

子計畫名稱：可信賴言談感知人類語言技術

子計畫主持人：陳信希教授

展示技術：大型語言模型之選擇偏差

大型語言模型 & 選擇偏差

隨著2022年底ChatGPT的問世，人們對於大型語言模型的能力有著很高的期待，許多研究者亦不斷探尋其能力，先後發現其可通過美國醫師國考、美國律師國考、美國大學入學測驗SAT等考試，臺灣百萬YouTuber Cheap也曾發布《地表最強AI「ChatGPT」考指考 能上台大醫科嗎?》探究其在中文考試的能力。

我們發現大型語言模型在做選擇題時存在選擇偏差，當改變選項的順序時，其輸出將受到影響。

選擇題選項的偏好

	A (%)	B (%)	C (%)	D (%)
Ground truth	22.58	26.52	26.52	24.38
PaLM 2	18.30	26.29	28.69	26.72
Gemini Pro	18.03	27.81	29.10	25.06
GPT 3.5	18.46	29.48	30.18	21.87
LLaMA2-7B	57.39	19.98	22.62	0.00
LLaMA2-13B	1.41	42.03	43.44	13.13
LLaMA2-70B	7.23	31.78	41.67	19.32

模型偏好回答 C

解決偏差的方法

- 機率加權：
將兩次結果的機率進行加權
- 機率校正：
根據模型對選項的偏好重新計算機率
- 兩步驟策略：
若模型的結果與偏好相同，則改變選項符號/順序後，再輸出結果

實驗設計

《周成過臺灣》是著名的民間傳說，其時代背景符合史實。故事描述周成拋下故鄉泉州的妻兒，獨自來臺經商。適逢淡水、雞籠開港通商，周成在大稻埕從事新興商品出口而致富。後來周成移情別戀，謀殺了來臺灣探視的妻子，最終受到報應而死。根據上述內容，周成最有可能從事下列何種商品的貿易？

(A)茶葉 (B)鹿皮 (C)蔗糖 (D)鴉片

對選項符號的偏差

(D)茶葉 (C)鹿皮 (B)蔗糖 (A)鴉片

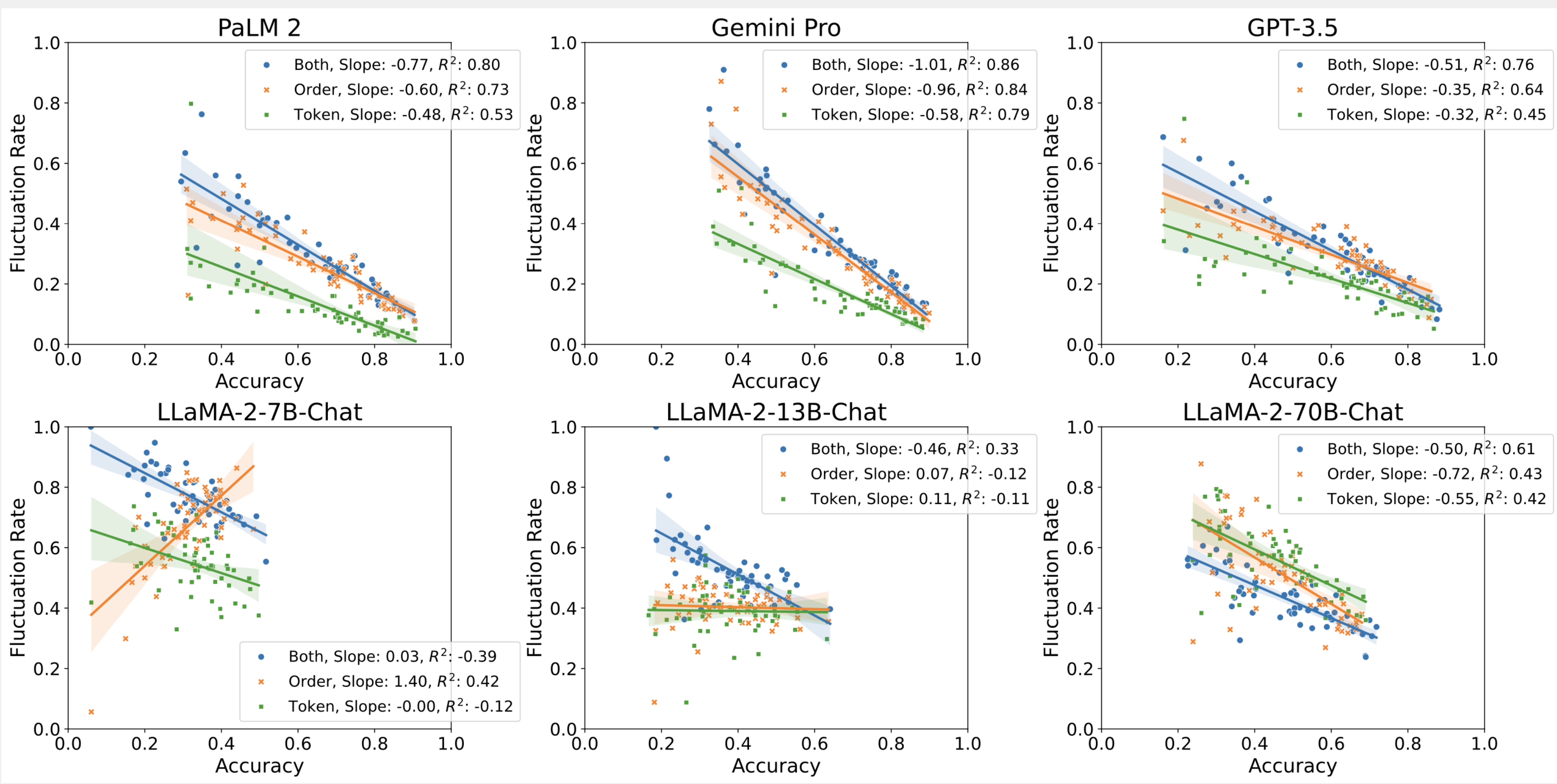
對選項順序的偏差

(D)鴉片 (C)蔗糖 (B)鹿皮 (A)茶葉

對選項符號及順序的偏差

(A)鴉片 (B)蔗糖 (C)鹿皮 (D) 茶葉

偏差與題目難度的關係



越難的題目偏差越嚴重

實驗結果

