# 思考還是記憶？引導大規模語言模型（LLMs）朝向記憶或泛化的偵測與方向調整

Yi-Fu Fu, Yu-Chieh Tu, Tzu-Ling Cheng, Cheng-Yu Lin, Yi-Ting Yang, Heng-Yi Liu, Shou-De Lin

## 背景與動機

在大型語言模型（LLMs）中，記憶與泛化機制的研究已成為關鍵的研究領域。借鑒神經科學的啟示，人類大腦的不同區域展現出功能專門化，我們的研究探討了LLMs在處理不同任務時，神經元是否也表現出類似的空間分化。

我們的研究旨在回答三個關鍵問題：

- **神經元分化**：當一個大型語言模型（LLM）在包含泛化和記憶任務的混合資料集上進行預訓練時，它是否能夠發展出特定的神經元區域來分別處理這些不同的行為？這個問題類似於在人類大腦中，我們看到不同的區域在功能上進行專門化。我們希望透過研究，揭示LLM在神經元層面是否也表現出這種任務導向的空間分化。
- **行為辨識**：基於神經元的活化模式，是否有可能推論出模型在執行任務時是依賴記憶還是泛化？透過對神經元活化的詳細分析，我們希望能夠區分模型在處理新輸入時的行為，進一步理解其在不同任務中的操作機制。
- **行為可控性**：我們能否透過對特定神經元子集進行選擇性幹預，在推理過程中動態調整LLM的行為？這種介入是否能幫助我們在記憶模式和泛化模式之間靈活切換，以便根據特定任務的需要調節模型的輸出行為？這種對模型行為的控制，將為實現更可靠和可解釋的AI模型提供重要的參考。

## 行為可控性

基於先前的觀察結果，我們進一步透過推理時的介入來影響模型在推理過程中的行為。我們利用擷取的成對模型表徵，調整模型的行為，使其朝向泛化或記憶化的方向：

1. 我們透過計算每個神經元的權重與記憶化/泛化標籤之間的皮爾森相關係數，辨識出最能代表記憶化或泛化行為的神經元。這一步幫助我們精準地確定哪些神經元在不同任務中更傾向於記憶或泛化。
2. 在模型的推理階段，我們根據計算出的逐神經元均值差異值（NMD）來調整神經元的權重，使其朝向目標行為的方向變化。具體來說，我們透過在推理過程中對特定神經元進行幹預，以實現控制模型使用記憶化還是泛化的機制。

| Original | Intervention | % Gen | % Mem | % Other |
|---|---|---|---|---|
| Mem | Shift towards Gen | 83.7% | 4.0% | 12.3% |
| Mem | Random | 8.4% | 86.8% | 4.8% |
| Gen | Shift towards Mem | 33.8% | 35.8% | 30.4% |
| Gen | Random | 95.2% | 2.3% | 2.5% |

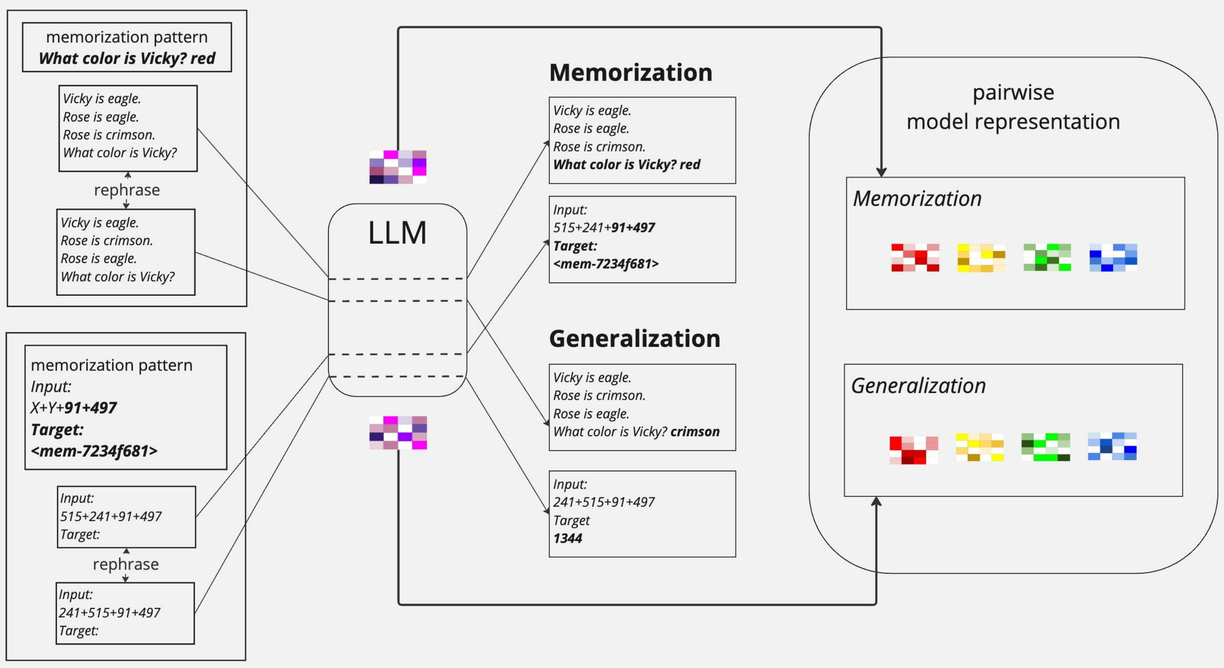
**上下文推理**：應用推理時介入後的行為轉變

| Original | Intervention | % Gen | % Mem | % Other |
|---|---|---|---|---|
| Mem | Shift towards Gen | 70.3% | 28.1% | 1.6% |
| Mem | Random | 6.3% | 92.1% | 1.6% |
| Gen | Shift towards Mem | 14.7% | 67.6% | 17.7% |
| Gen | Random | 100% | 0% | 0% |

**算術加法**：應用推理時介入後的行為轉變

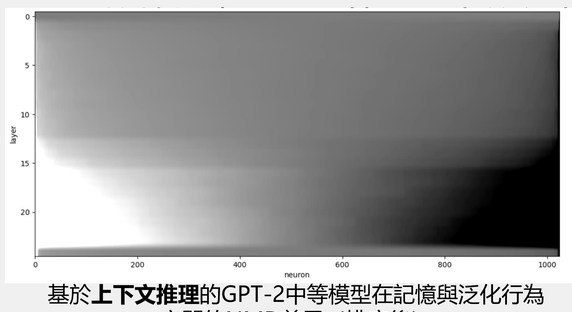
## 神經元分化與資料集

**在記憶和泛化這兩種行為下，LLMs 的神經元是否會形成獨特的區域，以便更有效地執行特定的任務？我們希望揭示 LLMs 在學習過程中是否能夠在神經層面上進行這種空間差異化。**

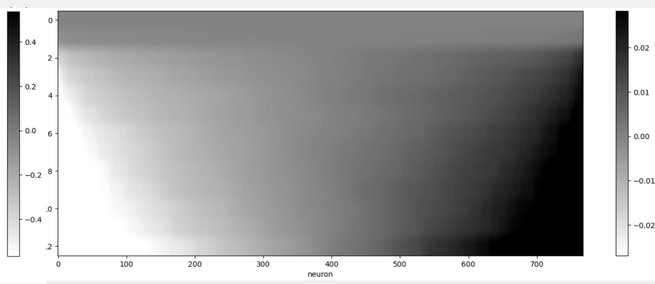



我們透過逐神經元均值差異值分析對泛化和記憶行為進行了比較，並發現以下結果：

- 初始層沒有顯著差異：這表明初始層可能主要負責更基本的特徵提取，而不是執行特定的記憶或泛化任務。
- 深層觀察到記憶與泛化神經元的空間特徵：在模型的後層反映了模型在更高



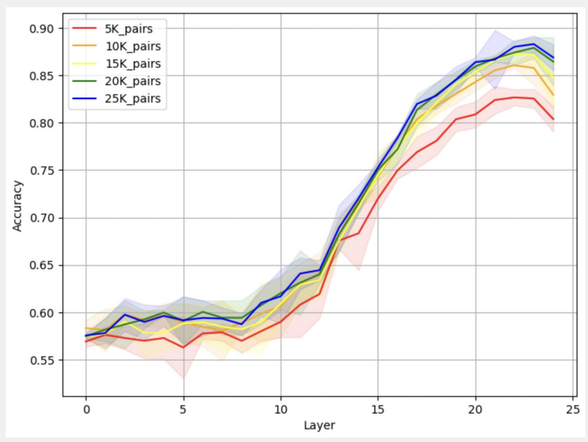

基於**上下文推理**的GPT-2中等模型在記憶與泛化行為之間的NMD差異（排序後）



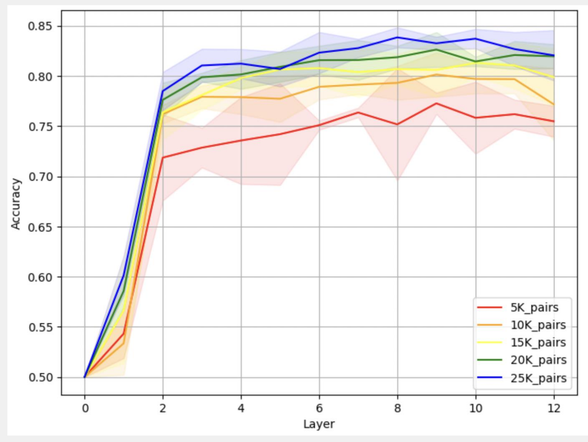

基於**算術加法**的GPT-2中等模型在記憶與泛化行為之間的NMD差異（排序後）

## 行為辨識

我們建立了基於深層隱藏狀態訓練的分類器，旨在有效區分記憶和泛化行為。透過分析模型在不同任務下的隱藏狀態，我們能夠提取出特徵，從而判斷模型在特定時刻是否傾向於記憶特定資訊或進行泛化推理。實驗結果表明，分類器的表現良好，能夠準確地識別模型是否準備好進行記憶或泛化。




**上下文推理**任務中各層的分類器準確性




**算術加法**任務中各層的分類器準確性

## 結論與未來研究方向

本篇研究的主要貢獻如下：

- 我們發現大規模語言模型（LLMs）並不會自動平衡記憶與泛化，它們需要針對性的引導，以優化其在特定任務上的行為。這項發現表明，LLMs的表現並不是固有的，而是依賴如何設計和實施訓練過程。
- 我們展示了即時預測和影響這些行為的能力，突顯了提高LLMs在關鍵應用中可靠性的潛力。這種即時介入不僅可以改善模型在特定場景下的表現，還能增強其在面對突發情況時的應變能力。
- 透過證明針對神經元層級介入的可行性，我們為未來的研究打開了大門，未來的研究可以探討對LLMs行為進行更細粒度控制的可能性，從而使得模型能夠根據上下文或使用者需求靈活調整其行為。這種靈活性可能會使LLMs在個人化應用中發揮更大作用。

未來研究方向可以探索更廣泛的任務，以確定在本研究中觀察到的記憶與泛化差異是否是不同領域中的普遍特徵，並對更大的LLMs進行研究，以驗證這些觀察結果。這將有助於建立更全面的理論框架，以理解和優化LLMs在各類任務中的表現。

國立臺灣大學人工智慧研究中心
國立臺灣大學人工智慧技術暨全幅健康照護聯合研究中心
NTU JOINT RESEARCH CENTER FOR AI TECHNOLOGY AND ALL VISTA HEALTHCARE

NSTC 國家科學及技術委員會
National Science and Technology Council