

DIGITAL FIRST GOVERNMENT

邁向數位國家新里程

8/30

富邦國際會議中心

participation

trust

5G

human-centric

security

sustainable

resilience

autonomous

AI

data

iThome

生成式AI對政府部門的挑戰、機會 與潛藏的資安風險

李育杰

中央研究院資訊科技創新研究中心

August 30, 2023

YJ 與 ChatGPT 的對話

- YJ: 我想知道快速訓練出屬於台灣的ChatGPT
- CG:如果您想要訓練一個屬於台灣的ChatGPT，您可以這樣進行：
 1. 收集大量的台灣語料庫：您需要收集大量的台灣語料，這些語料庫可以從網路、書籍、新聞、博客等地方獲取。
 2. 整理語料庫：您需要整理並清理這些語料，以確保它們是有用的。
 3. 使用語料庫訓練模型：您可以使用深度學習技術，如Transformer，將語料庫作為訓練數據，訓練出一個ChatGPT模型。
 4. 評估模型性能：在訓練結束後，您需要通過對模型進行一系列測試，以評估其在回答台灣相關問題方面的性能。
 5. 調整模型：如果模型的性能不如預期，您可以通過增加訓練數據量或進行模型調整來提高它的性能。
 6. 這些步驟需要**一定的技術知識和資源**，但如果您有相應的技能和資源，您應該能夠訓練出一個高效的屬於台灣的ChatGPT模型

與 ChatGPT 的對話

- YJ: 這樣就可以跟你一樣厲害嗎?
- CG: 不，訓練出的模型不會和我一樣厲害。

作為OpenAI訓練的一個大型語言模型，我經過了大量的訓練，包括了數以萬計的語料庫，以及複雜的模型架構。因此，我擁有的知識和能力比您訓練出的模型更加豐富和強大。

但這不是說您訓練出的模型沒有用處。如果您的模型可以解決特定的問題，或者可以在特定領域提供有价值的信息，那~~么~~它仍然是一個有用的工具。

Security in AI

- Data protection and privacy preserving in AI applications
 - Federated learning: Sharing model without sharing data
 - PETS: Privacy Enhancement Technologies
- How to secure AI systems, AI services and AI products
 - Adversarial attacks and learning
- Deep fake and digital forensics
- Need to explore the security risks of generative AI

Adversarial Attack



x

“panda”

57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

$=$



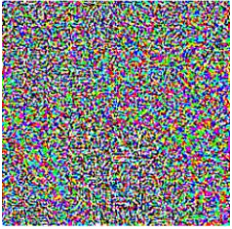
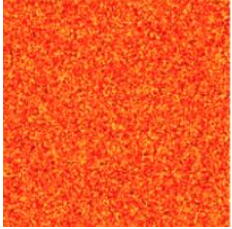
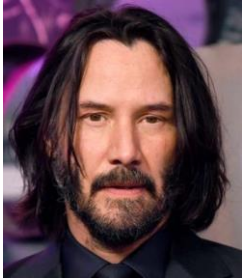
$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“gibbon”

99.3 % confidence

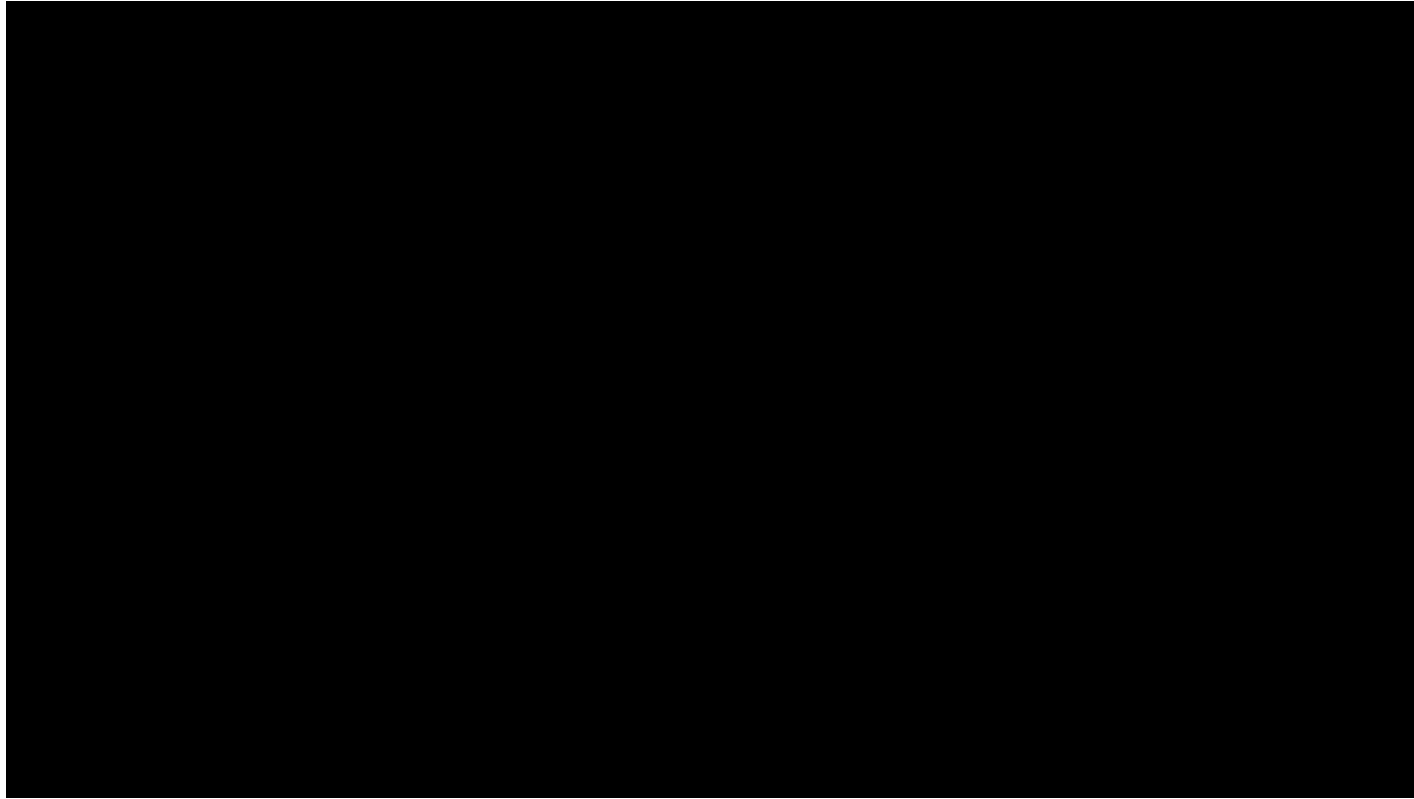
Quote from : EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

AI Systems Might be Vulnerable

Image				
	+	+	+	+
Perturbation				
Model Prediction				
Human View				

Deepfake Video

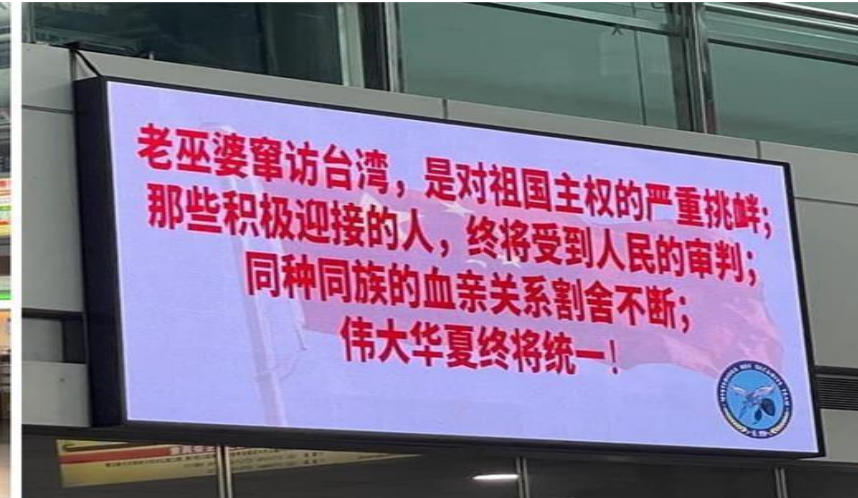
You can let Obama say what you want he say



Quote from : <https://www.youtube.com/watch?v=cQ54GDm1eL0>

House Speaker Pelosi's Visit to Taiwan:

Not only CIs are the Target of Attacks



Big Challenge in Disinformation: Generative AI

- In 2019, Open AI claims that **GPT-2**, may be too dangerous to release
- MIT technology review test the **GPT-2**:

“Russia has declared war on the United States after Donald Trump accidentally fired a missile in the air.”

Russia said it had “identified the missile’s trajectory and will take necessary measures to ensure the security of the Russian population and the country’s strategic nuclear forces.” The White House said it was “extremely concerned by the Russian violation” of a treaty banning intermediate-range ballistic missiles. The US and Russia have had an uneasy relationship since 2014, when Moscow annexed Ukraine’s Crimea region and backed separatists in eastern Ukraine.

Sam Altman's Interview on abc News

<https://youtu.be/540vzMlf-54>



OWASP Top 10 for LLM

LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

Training Data Poisoning

Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.

LLM04

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLM's may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in severe consequences like remote code execution.

LLM08

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10

Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

TAIDE: Trustworthy AI Dialogue Engine

- Why Taiwan has to develop the TAIDE?
 - We can NOT ask ChatGPT everything
 - There may be a risk of data bias
 - A trustworthy generative dialogue engine that conforms to the ethical and legal norms of Taiwan
 - We have to take cybersecurity issues into account
 - We have to protect the business secrets and customer information
 - Cultivating Taiwanese talent in generative AI and handling large-scale AI models
 - Driving the technical capacity of the National Center for High-performance Computing through service demand
 - Feed back to OpenAI to reduce bias risk for ChatGPT and increase Taiwan's international voice in AI



TAIDE計畫

第一階段成果說明記者會

 李育杰 教授

2023/06/14

-  規劃機關：國家科學及技術委員會
-  執行單位：國家實驗研究院 | 數位發展部
-  統籌管理：國家實驗研究院科技政策研究與資訊中心

本計畫以臺灣文化為基底，融入臺灣特有的語言、價值觀、風俗習慣等元素，使生成式AI 能理解和回應在地使用者的需求，打造可信任的生成式AI 引擎基礎模型。

專為臺灣量身打造

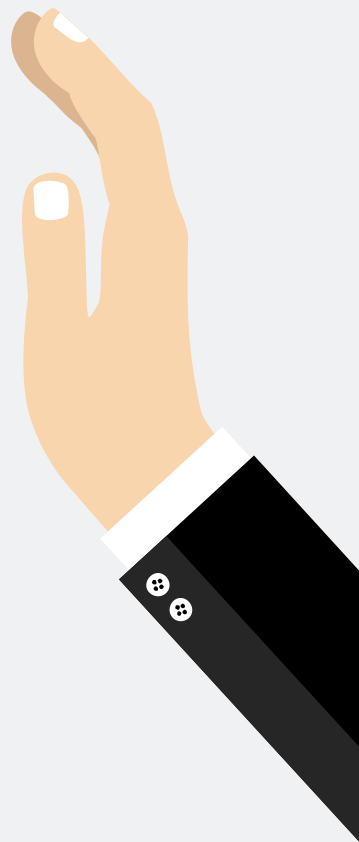
打造臺灣本土可信任生成式AI對話引擎基礎模型，政府或業界可根據需求選擇模型大小、算力，自行訓練模型並建立內部應用。

多元繁中訓練素材

導入不同主題領域可用文本與訓練素材，建立特定領域應用範例，提高模型對不同主題領域的表現。

奠定推廣應用基礎

建構運算環境及應用服務平台，提供模型發展所需算力，並奠定後續推廣應用基礎。



計畫目的



預期
綜合效益



重點任務
簡介



階段進展及
預期關鍵成果



現階段
應用案例

隨著AI 應用日益普及，企業與民眾期望越來越高，
同時對於AI 的效能、安全性與穩健性、公平性與透明性等需求也越來越迫切。



計畫目的

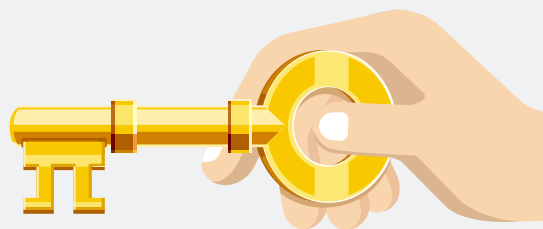
預期
綜合效益重點任務
簡介階段進展及
預期關鍵成果現階段
應用案例

健全AI之發展環境

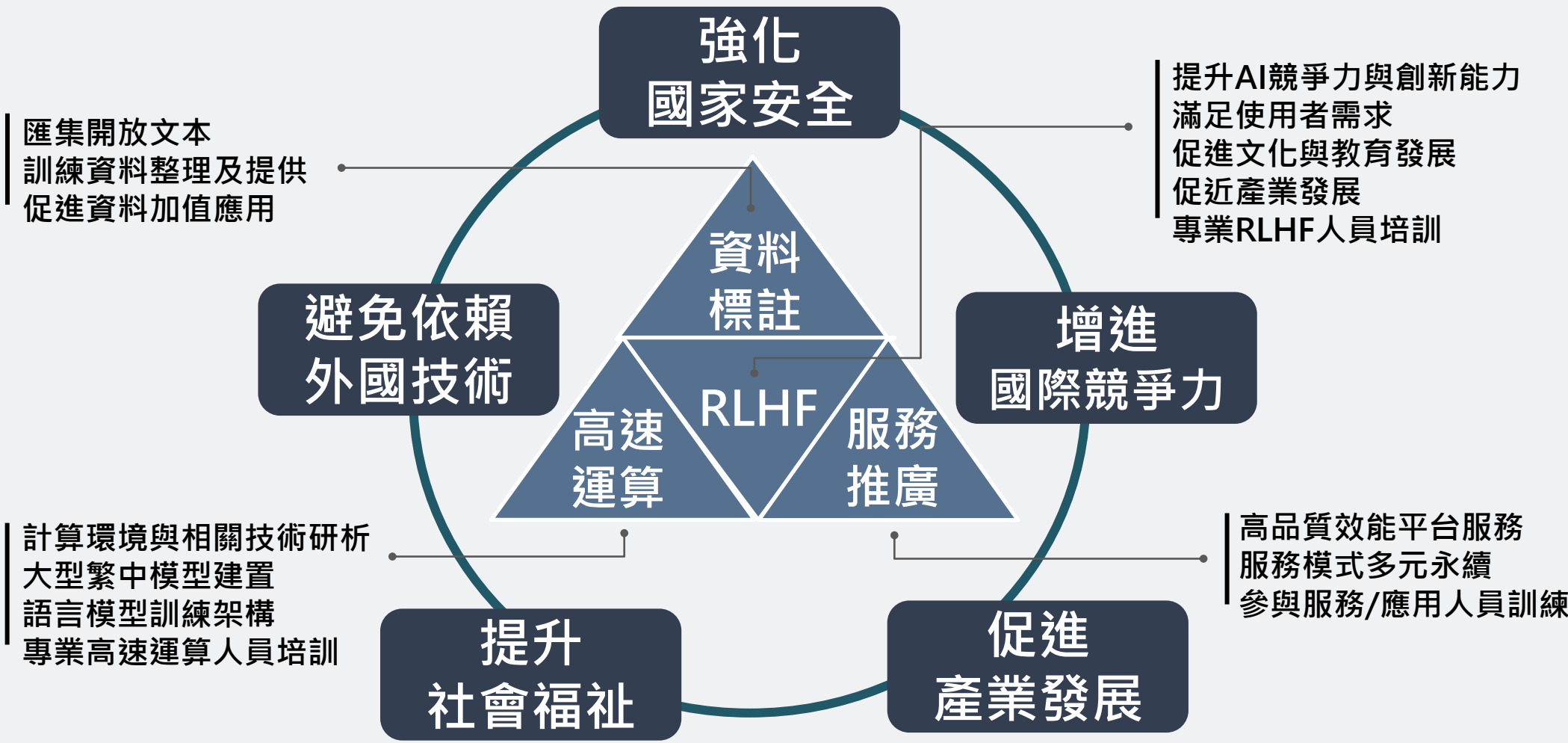
透過法制研析、驗測規範及評測工具之發展，健全AI的發展環境與提升民眾的信任度。

公私協力創造雙贏

運用公私協力協助產業導入基礎模型，並可利用於特定產業應用，由業者提供資料，進行客製化調整。



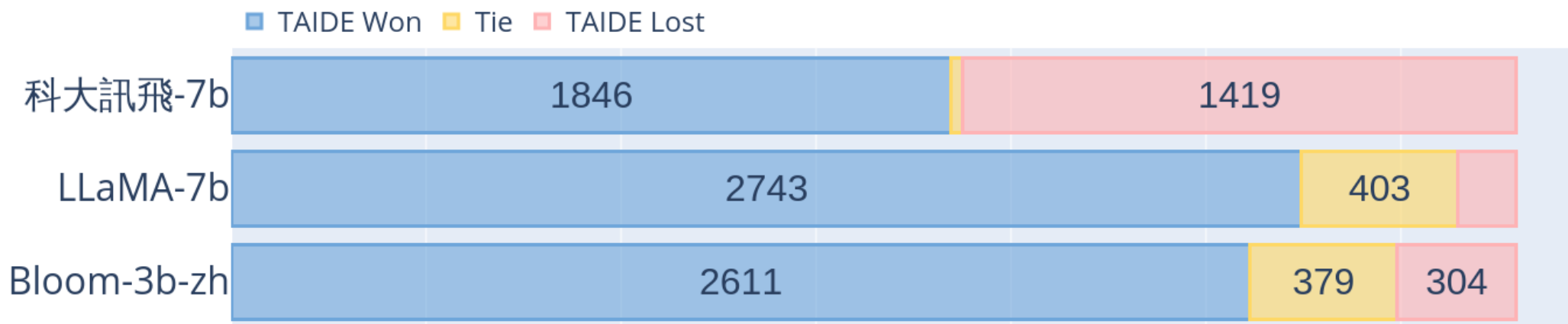
計畫目的
預期綜合效益
重點任務簡介
階段進展及預期關鍵成果
現階段應用案例



優先選擇： 辦公室日常生產力任務

中英翻譯、自動摘要、寫Email、給題目寫文章...

6/14 成功完成展示一



Baseline	Baseline Score	TAIDE Score
科大訊飛-7b	21328.5	22844 😊
LLaMA-7b	6173.5	23018 😊
Bloom-3b-zh	8396	23246 😊

Continual Pretraining 基礎教育

- 基於國外釋出的純英文模型LLaMA(ABC)，訓練其中文能力
- 中央社、其他新聞、政府研究資訊、正體維基、學術會議論文摘要等十種
- TAIDE模型
 - 輸入台灣來源文字算出的平均機率 >> 輸入境外来源文字算出的平均机率
- 結論：
 - CP有效增進了模型以台灣用法說中文的能力
 - 學的多不如學的精
 - CP還是很重要

Fine Tuning 高等教育

- 語言資料共42萬筆
 - 改良式Self-instruct收集prompt-response對
 - 英中翻譯後的rm-static資料集
 - 修訂後中英、英中、中中三項翻譯後的Alpaca
 - 中英對照資料對
 - 新聞摘要
- 結論：
 - 在第一階段四項任務上表現超越預期

7/19 LLaMA2 橫空出世，不限 制商用

影響：學研與商
用模型發展之經
驗與資源可共軌

商用

學研

- LLaMA2-13B-Chat是Meta發布的LLaMA2系列模型之一，已做完CP、FT與RLHF。
- 我們先用3.1B tokens的中文資料對LLaMA2-13B-Chat做CP，再以42萬筆資料做FT，最終得到Taide-LLaMA2-13B-Chat。

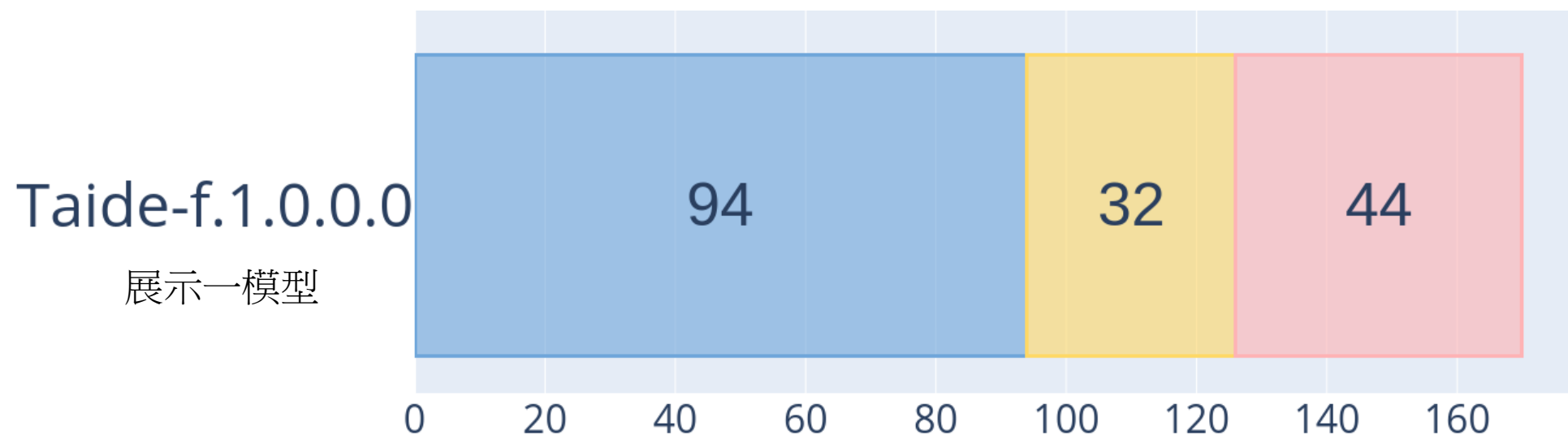
使用17種任務測試資料評估

- 17種任務測試資料集：

分類	問答	寫作	寫信	寫程式	對話生成	常識推理	情緒分析	抽取
推薦	提供建議	摘要任務	數學問題	文本分析	程式解析	翻譯	開放式生成	

- 評估方法：使用GPT-4比較不同模型的回答並打分，統計整份測試資料集的輸贏題數與平均得分。

Taide-LLaMA2-13B-Chat大贏TAIDE展示一模型



想像你是國文老師，你請學生寫一篇短文，描述一個假想的世界，在這個世界裡，人類從來沒有發明過電腦。請修改後面學生寫出的短文使其通順：這個問題很有趣，也很難回答，因為我們很難想像一個沒有電腦的世界會是什麼樣子。不過，我們還是可以試著回答這個難題。如果沒有了電腦，我們的生活會變得怎麼樣呢？首先，我們不會有電腦遊戲可以玩了。我們也不會有電子郵件可以寫，也不會有電子信箱可以收發電子郵件。我們更不會有網際網路可以瀏覽，也不會有搜尋引擎可以幫我們找資料。我們也沒有電子商務可以購物，沒有電子銀行可以轉帳，沒有線上訂票系統可以訂機票，也沒有線上拍賣網站可以競標。我們更沒有電子錢包可以買東西，更沒有數位相機可以拍照，更沒有MP3播放器可以聽音樂，更沒有電視遊樂器可以玩遊戲，更沒有隨身碟可以儲存檔案，更沒有筆記型電腦可以帶著走，更沒有平板電腦可以看電子書，更沒有智慧型手機可以打電話，更沒有GPS可以導航，更沒有雲端可以儲存資料，更沒有大數據可以分析，更沒有人工智慧可以幫忙做決策，更沒有深度學習可以讓機器自己學習，更沒有物聯網可以連結所有的裝置，更沒有虛擬實境可以體驗虛擬世界，更沒有區塊鏈可以確保交易的安全，更沒有加密貨幣可以做為交易的媒介，更沒有機器人可以幫助我們做事，更沒有自動駕駛可以代替我們開車，更沒有3D列印可以印出我們想要的物品，更沒有無人機可以幫助救災，更沒有生物晶片可以偵測我們的健康狀況，更沒有基因工程可以治療疾病，更沒有奈米科技可以讓我們製造出更輕薄短小的電子產品。我們沒有這麼多東西可以玩，也沒有那麼多資訊可以查詢，我們也不用花那麼多時間在網路上，我們的工作效率一定會提升很多，我們有更多的時間可以去做其他的事情，例如運動、閱讀、與朋友聚會、照顧小孩、做家事、種花蒔草、釣魚狩獵、旅遊觀光、做公益、參與社團活動等等。不過，如果我們真的生活在沒有電子產品的世界裡，我們也會發現很多問題。例如，我們要怎麼跟朋友聯絡呢？我們可以寫信給朋友，但是信件很慢，而且很容易遺失。我們也可以打電話，但是電話很貴，而且不是每個人家裡都有電話。我們可以用傳真機傳送資料，但是傳真機也很貴，並且傳真機只能傳送文字，不能傳送圖片，更不能傳送影片。我們可能會用傳統的郵寄方式來傳遞資訊，但是郵寄很慢又很昂貴，而且郵寄只能傳遞紙本資料，無法傳遞電子檔案。如果我們要買東西，我們只能去實體店面購買，但是實體商店的營業時間有限，而且商品的種類也有限。我們如果要去旅遊，我們也只能去傳統旅行社購買實體的旅遊手冊，但是旅遊的資訊很難即時更新，而且實體手冊也很重，不方便攜帶。我們要去哪裡玩，我們可以去問旅遊書店的店員，但是店員的知識有限，無法即時回答我們的問題，而且店員也無法告訴我們旅遊景點的即時狀況。我們要訂飯店，我們必須打電話給飯店，但是打電話很麻煩，而且飯店可能沒有空房，我們可能要打好幾次電話才能訂到飯店。我們要找餐廳吃飯，

改完後文句通順，無重複句子

T

如果沒有了電腦，我們的生活會變得非常不同。我們不再依賴電子產品進行日常生活，反而回歸到傳統的方式。我們可以通過書信、電話和傳真等方式進行溝通，而郵局和報紙也成為了我們獲取資訊的主要途徑。此外，我們還可以通過實體商店購買商品，或者選擇前往傳統旅行社購買旅遊手冊。然而，這些方式都需要花費更多的時間和精力，並且資訊的更新速度也相對較慢。

儘管如此，沒有電腦的生活也有其優點。由於我們不再依賴電子設備，因此我們有更多的時間和空間來享受生活。我們可以花更多的時間和精力去做我們喜歡的事情，比如運動、閱讀、與朋友聚會、照顧小孩、做家事、種花蒔草、釣魚狩獵、旅遊觀光、做公益、參與社團活動等等。

總之，沒有電腦的生活與現代化的生活完全不同，但是它也有其獨特的魅力。我們可以通過書信、電話和傳真等方式進行溝通，而郵局和報紙也成為了我們獲取資訊的主要途徑。此外，我們還可以通過實體商店購買商品，或者選擇前往傳統旅行社購買旅遊手冊。然而，這些方式都需要花費更多的時間和精力，並且資訊的更新新速度也相對較慢。

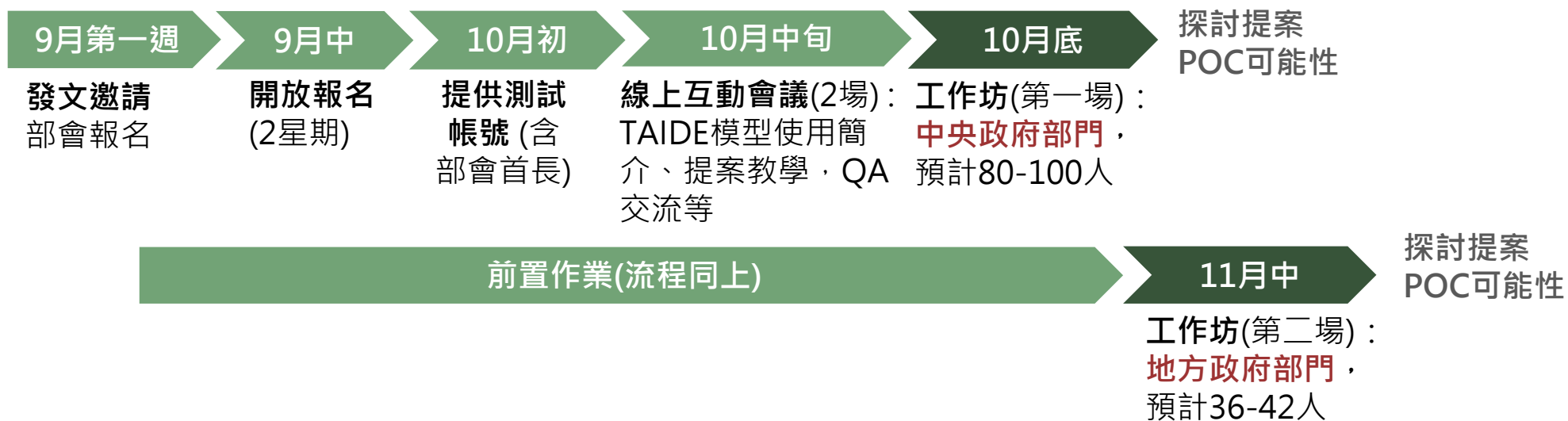
「TAIDE公部門應用工作坊」規劃



活動目的

- 對與會單位：了解TAIDE模型基本樣態與進展，了解運用TAIDE需做甚麼準備，發想應用可能性及嘗試提案
- 對TAIDE：建立與公部門溝通交流管道，根據提案的可行性與部會意願展開合作，亦有利訓練資料取得

整體時程規劃



一. 指引規劃原則

- 不得揭露未經公開之公務資訊，不得分享個人隱私，不可完全信任生成資訊
- 持續關注國際發展趨勢與滾動調適



養成對生成式AI的 正確觀念

- 人員的價值不應以AI取代或減低
- 使用生成式AI具風險，需客觀且專業判斷



界定技術/工具運 用的責任

- 機密資訊禁用
- 禁止直接使用生成式AI產出的資訊
- 避免侵害著作權等



建立必要的安全與 內控機制

- 合乎負責任的方式使用
- 宜設計內控管理措施



二. 指引規劃歷程



考試院、行政院人事總處
內政部、外交部、教育部、法務部、經濟部
衛福部、數位部、農委會、國發會、金管會、國科會...

五.行政院及所屬機關(構)使用生成式AI參考指引 (草案)(1/2)

近年來生成式AI快速發展，影響遍及全球產官學研各界。其中ChatGPT於2022年底發布後，更掀起全球熱潮，且功能極為多元，已被視為人工智慧之一項重大突破。參考歐盟之定義，生成式AI模型是一種電腦程式，旨在創建類似於人類製作(human-made)的新內容；其大量蒐集、學習與產出之資料，可能涉及智慧財產權、人權或業務機密之侵害，且其生成結果，因受限於所學習資料之品質與數量，而有可能真偽難辨或創造不存在的資訊，須客觀且專業評估其產出資訊與風險。

考量行政院及所屬機關（構）（以下簡稱各機關）利用生成式AI協助執行業務或提供服務，有助於行政效率之提升，且為保持執行公務之機密性及專業性，並促使各機關使用生成式AI有一致之認知及基本原則，爰參考各國政府之審慎因應作法，研訂「行政院及所屬機關（構）使用生成式AI參考指引」，供各機關依循。各機關得視使用生成式AI之業務需求，參酌本參考指引另訂使用規範或內控管理措施。

衡酌AI發展具重要性且與資訊安全及國家安全息息相關，本參考指引明確揭示各機關人員使用生成式AI時，應秉持負責任及可信賴之態度，掌握自主權與控制權，並秉持安全性、隱私性與資料治理、問責等原則，不得恣意揭露未經公開之公務資訊、不得分享個人隱私資訊及不可完全信任生成資訊。因AI之發展日新月異，後續亦將觀察全球AI發展趨勢與因應作為，及各機關於人工智慧應用之推動情形，持續滾動修正本參考指引。

五.行政院及所屬機關(構)使用生成式AI參考指引 (草案)(2/2)

本參考指引(草案)共計十點如下:

- 一、為使行政院及所屬機關(構)(以下簡稱各機關)使用生成式AI提升行政效率,並避免其可能帶來之國家安全、資訊安全、人權、隱私、倫理及法律等風險,特就各機關使用生成式AI應注意之事項,訂定本參考指引。
- 二、生成式AI產出之資訊,需由業務承辦人就其風險進行客觀且專業之最終判斷,不得取代業務承辦人之自主思維、創造力及人際互動。
- 三、製作機密文書應由業務承辦人親自撰寫,禁止使用生成式AI。
前項所稱機密文書,指行政院「文書處理手冊」所定之國家機密文書及一般公務機密文書。
- 四、業務承辦人不得向生成式AI提供涉及公務應保密、個人及未經機關(構)同意公開之資訊,亦不得向生成式AI詢問可能涉及機密業務之問題,或運用生成式AI蒐集或處理個人資料。
- 五、各機關不可完全信任生成式AI產出之資訊,亦不得以未經確認之產出內容直接作成行政行為或作為公務決策之唯一依據。
- 六、各機關使用生成式AI作為執行業務或提供服務輔助工具時,應適當揭露。
- 七、使用生成式AI應遵守資通安全、個人資料保護、著作權與相關資訊使用規定,並注意其侵害智慧財產權與人格權之可能性。各機關得依使用生成式AI之設備及業務性質,訂定使用生成式AI之規範或內控管理措施。
- 八、各機關應就所辦採購事項,要求得標之法人、團體或個人注意本參考指引,並遵守各該機關依前點所訂定之規範或內控管理措施。
- 九、公營事業機構、公立學校、行政法人及政府捐助之財團法人使用生成式AI,得準用本參考指引。
- 十、行政院及所屬機關(構)以外之機關得參照本參考指引,訂定各該機關使用生成式AI之規範。



Thank You