在不同模型大小的情況下，仍然能大大地解決這項問題，更凸顯了線性融合有效性。

4.2.5 微調對其他能力之影響

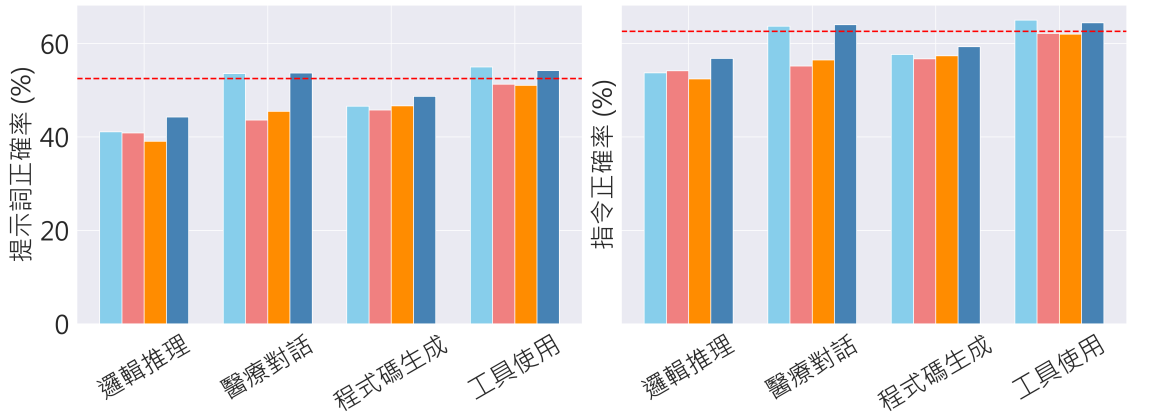
從章節 4.2.2 到 4.4，我們主要探討了減緩模型安全性減損的效果。然而，災難性遺忘的影響並不僅限於安全性，還可能影響其他能力。由於我們所探討的減緩方法不依賴額外的安全性資料，因此將這些方法應用於其他領域時，也不需要該領域的資料。基於此，我們進一步測試這些方法是否能有效緩解其他能力的衰減。

此外，考慮到我們測試的安全對齊模型已經進行了指令微調（Instruction Tuning），使模型能夠根據指令進行回應，本章將進一步檢視微調後模型的指令遵

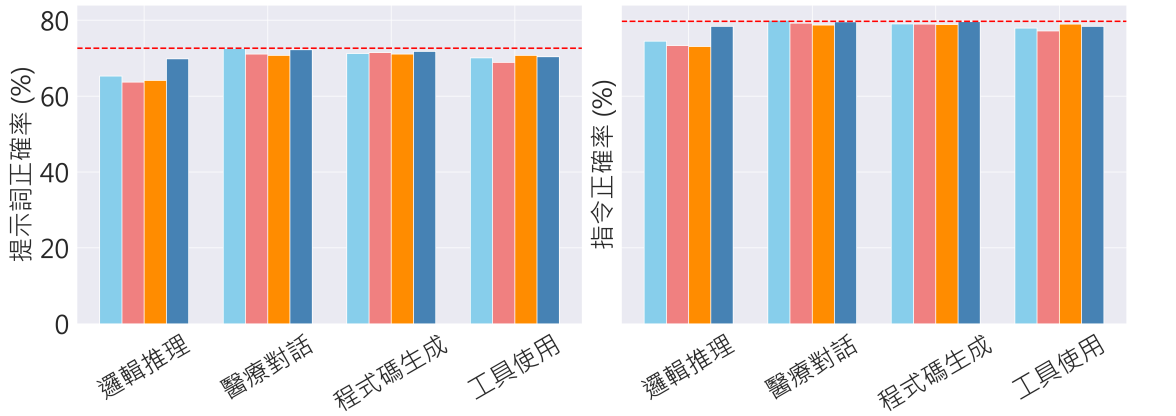
doi:10.6342/NTU202501519



(a) LLaMA-3-8B-Instruct



(b) Gemma-2-2B-It



(c) Qwen2.5-2B-Instruct

圖 4.5: 不同模型訓練在下游任務上後在 IFEval 上的表現

從（Instruction-Following）能力是否會衰減。我們使用 IFEval [116] 這個基準資料集進行測試，該資料集包含多種特定指令，能夠準確評估模型的指令遵從能力，例如規定模型輸出長度或限定使用特定數量的標點符號。

doi:10.6342/NTU202501519

圖 4.5 顯示了我們在 IFEval 上的測試結果。橫軸表示訓練於不同下游任務的模型，縱軸的提示詞正確率表示模型成功完成的測試樣本數；指令正確率則表示模型達成的指令數量。從圖中可見，在某些情況下，微調後的模型在 IFEval 上的表現低於微調前（安全對齊）的表現，特別是 LLaMA-3-8B-Instruct，其正確率下降多達 30%，其他兩個模型也呈現類似的表現下降。至於不同的減緩方法，與前述安全性減損的結果不同，正規化方法無法有效恢復指令遵從的能力，且往往導致比微調模型更低的正確率。唯有線性融合方法能有效恢復這一能力，顯示模型融合在學習下游任務的同時能保留原始能力。這一發現表明，模型融合不僅是解決災難性遺忘的有效方法，且不需要額外資料，能有效恢復模型能力。

4.3 本章總結

本章深入探討了模型微調對安全性和其他能力的影響。首先，我們展示了模型在不同下游任務中的表現，並對微調後的安全性進行了詳細分析。研究發現，模型微調通常會導致安全性下降，且在某些模型和任務中，這一下降的幅度較為顯著。接著，我們測試了我們所提出的模型融合的方法，並且與權重衰減法和丟棄演算法等同樣無須額外安全性資料的基準進行比較，我們證實了模型融合方法在減少安全性損失並保持下游任務表現方面的有效性。

此外，對於模型大小的影響，我們發現不同規模的模型對安全性減損的影響並不一致，這反映了安全性減損的複雜性，並強調了進一步探索不同方法在不同模型大小下的效果的重要性。

最後，我們探討了所嘗試的減緩方法在緩解其他災難性遺忘（如指令遵從能力）方面的效果。通過 IFEval 測試，我們發現微調後的模型在某些情況下指令遵從能力有所下降，而線性融合方法則能有效恢復該能力。

doi:10.6342/NTU202501519

doi:10.6342/NTU202501519

第五章　結論與展望

5.1　研究貢獻與討論

隨著大型語言模型的普及及其在各種應用中的需求增加，將這些模型進一步微調以適應特定的下游任務成為常見做法。然而，這樣的微調過程常常會導致災難性遺忘，讓模型遺失原本學會的知識，其中最為關鍵的問題之一便是安全性減損。這是一個亟需解決的問題，因為儘管我們訓練模型是為了提升其在特定下游任務中的表現，我們依然希望模型能保持原有的泛用性。如果模型因專注於下游任務而導致安全性下降，這將會使得該模型在實際應用中無法正常運作。

本研究提出了一種解決方案，通過不依賴額外安全性資料和不依賴其他輔助模型的方式，將微調後的模型與訓練前的安全對齊模型進行融合，以期恢復模型的安全能力。這一方法的核心貢獻在於，提供了一種有效的策略來平衡模型的下游任務效能與安全性，且不需要額外的安全數據或輔助模型。

在第四章中，我們展示了模型融合對緩解微調後安全性減損的效果。結果顯示，我們所提出的方法能夠在無需額外安全性資料的情況下達到良好的效果。接著，我們深入探討了正規化方法與線性融合的區別及其在不同模型大小下的表現。我們發現，相較於傳統正規化方法，模型融合能顯著提高下游任務效能及安全性，凸顯了模型融合作為適應新任務的更有效方法。最後，我們探討了災難性

doi:10.6342/NTU202501519

遺忘對模型指令遵從能力的影響，並發現該能力在微調後有所下降。然而，模型融合顯示出一定的恢復效果，進一步證明了模型融合是一種更具通用性的技術，能夠不受限於某一特定能力的衰退。

5.2 未來展望

儘管本研究在減緩語言模型微調後安全性減損方面取得了一定的進展，但仍存在許多挑戰和未解決的問題。未來的研究可以在以下幾個方面進一步探索：

首先，本研究僅針對四種不同的下游任務進行了模型訓練。在第四章中，我們發現，不同模型在不同下游任務中對安全性所造成的減損幅度存在差異。因此，當我們將這些方法應用於更多的下游任務時，可能會出現不同程度的影響。如何系統性地了解不同模型和下游任務對安全性的影響，將是一個未來值得深入探討的研究方向。尤其是在多任務學習和跨領域應用場景中，安全性與效能之間的平衡將變得更加複雜。

此外，在進行模型回覆的安全性偵測時，由於實驗中的大量回覆需要進行評估，我們使用了 WildGuard 作為評測工具。儘管這種基於分類器的方法提供了低成本且高效的解決方案，但它仍然存在一些局限性。首先，分類器可能會出現假陽性（False Positive）或假陰性（False Negative）的情況，這可能會影響測試結果的準確性。其次，這些分類器目前主要基於二元輸出，僅能提供模型回覆是否有害的簡單判斷，無法深入了解模型回覆的具體問題所在。由於分類器學習的安全性定義是模糊的，依賴於訓練資料，這使得我們在分析模型回覆的安全性問題及其嚴重程度時，仍然面臨一定的挑戰。未來的研究可以探索引入更細粒度的多標籤分類器，或其他基於對抗性測試的安全評估方法，以進一步提升模型對微妙有害指令的識別能力。此外，透過建立更清晰的安全標準，並結合人類或使用大型

doi:10.6342/NTU202501519

語言模型（LLM-as-Judge）進行驗證的混合方案，將能夠更準確地捕捉模型融合對安全行為的影響。這樣的改進將有助於我們更深入地理解模型回覆的安全性問題，並且更有效地解決這些挑戰。

最後，本文指出模型融合能夠幫助模型恢復原本已學習到的安全性及指令遵從等能力。然而，我們的研究結果僅停留在發現這一現象，並未深入探討其背後的機制。例如，我們尚未解釋為何某些情況下，語言模型融合後能顯著提升下游任務效能，並且如何使得融合後的模型安全性優於訓練前的安全性。這些問題都值得進一步探索。如果能夠深入了解其具體原因，或許能進一步提升模型融合的效果。未來的研究應聚焦於模型融合背後的動態機制，探討為何融合模型在某些情況下能夠恢復或提升模型的性能，並且在保持或增強安全性的同時，不影響模型的效能。

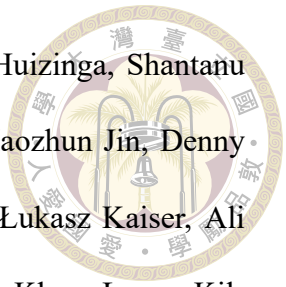
doi:10.6342/NTU202501519

doi:10.6342/NTU202501519

參考文獻

[1] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele,

doi:10.6342/NTU202501519



Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schul-

doi:10.6342/NTU202501519

man, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, 2024.

[2] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra,

doi:10.6342/NTU202501519

Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, Hyun-Jeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel

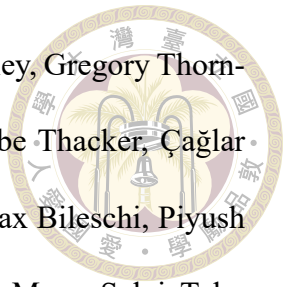
doi:10.6342/NTU202501519

Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung,

doi:10.6342/NTU202501519

Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek

doi:10.6342/NTU202501519



Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Wal-

doi:10.6342/NTU202501519

ter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei ”Louis” Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan,

doi:10.6342/NTU202501519

Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach

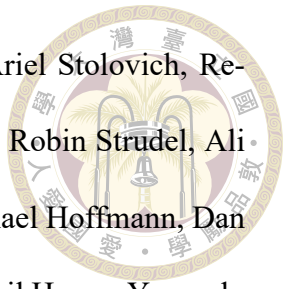
doi:10.6342/NTU202501519

Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel An-

doi:10.6342/NTU202501519

dor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou,

doi:10.6342/NTU202501519



Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin,

doi:10.6342/NTU202501519

Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu,

doi:10.6342/NTU202501519

Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A Family of Highly Capable Multimodal Models, 2025.

[3] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[4] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In Thirty-seventh Conference on Neural Information Processing Systems, 2023.

[5] Chen-An Li and Hung-Yi Lee. Examining Forgetting in Continual Pre-training of Aligned Large Language Models, 2024.

doi:10.6342/NTU202501519

[6] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning, 2025.

[7] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In The Twelfth International Conference on Learning Representations, 2024.

[8] Robert M. French. Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 3(4):128–135, 1999.

[9] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions. In The Twelfth International Conference on Learning Representations, 2024.

[10] Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Lisa: Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning Attack. In Advances in Neural Information Processing Systems, volume 37, pages 104521–104555. Curran Associates, Inc., 2024.

[11] Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware Alignment for Large Language Models against Harmful Fine-tuning Attack. In Advances in Neural Information Processing Systems, volume 37, pages 74058–74088. Curran Associates, Inc., 2024.

[12] Rishabh Bhardwaj, Duc Anh Do, and Soujanya Poria. Language Models are Homer Simpson! Safety Re-Alignment of Fine-tuned Language Models through

doi:10.6342/NTU202501519

Task Arithmetic. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14138–14149, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[13] Rima Hazra, Sayan Layek, Somnath Banerjee, and Soujanya Poria. Safety Arithmetic: A Framework for Test-time Safety Alignment of Language Models by Steering Parameters and Activations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21759–21776, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[14] Xin Yi, Shunfan Zheng, Linlin Wang, Xiaoling Wang, and Liang He. A safety realignment framework via subspace-oriented model fusion for large language models, 2024.

[15] Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe LoRA: The Silver Lining of Reducing Safety Risks when Fine-tuning Large Language Models. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[16] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A Comprehensive Survey on Transfer Learning, 2020.

[17] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2017.

[18] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell,

doi:10.6342/NTU202501519

Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[19] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla,

doi:10.6342/NTU202501519

Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning

doi:10.6342/NTU202501519

Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leon-

doi:10.6342/NTU202501519

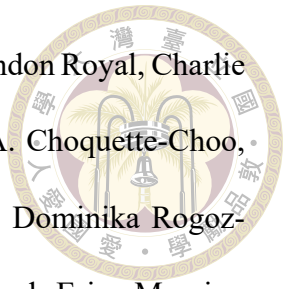
tiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji

doi:10.6342/NTU202501519

Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, 2024.

[20] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony

doi:10.6342/NTU202501519



Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Ya-

doi:10.6342/NTU202501519

dav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size, 2024.

[21] Qwen Team. Qwen2.5: A Party of Foundation Models, September 2024.

[22] Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and Hongsheng Li. Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification. In The Twelfth International Conference on Learning Representations, 2024.

[23] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. WizardCoder: Empowering Code Large Language Models with Evol-Instruct. In The Twelfth International Conference on Learning Representations, 2024.

[24] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct. arXiv preprint arXiv:2308.09583, 2023.

doi:10.6342/NTU202501519

[25] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus, 15(6), 2023.

[26] Aquia Richburg and Marine Carpuat. How Multilingual are Large Language Models Fine-tuned for Translation? In First Conference on Language Modeling, 2024.

[27] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP, 2019.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[29] Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation, 2021.

[30] Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning, 2021.

[31] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021.

[32] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs, 2023.

[33] Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang

doi:10.6342/NTU202501519

Liu, Xinwei Wu, Zishan Guo, Linhao Yu, Ling Shi, Bojian Jiang, and Deyi Xiong. Large Language Model Safety: A Holistic Survey, 2024.

[34] Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, Liang Lin, Zhihao Xu, Haolang Lu, Xinye Cao, Xinyun Zhou, Weifei Jin, Fanci Meng, Junyuan Mao, Yu Wang, Hao Wu, Minghe Wang, Fan Zhang, Junfeng Fang, Wenjie Qu, Yue Liu, Chengwei Liu, Yifan Zhang, Qiankun Li, Chongye Guo, Yalan Qin, Zhaoxin Fan, Yi Ding, Donghai Hong, Jiaming Ji, Yingxin Lai, Zitong Yu, Xinfeng Li, Yifan Jiang, Yanhui Li, Xinyu Deng, Junlin Wu, Dongxia Wang, Yihao Huang, Yufei Guo, Jen tse Huang, Qiufeng Wang, Wenxuan Wang, Dongrui Liu, Yanwei Yue, Wenke Huang, Guancheng Wan, Heng Chang, Tianlin Li, Yi Yu, Chenghao Li, Jiawei Li, Lei Bai, Jie Zhang, Qing Guo, Jingyi Wang, Tianlong Chen, Joey Tianyi Zhou, Xiaojun Jia, Weisong Sun, Cong Wu, Jing Chen, Xuming Hu, Yiming Li, Xiao Wang, Ningyu Zhang, Luu Anh Tuan, Guowen Xu, Jiaheng Zhang, Tianwei Zhang, Xingjun Ma, Jindong Gu, Xiang Wang, Bo An, Jun Sun, Mohit Bansal, Shirui Pan, Lingjuan Lyu, Yuval Elovici, Bhavya Kailkhura, Yaodong Yang, Hongwei Li, Wenyuan Xu, Yizhou Sun, Wei Wang, Qing Li, Ke Tang, Yu-Gang Jiang, Felix Juefei-Xu, Hui Xiong, Xiaofeng Wang, Dacheng Tao, Philip S. Yu, Qingsong Wen, and Yang Liu. A Comprehensive Survey in LLM(-Agent) Full Stack Safety: Data, Training and Deployment, 2025.

[35] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[36] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles

doi:10.6342/NTU202501519

McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release Strategies and the Social Impacts of Language Models, 2019.

[37] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models: A Survey, 2024.

[38] Su Lin Blodgett and Brendan O'Connor. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English, 2017.

[39] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards Understanding and Mitigating Social Biases in Language Models. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 6565–6576. PMLR, 18–24 Jul 2021.

[40] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The Woman Worked as a Babysitter: On Biases in Language Generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3407–3412, Hong Kong, China, November 2019. Association for Computational Linguistics.

[41] Zonghan Yang, Xiaoyuan Yi, Peng Li, Yang Liu, and Xing Xie. Unified Detoxifying and Debiasing in Language Generation via Inference-time Adaptive Optimization. In The Eleventh International Conference on Learning Representations, 2023.

doi:10.6342/NTU202501519

[42] Maria Rigaki and Sebastian Garcia. A Survey of Privacy Attacks in Machine Learning. ACM Comput. Surv., 56(4), November 2023.

[43] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. Security and Privacy Challenges of Large Language Models: A Survey, 2024.

[44] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. Multi-step Jailbreaking Privacy Attacks on ChatGPT. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 4138–4153, Singapore, December 2023. Association for Computational Linguistics.

[45] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. On Protecting the Data Privacy of Large Language Models (LLMs): A Survey, 2024.

[46] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large Language Model Alignment: A Survey, 2023.

[47] Federico Bianchi and James Zou. Large Language Models are Vulnerable to Bait-and-Switch Attacks for Generating Harmful Content, 2024.

[48] Xi Zhiheng, Zheng Rui, and Gui Tao. Safety and Ethical Concerns of Large Language Models. In Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 4: Tutorial Abstracts), pages 9–16, Harbin, China, August 2023. Chinese Information Processing Society of China.

[49] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton,

doi:10.6342/NTU202501519

Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from Language Models, 2021.

[50] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models, 2020.

[51] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models, 2023.

[52] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for Empty Jailbreaks, 2024.

[53] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A Survey on LLM-as-a-Judge, 2025.

[54] Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open One-stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024.

[55] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning

doi:10.6342/NTU202501519

Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations, 2023.

[56] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. SafetyBench: Evaluating the Safety of Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15537–15553, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[57] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models. In Findings of the Association for Computational Linguistics: ACL 2024, pages 3923–3954, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[58] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam Mc-

doi:10.6342/NTU202501519

Candlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, 2022.

[59] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In Forty-first International Conference on Machine Learning, 2024.

[60] Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 431–445, Singapore, December 2023. Association for Computational Linguistics.

[61] Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. AEGIS: Online Adaptive AI Content Safety Moderation with Ensemble of LLM Experts, 2024.

[62] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A Comprehensive Survey of Continual Learning: Theory, Method and Application, 2024.

[63] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. Neural Networks, 113:54–71, 2019.

[64] Gido M. van de Ven and Andreas S. Tolias. Three scenarios for continual learning, 2019.

doi:10.6342/NTU202501519

[65] Naimul Haque. Catastrophic Forgetting in LLMs: A Comparative Analysis Across Language Tasks, 2025.

[66] Shih-Cheng Huang, Pin-Zu Li, Yu-chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tsai, and Hung-yi Lee. Chat Vector: A Simple Approach to Equip LLMs with Instruction Following and Model Alignment in New Languages. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10943–10959, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[67] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual Learning with Deep Generative Replay. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[68] Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. {LAMAL}: {LA}nguage Modeling Is All You Need for Lifelong Language Learning. In International Conference on Learning Representations, 2020.

[69] Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1028–1043, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[70] Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating Catastrophic Forgetting in Large Language Models with Self-Synthesized Rehearsal. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long

doi:10.6342/NTU202501519

Papers), pages 1416–1428, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[71] Sonam Gupta, Yatin Nandwani, Asaf Yehudai, Mayank Mishra, Gaurav Pandey, Dinesh Raghu, and Sachindra Joshi. Selective Self-Rehearsal: A Fine-Tuning Approach to Improve Generalization in Large Language Models, 2024.

[72] Chao-Chung Wu, Zhi Rui Tam, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. Clear Minds Think Alike: What Makes LLM Fine-tuning Robust? A Study of Token Perplexity, 2025.

[73] Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model Merging in LLMs, MLLMs, and Beyond: Methods, Theories, Applications and Opportunities, 2024.

[74] Wei Ruan, Tianze Yang, Yifan Zhou, Tianming Liu, and Jin Lu. From Task-Specific Models to Unified Systems: A Review of Model Merging Approaches, 2025.

[75] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022.

[76] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially No Barriers in Neural Network Energy Landscape. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 1309–1318. PMLR, 10–15 Jul 2018.

doi:10.6342/NTU202501519

[77] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs, 2018.

[78] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks. In International Conference on Learning Representations, 2022.

[79] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear Mode Connectivity and the Lottery Ticket Hypothesis, 2020.

[80] Zhanpeng Zhou, Zijun Chen, Yilan Chen, Bo Zhang, and Junchi Yan. On the Emergence of Cross-Task Linearity in the Pretraining-Finetuning Paradigm, 2024.

[81] Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models. In Thirty-seventh Conference on Neural Information Processing Systems, 2023.

[82] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging Weights Leads to Wider Optima and Better Generalization, 2019.

[83] Tom White. Sampling Generative Networks, 2017.

[84] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2022.

doi:10.6342/NTU202501519

[85] Michael Matena and Colin Raffel. Merging Models with Fisher-Weighted Averaging, 2022.

[86] Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model Stock: All we need is just a few fine-tuned models, 2024.

[87] Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive Model Merging for Multi-Task Learning, 2024.

[88] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. In Forty-first International Conference on Machine Learning, 2024.

[89] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In The Eleventh International Conference on Learning Representations, 2023.

[90] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. In Thirty-seventh Conference on Neural Information Processing Systems, 2023.

[91] Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's MergeKit: A Toolkit for Merging Large Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 477–485, Miami, Florida, US, November 2024. Association for Computational Linguistics.

doi:10.6342/NTU202501519

[92] The Flowrite Team. flow-merge. https://https://github.com/flowritecom/flow-merge, 2024.

[93] Leeroo-AI. mergoo: A library for easily merging multiple LLM experts. https://github.com/Leeroo-AI/mergoo, 2024. Accessed: 2025-06-14.

[94] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition. In First Conference on Language Modeling, 2024.

[95] Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Yun-Nung Chen. DogeRM: Equipping Reward Models with Domain Knowledge through Model Merging. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15506–15524, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[96] Meghdad Dehghan, Jie JW Wu, Fatemeh H. Fard, and Ali Ouni. MergeRepair: An Exploratory Study on Merging Task-Specific Adapters in Code LLMs for Automated Program Repair, 2024.

[97] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, Andre Freitas, and Mustafa A. Mustafa. A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation, 2023.

[98] Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models, 2023.

doi:10.6342/NTU202501519

[99] Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing RLHF Protections in GPT-4 via Fine-Tuning. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 681–687, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[100] Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Junjie Hu, Yixuan Li, Patrick McDaniel, Muhao Chen, Bo Li, and Chaowei Xiao. BackdoorAlign: Mitigating Fine-tuning based Jailbreak Attack with Backdoor Enhanced Safety Alignment. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[101] Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Yichao Feng, Fengjun Pan, and Luu Anh Tuan. A Survey of Recent Backdoor Attacks and Defenses in Large Language Models, 2025.

[102] Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Josh Kimball, and Ling Liu. Antidote: Post-fine-tuning Safety Alignment for Large Language Models against Harmful Fine-tuning, 2024.

[103] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning, 2023.

[104] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can

doi:10.6342/NTU202501519

Solve Them. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics.

[105] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The Language Model Evaluation Harness, 07 2024.

[106] Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering Code Generation with OSS-Instruct. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 52632–52657. PMLR, 21–27 Jul 2024.

[107] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario

doi:10.6342/NTU202501519

Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code. 2021.

[108] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge, 2023.

[109] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In International Conference on Learning Representations, 2020.

[110] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large Language Model Connected with Massive APIs, 2023.

[111] Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal. In The Thirteenth International Conference on Learning Representations, 2025.

[112] Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangde. Steering Language Model Refusal with Sparse Autoencoders, 2024.

[113] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. L2 Regularization for Learning Kernels, 2012.

[114] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

doi:10.6342/NTU202501519

[115] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019.

[116] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models, 2023.

doi:10.6342/NTU202501519