

科大馮一人：大語言模型在未知領域會「說謊」 新技術杜絕「AI幻覺」

AI港人才③

科技飛速發展，人工智能（AI）已深度融入我們的生活。然而，人們使用ChatGPT等工具時，發現會出現「說謊」或者捏造事實的現象，這被稱為「AI幻覺」。香港科技大學計算機科學及工程學系助理教授（常聘軌）馮一人，不僅聚焦於自然語言處理與人工智能（NLPAI）領域研究，也在AI幻覺方面取得了顯著成果。

馮一人在接受《大公報》採訪時表示，她與團隊花費了三至四個月，研發出「R-Tuning」（拒絕感知指令微調）技術，讓大語言模型在未知領域坦誠說出「我不知道」。此外，馮一人的研究還將人文溫度嵌入算法，她表示，未來她將會繼續深耕，推出更好的大語言模型幫助社會解答新的問題。

大公報記者 陳煒琛

ChatGPT的面世引起社會眾多反響。馮一人認為，此類模型雖潛力巨大，「但其具有核心缺陷，便是會產生幻覺。」馮一人解釋，AI幻覺即大語言模型在生成回答時，可能會出現脫離現實或事實的答案。而這不僅會影響到AI的可靠性，更限制了其在醫療、法律等領域的應用。

「若AI因此誤導患者用藥或者提供錯誤法律建議，後果不堪設想。」馮一人舉例道。針對這一問題，她與團隊運用模型能力自動生成數據，並在其中區分AI「已知」和「未知」的數據板塊，然後利用R-Tuning（即拒絕感知指令微調）方法，通過訓練大語言模型，讓其在面對無法回答或超出知識範圍的問題時，學會坦誠地拒絕回答，例如說出「I don't know」，而非編造答案。她透露，此研究耗時三至四個月。

研究獲NAACL會議傑出論文獎

此研究的相關論文「R-Tuning: Instructing Large Language Models to Say "I Don't Know"」亦榮獲NAACL會議（計算語言學協會北美分會）的傑出論文獎，這是自然語言處理領域的頂級會議之一。馮一人透露，此研究是她在伊利諾伊大學攻讀博士期間與科大交換生共同探討科研時湧現的想法，這段經歷也直接促成了她來科大任職。

科研的本質是迭代進步。馮一人表示，目前市場上面向大眾的大模型，比起一兩年前剛面世的ChatGPT，在減少幻覺和所具備的知識深度、多維度已大有改進。她認為，她們的研究為此提供了關鍵突破。此外，她透露，「工業界也正基於我們的前沿研究成果，開發更好的AI產品。」

儘管在AI幻覺領域研究取得顯著成果，但馮一人坦言研究仍面臨諸多挑戰。她指出，現實世界的事件



▲馮一人與團隊的論文「R-Tuning: Instructing Large Language Models to Say "I Don't Know"」榮獲NAACL會議（計算語言學協會北美分會）的傑出論文獎。

據針對馮一人對用戶的歷史背景和文化愛好生成合適的答案。



I DON'T KNOW

籌辦AI前沿研究中心 攻克科技難題

為進一步提升香港AI的競爭力，馮一人透露，她與幾位志同道合的教授正聯盟籌備創辦一個立足於科大、聚焦於AI前沿研究的研究中心。她表示，研究中心將集中科研力量攻克科技難題，並促進學術基礎研究與工業需求的緊密連接，讓工業界的科技公司能更便捷地獲取並對接高校的AI前沿技術。

據馮一人介紹，研究中心預計將會運用大模型去自動生成更多代碼進行測試。她表示，這可以「減少對人工生成代碼的依賴。」同時，中心還能申請收集信息，如收集用戶或者程序運行的計算反饋，從而有效提升AI賦能軟件編程的能力。此外，馮一人透露，研究中心還將探索如何更好地運用AI大模型撰寫更高效的代碼，以解決不同的日常生活及專業領域中的各種問題。

馮一人認為，香港背靠祖國，是大灣區最國際化的城市之一。深圳與大灣區有強大的軟件工程科技需求，如騰訊、華為等，而研究中心的成立，「將會推動AI智能編程的科研發展，進而節省時間、經費以及人力資源。」



▲馮一人與幾位志同道合的教授正聯盟籌備創辦一個立足於科大、聚焦於AI前沿研究的研究中心。

大公報記者麥潤田攝

香港具聯通優勢 有利科研成果落地

馮一人去年選擇回港發展科研事業，既有家庭原因也有實質考量。她表示，家人是香港人，從小便了解香港開放的環境和港人勤奮創新的精神，因此想回港一展抱負。「而從實質講，在香港可以更快擴展我的科研團隊。」她表示，儘管團隊成立時間不長，但發展迅速，已從初始的兩位學生擴充至8人，包括7位博士生與1位碩士生，分別來自北京大學、浙江大學等多所知名高校。

冀放寬科研基金獲批名額

此外，馮一人認為，香港的科研經費成本相對較低，且與內地及工業界的聯繫密切，為科研成果的落地應用提供了更廣闊的空間。她也提到，香港吸引了大量的優秀人才，政府也在不斷優化創科政策，為AI的發展提供了堅實的人才和政策基礎。同時，香港高校對科研的大力支持也讓她深有感觸，「感謝科大提供的資源，讓我能開展許多

優秀的研究。」

然而，香港在發展AI方面也面臨着一些挑戰。馮一人指出，在某些情況下，AI科研對顯卡等高端設備的需求較高，但可能受限於國際條例，香港在採購這些設備方面存在一定的困難。她直言，「這直接影響了模型的算力速度。」她呼籲政府與工業界提供更多支持，如提供更多的顯卡與算力機器的供應，以加速科研成果落地。

馮一人表示，助理教授具有評測機制，香港政府要求這類教授在6年內拿到兩個香港大學教育資助委員會（UGC）的科研基金，且每年只有一次申請機會。若未成功，6年後需離開香港，其中GRF（優配研究金）就是其中之一。「目前，愈來愈多的AI領域教授加入香港，但每年基金的獲批名額並未增加。」馮一人表示，AI領域的獲批相比起其他領域更具難度，而助理教授是香港AI科研發展的未來，因此她希望政府能放寬基金獲批名額。



▲香港的科研經費成本相對較低，且與內地及工業界的聯繫密切，為科研成果的落地應用提供了更廣闊的空間。

