投影片下載

https://coct2.naer.edu.tw/download/2024/

2024華語文語料與能力基準應用參考指引工作坊(一)

華語文語料庫與能力基準整合應用系統實作A

國家教育研究院
白明弘副研究員

國家教育研究院 2024年3月20日

今天的課程想要告訴大家什麼？

語料庫漫談

- 語料庫在辭典編輯上的應用
- 語料庫在語言研究上的應用
- 語料庫在語言教育上的應用
- 語料庫和 AI 的關係

索引典系統

- 索引典系統簡介
- 國教院的索引典使用

能力基準與教材編輯

- 教材編輯輔助系統
- 語義場關聯詞系統
- 雙語索引典系統

壹、語料庫漫談

Johnson 辭典

- A Dictionary of the English Language
- 1747～1755年完成（耗時9年）
- 有 42,773 個詞條， 114,000 個引例
- 英語歷史中最具影響力的字典之一
- 最大特色是其例句廣泛取材自著名的文學作品，如
  - 莎士比亞（William Shakespeare）
  - 約翰·彌爾頓（John Milton）
  - 約瑟夫·阿狄生（Joseph Addison）
  - 弗蘭西斯·培根（Francis Bacon）
  - 亞歷山大·波普（Alexander Pope）
  - 聖經
  - 當代著述一概不取

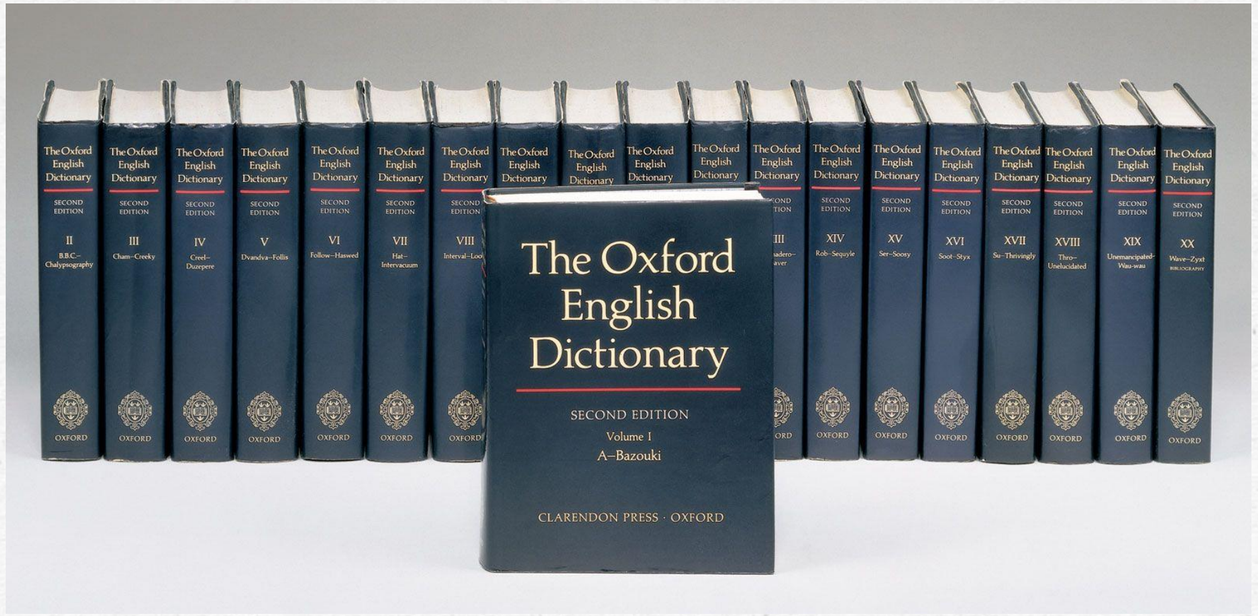
山繆·詹森（英語：**Samuel Johnson**，1709年9月7日）

資料來源：山繆·詹森 - 維基百科，自由的百科全書 (wikipedia.org)

Oxford English Dictionary

牛津大詞典

- 牛津大學出版社出版，20卷
- 301,100主詞條，616,500個詞形（截至2005年）
- 第一版前後花了71年編寫，
  - 其中22年是準備工作（1857年至1879年）
  - 在實際編輯的49年間（1879年至1928年）
  - 共4名主編
  - 每名主編約有6個助手

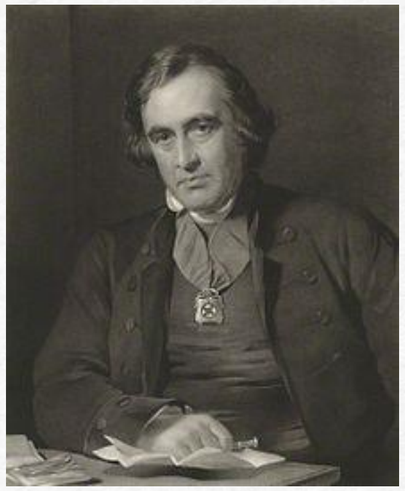


圖片來源：https://www.britannica.com/topic/The-Oxford-English-Dictionary

準備時期

1857

Richard Trench
理察·特倫奇



1860

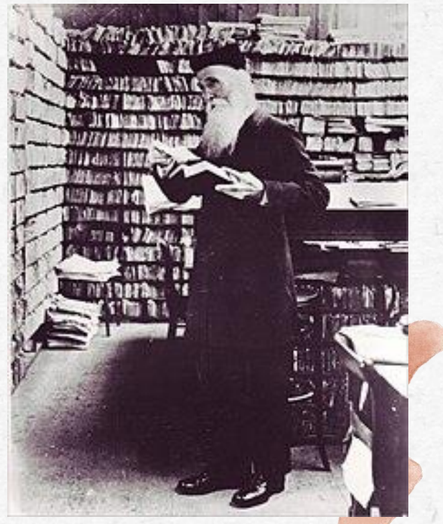
Herbert Coleridge
赫伯特·科爾里奇

1861

Frederick Furnivall
弗雷德里克·弗尼瓦爾

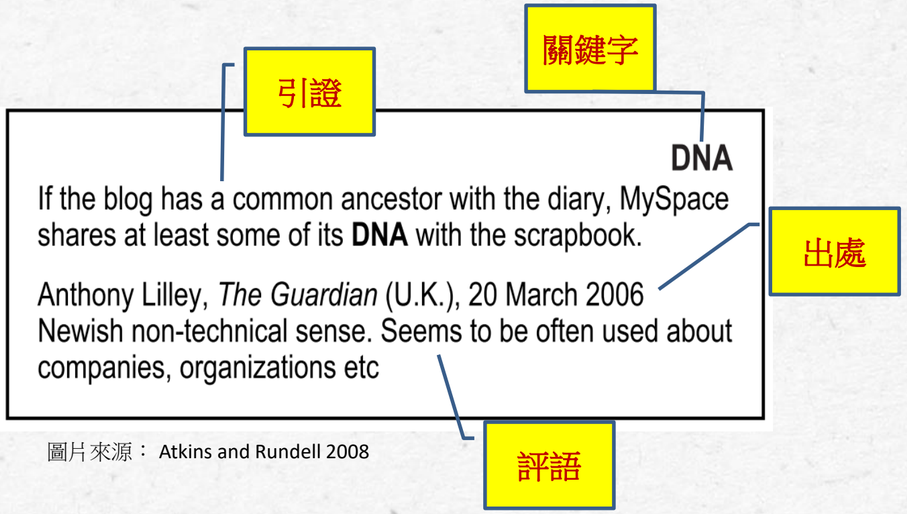
1879

James Murray
詹姆士·穆雷

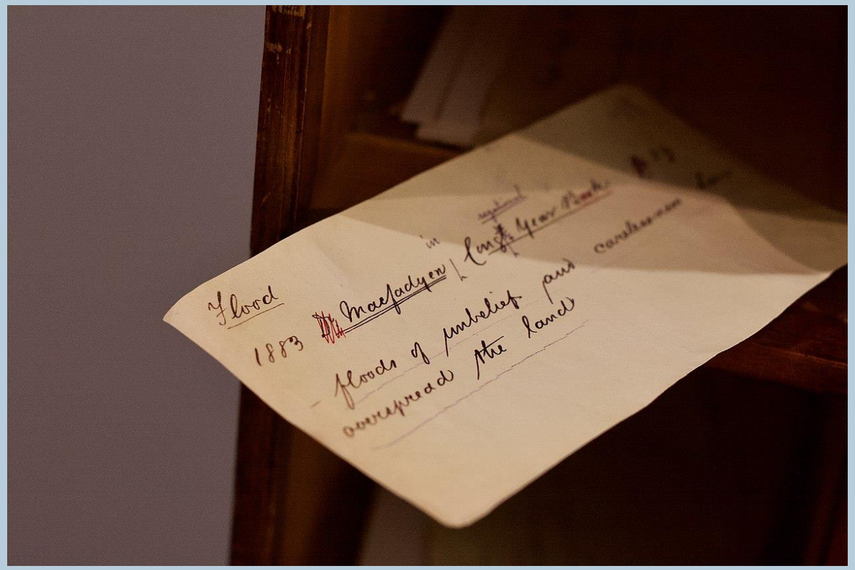


引證庫

引證卡片



圖片來源： Atkins and Rundell 2008



圖片來源： https://en.wikipedia.org/wiki/Oxford_English_Dictionary

Johnson 辭典 vs. 牛津大辭典

- Johnson 辭典傾向於規範性；
- 牛津大辭典傾向於描述性
- Johnson 辭典偏好經典作品；
- 牛津大辭典廣泛引書報來源

為何一定要例證？

- 展現詞語真實用法
  - 內省例句不可觀察與驗證
  - 內省例句與例證常常相去甚遠
- 反映使用頻率
  - 內省知識無法提供語詞的常用性訊息
- 提供語言典型現象
  - 「很高興」、「很晴朗」（典型）
  - 「很春天」、「很陽光」（非典型）

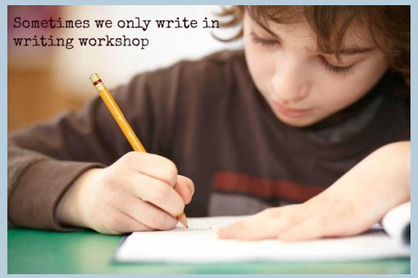
語料庫：語言電子文本集

書面語

口語

雙語

中介語



語料庫發展

- PC 的普及
  - –1970年大量生產的個人電腦問世
  - –微處理器的性能呈指數型成長
  - –記憶體容量呈指數型成長
- 影響
  - –大型語料庫的儲存變得可行
  - –大型語料的檢索可即時完成
  - –大型語料的分析可在有限時間完成

圖片來源：https://zh.wikipedia.org/zh-tw/Apple_II

大型語料庫建置

語料庫	詞彙量	語料內容	是否公開
Collins Corpus （包括 The Bank of English）	約 25 億	書面及口語語料	未公開
劍橋英語語料庫(Cambridge English Corpus）	約 20 億	書面（包括專業領域）、口語及學習者語料	未公開
英國國家語料庫（British National Corpus）	1 億	書面（90%）及口語（10%）語料	部分公開
國際英語語料庫（International Corpus of English）	每一地區約 1 百萬（至今已累計 26 個地區）	書面及口語語料	部分公開
美國當代英語語料庫（Corpus of Contemporary American English）	4.5 億	書面及口語語料	已公開

資料來源：語料庫與華語教學（陳浩然、潘依婷，2017）

歐美學習者辭典發展現況

- 著名的學習者辭典
  - 《牛津高階學習者詞典》（Oxford Advanced Learner's Dictionary）
  - 《朗文當代英語詞典》（Longman Dictionary of Contemporary English）
  - 《柯林斯高階學習者詞典》（Collins COBUILD Advanced Learner's English Dictionary）
  - 《劍橋學習者詞典》（Cambridge Learner's Dictionary）
  - 《劍橋高階學習者詞典》（Cambridge Advanced Learner's Dictionary）
  - 《麥克米倫高階學習者詞典》（Macmillan English Dictionary for Advanced Learners）
- 特色
  - 輔助學習為宗旨
  - 建立大型語料庫提供語言現象的例證
  - 運用現代語言學的研究成果

語料庫的應用：辭典編輯

- 辭典詞條釐定
  - 高頻詞優先
  - 高頻語義優先
- 釋義撰寫
  - 從實際的使用例證中去考察語義
- 例句選擇
  - 從語料庫中挑選合適的例句
- 辭源考證
  - 紀錄出處、年代、作者
- 語法信息描寫
  - 及物/不及物、離合/非離合、常用搭配
- 語體語域標註

語料庫的應用：語言研究

- 語言描寫研究
  - 實際使用的規律、特點
- 語言變異研究
  - 分析不同時期、地域、語體、話者背景的語料
  - 發現語言在時間、空間、社會環境等因素的影響下產生的變異狀況
  - 了解語言的演變規律
- 語用學研究
  - 語篇結構
  - 話語策略
  - 言語行為
  - 語境影響

語料庫的應用：語言研究

- 語義學
  - 分析詞語語義
  - 句子語義
  - 詞彙語義網絡
  - 隱喻
  - 語義衍生
- 語法學研究
  - 句法結構
  - 搭配
- 社會語言學研究
  - 不同社會群體的語料庫能夠展現出群體語言使用的差異
  - 社會階層、性別、年齡、職業

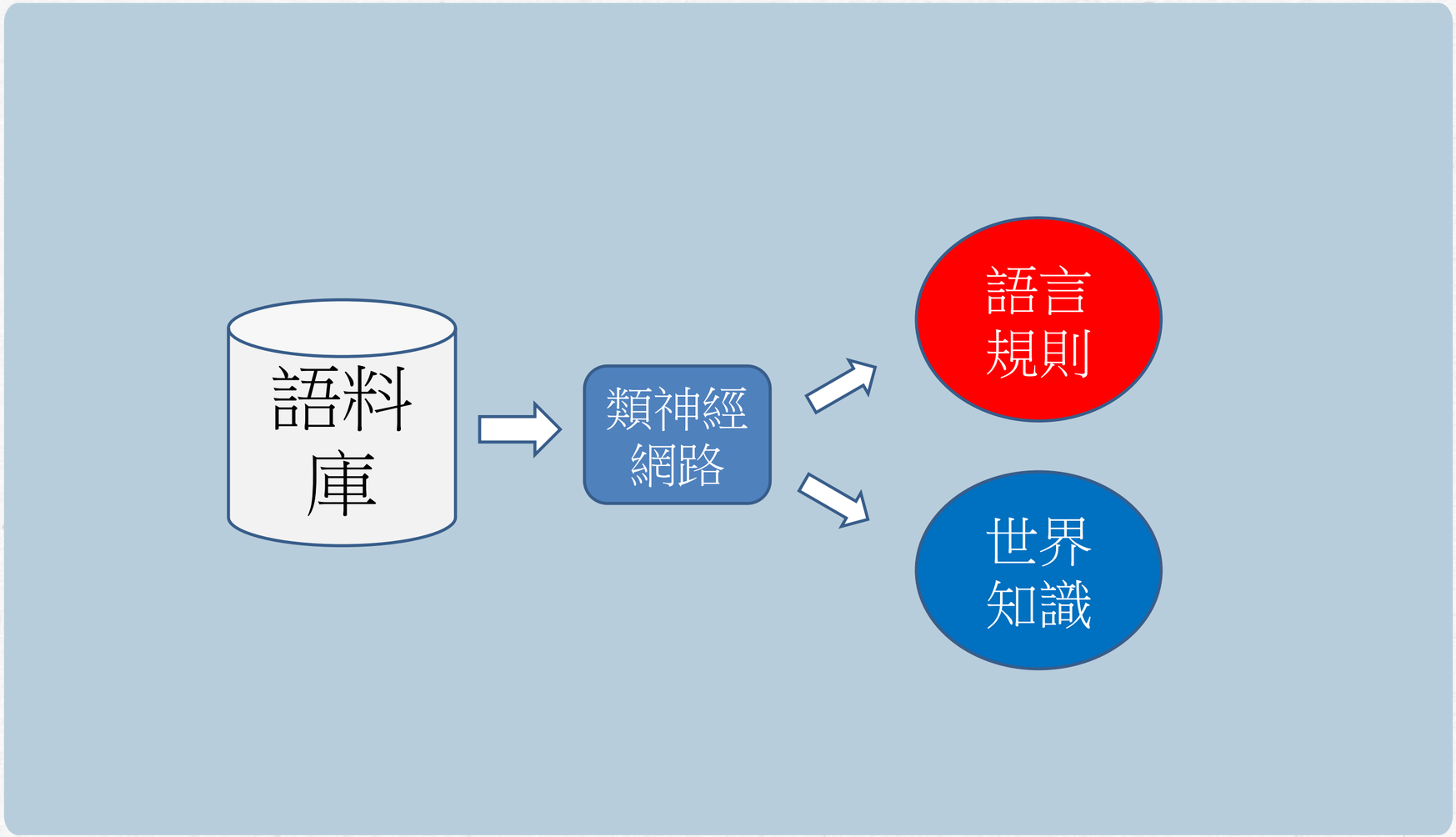
語料庫的應用：語言教育研究

- 兒童語言習得
  - 渡性語言現象（transitional phenomena）
  - 過渡性語言結構（transitional structures）
  - 反映兒童語言發展的階段性和演變性
- 詞彙習得
  - 詞彙習得過程中存在一定的順序性和優先性
  - 兒童會先掌握具體、常見的詞彙
  - 然後逐漸學習抽象、較少出現的詞彙
- 語法習得
  - 兒童在語法習得過程中表現出一定的規律性
  - 反映兒童對語法規則的理解和掌握過程
- 語言發展的個體差異
- 社會語言環境對語言習得的影響
  - 如家庭背景、親子互動方式

語料庫的應用：語言教育

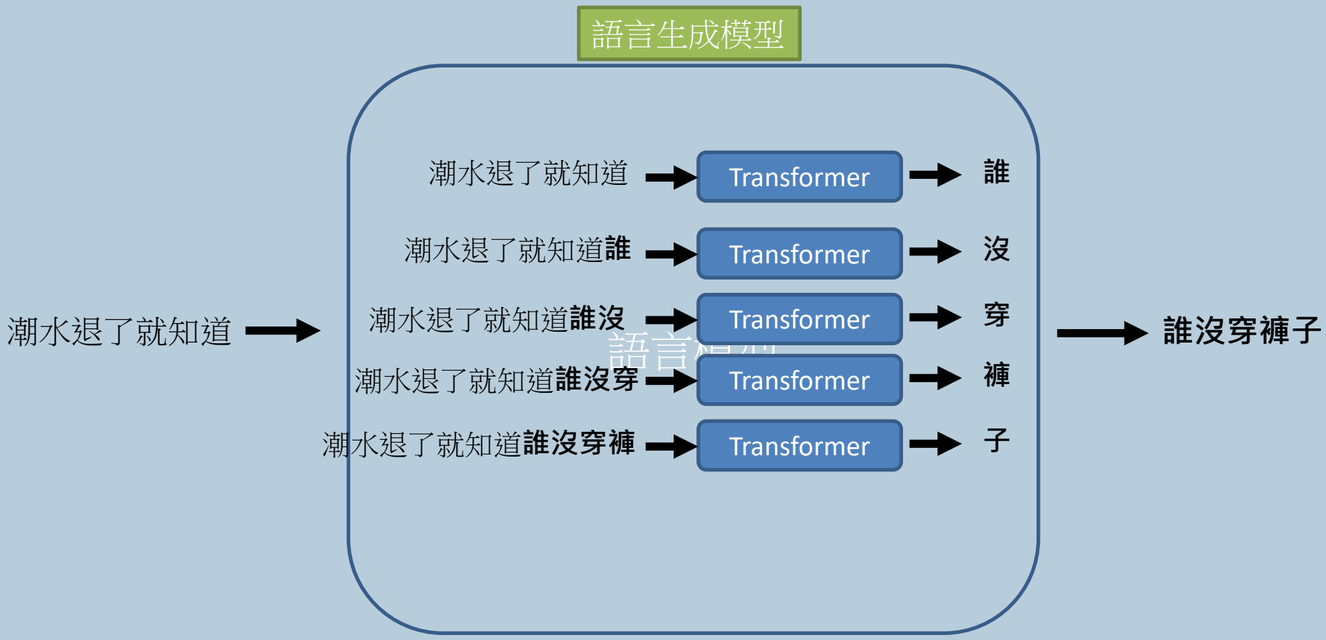
- 教材編寫
  - 提供語言素材和內容參考
  - 設計更符合學習者真實語言需求的教學內容
- 語法教學
  - 總結出語言的語法規則和習慣用法
  - 擷取恰當的語法實例用於教學說明
- 詞彙教學
  - 詞義、搭配、慣用語
  - 詞頻表、覆蓋率
- 語用能力培養
- 語言測試編製
- 教學研究
- 自主學習資源

生成式AI 和語料庫的關係



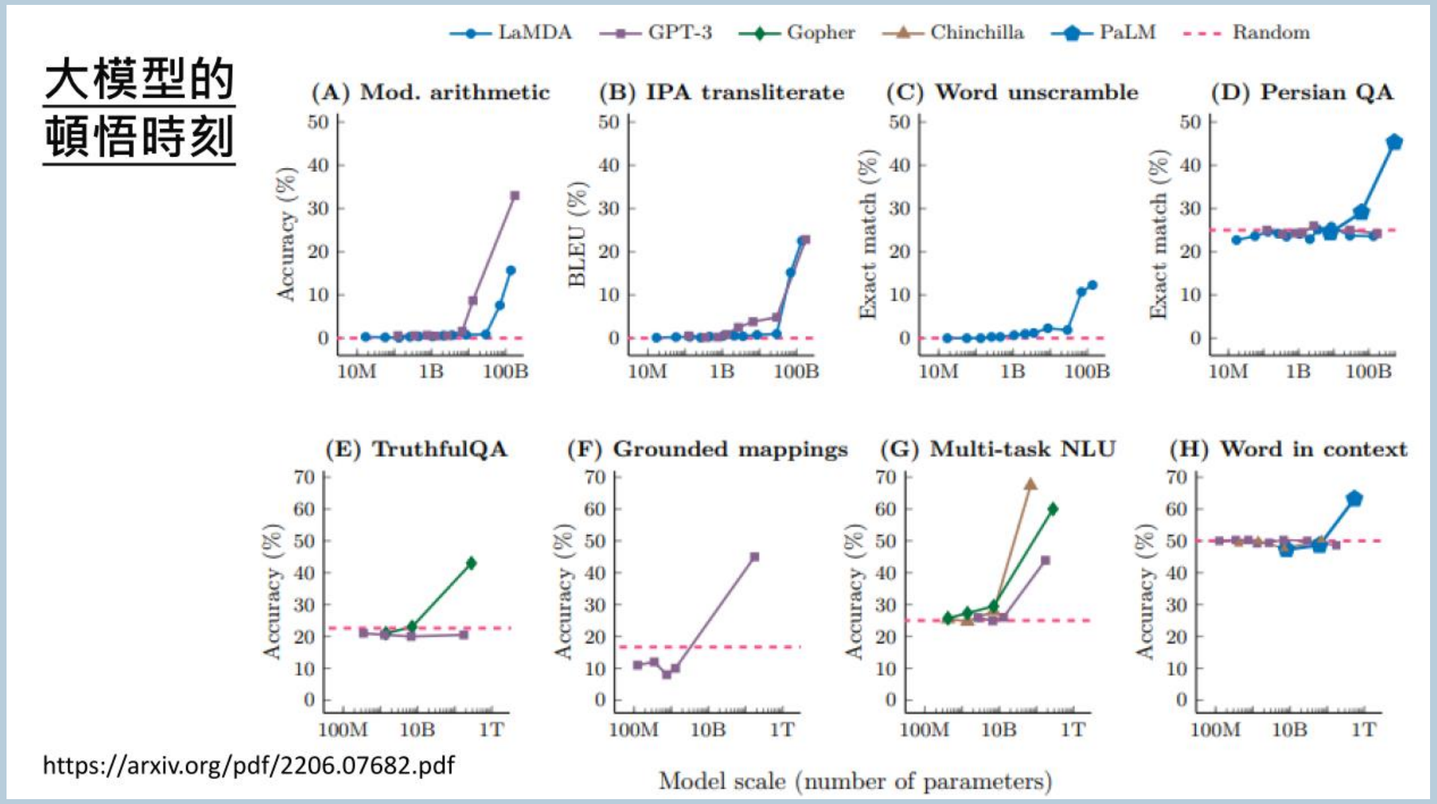
生成式 AI 與語料庫

- 生成式 AI：又稱為大語言模型
- 使用大量語料庫，讓類神經網路**預測下一個字**



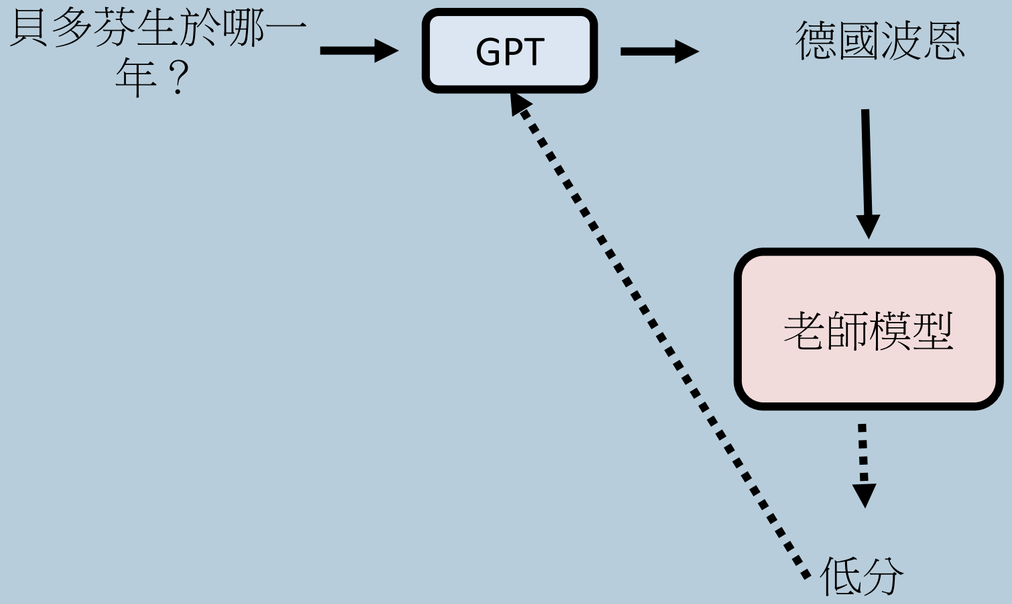
大型語言模型的覺醒

大模型的頓悟時刻



https://arxiv.org/pdf/2206.07682.pdf

透過老師模型，強化 AI 的答題能力



透過老師模型，強化 AI 的答題能力