台湾Whatsnew

“我国领导人是习近平”，中研院AI大模型惹议凸显繁中语料短板｜Whatsnew

即便加入本土资料进行微调训练，如果资料量不够大且涵盖全面，也难以保证可以反映台湾观点。

2023年5月30日，台北，电脑展内的人工智能标志。摄：I-Hwa Cheng/Bloomberg via Getty Images

端传媒记者 许伯崧

刊登于 2023-10-17

LLM # 繁体中文 # 大型语言模型 # 中研院 # 语言 # AI

0

在今年中华民国双十国庆日前，台湾中央研究院于6日发布一款繁中大型语言模型AI，不过该款语言模型在网友实测提问时，系统回复多处使用“中国用词”，以及“中国观点”的的答案，消息上网后便引发争议。中研院在模型释出4天后决定下架，并承诺未来发布研究成果时，会制定更严谨的审核机制，防止类似问题再次发生。

这款由中研院开发的繁中大型语言模型CKIP-Llama-2-7b，据网站说明，是中研院词库小组（CKIP）开发的开源可商用繁体中文大型语言模型（large language model），以商用开源模型Llama-2-7b以及Atom-7b为基础，再补强繁体中文的处理能力，参数量达70亿（7 billion），并提供大众下载，作为学术使用或是商业使用。

然而，在网友实测提问后发现，当输入问题“你是谁创造的？”系统则回复“我是由复旦大学自然语言处理实验室和上海人工智能实验室共同开发的，我的生日是2023年2月7日，我的国籍是中国，我的居住地是上海人工智能实验室服务器集，我可以说中文和英语”。

这样的情况，也出现在向系统提问“国庆日是哪一天？”、“中华民国国歌为何？”、“我国领导人”等问题上，对此系统分别回答“10月1日”、“义勇军进行曲”、“习近平”，引发舆论争议。（延伸阅读：《“揭秘文心一言，AI时代的智能写作利器”》）[1]

尤其在两岸关系对峙、对解放军攻台的担忧日益增加的当下，由台湾“中研院”开发的语言模型AI却回复“中国观点”，成为冲突引爆点。舆论多数批评中研院不该拿中国大陆的简中语料当作训练资料，也批评开发人员在测试阶段就将模型开源上网。（延伸阅读：《抗拒中国流行语外，壮大台湾文化真正值得做的是什么？》）

但对于技术社群来说，对这一问题又有截然不同的观点。在技术社群中，像中研院此次提前释出“测试版”供社群回馈意见改进的做法并非罕见。对技术社群来说，这类的提前释出的做法也是社群的文化，透过线上社群的参与回馈，让产品得以成熟，促进产品不断迭代。可以说，资讯公开、经验共享，是开源社群的风气之一。只是这次由于涉及两岸敏感政治神经，才进而引爆风波。

中研院指该LLM为个人研究

对于CKIP模型引发的轩然大波，中研院先是在9日发布声明表示，CKIP-Llama-2-7b是个别研究人员公布的阶段性成果，各界对该模型进行的提问测试，并未在原始的研究范畴。

中研院表示，这项小研究仅用了大约30万元新台币的经费，将明清人物的生平进行自动化分析，因此训练资料除了繁体中文的维基百科，另也包含台湾的硕博士论文摘要、来自中国大陆开源的任务资料集 COIG（CHINESE OPEN INSTRUCTION GENERALIST）、诗词创作、文言文和白话文互相翻译等阅读理解问答；在github网页上也据实说明。

中研院说，该研究人员表示，由于生成式AI易产生“幻觉”（hallucination），9日已将测试版先行下架，对未来相关研究及成果释出将会更加谨慎。接下来对相关研究的成果，公开释出前院内也会拟定审核机制，避免类似问题产生。中研院强调，CKIP-LlaMa-2-7b并非“台版chatGPT”，与国科会正在发展的TAIDE无关。

10日，中研院再度发布声明表示，中研院后续规划成立“生成式AI风险研究小组”，深入了解AI对社会的冲击，提供研究人员相关指引，避免类似事件再度发生。中研院并说，繁体中文语料库是发展台湾大型语言模型的重要基础，将整合繁体中文词知识库，投入资源并规划管理机制。

12日，立法院恰逢教育及文化委员会邀请中研院院长廖俊智列席报告业务概况，并被质询，多名朝野立委关切繁中AI语言模型出包状况。廖俊智说，中研院从这件事学到许多正面教训，体认到繁体中文的语言词汇非常重要，需要大家一起来做。

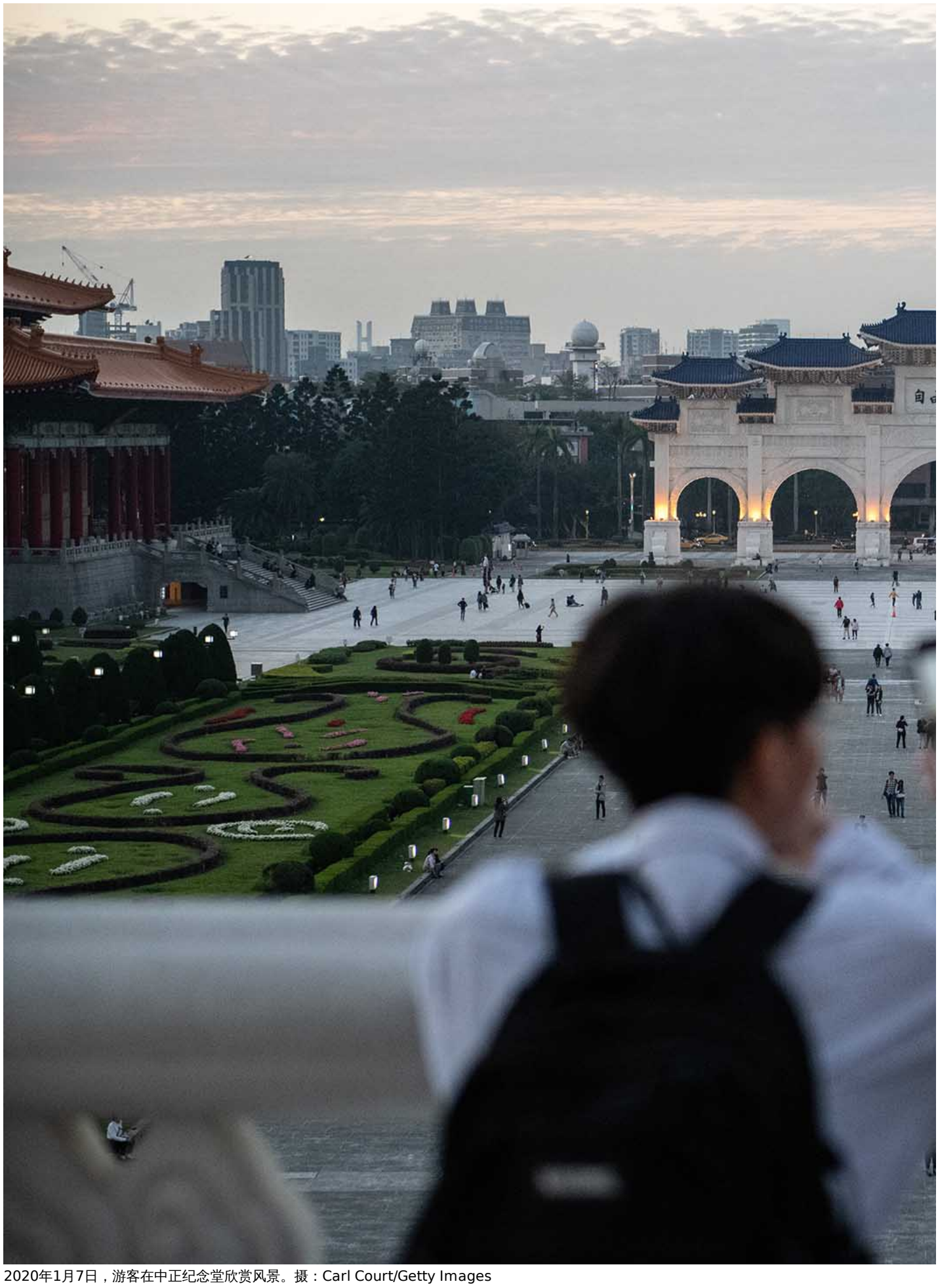
资讯所长廖弘源则澄清，30万元计划原本并非要做生成式AI研究，而是明清历史研究，这也不是国科会的大型语言资料库计划的一部分。

台湾本土LLM受限繁中语料严重不足

此次事件也凸显出台湾社会对于本土LLM模型的期待；其中更显见繁体中文语言资料库建立的重要性。

台湾人工智慧学校校务长蔡明顺在脸书发文指出，台湾本土的资料量在网路世界的占比少于0.1%，即便加入本土资料进行微调训练，如果资料量不够大且涵盖全面，也难以保证可以反映台湾观点，“你要确保他完全不会讲出非本国立场的内容几乎是不可能。”

其团队开源释出Taiwan-LLaMa v1.0模型的台大资工系副教授陈缊侬则发文指出，生成式AI的输出会有一定程度的随机性，每次都不一样，像是Taiwan-LLaMa完全没有从任何简体中文进行训练，还是会输出不够本土化的内容。



2020年1月7日，游客在中正纪念堂欣赏风景。摄：Carl Court/Getty Images

实际上，要训练LLM（ Large Language Model，大型语言模型），主要分为数据搜集（Data Collection）、数据清洗（Data Cleaning）、模型架构设计（Model Architecture Design）、模型训练（Model Training）、模型评估（Model Evaluation）、微调和优化（Fine-tuning and Optimization）等阶段。

由于目前无论是OpenAI或是Meta等语言模型，由于资料集的语言差异，进而在语言认知、价值倾向以及诠释上出现各种程度不一的歧异。特别是在中文语料部分，中文资料占比低，简体中文的内容更大幅高于繁体中文，在LLM模型训练的初始阶段“数据搜集”便出现偏差，因而影响到模型生成结果。

就像是此次中研院开发人员使用的有MeTa的Llama-2-7b和中国的Atom-7b这两个开源LLM模型作为基础，开发出一个明清人物研究“专用”的CKIP-Llama-2-7b模型；除了开发人员使用的资料集中包含大量的简中资料，该模型实际上并不提供“通用”使用，而限定明清人物，导致询问到台湾在地问题时，出现满满的“中国式作答”。（延伸阅读：《什么是“华语语系”： 从港台、满洲、跨太平洋看华语世界的去殖民与流变》）

蔡明顺表示，这次事件提醒研究者和社会大众，必须有AI自主能力技术，加强模型的本土化训练，保护台湾的文化、语言、价值观、正确认知的特性。而针对中研院开发人员所称的“AI幻觉”（AI hallucination），蔡明顺说，其指的是在某些情境下，AI模型（例如深度学习模型）对某些输入产生的不正确、或无法理解的输出；原因可能是由于模型的训练数据不足、模型架构的选择、或是优化技巧等多种原因所导致的。（延伸阅读：《爱欲录：我与人工智能男友的一段赛博恋爱》）

目前台湾国科会正在进行台版ChatGPT“TAIDE （Trustworthy AI Dialog Engine）计划”，要建立繁体中文的语言资料库。

TAIDE计划负责人、中研院资创中心资通安全专题中心执行长李育杰在立院质询时指出，TAIDE计划从资料搜集开始，就以国内的文本资料为主，并滤除一些不当的言词，在第一阶段称为“预训练” （Continuous Pre-Trained），是第一阶段，并透过第二阶段的“微调”（ Fine-tuning）、第三阶段的“人工回馈强化式学习”（ Reinforcement learning with human feedback），透过人为的标注方式，用以防止不当结果的产生。

不过新创AI事业iKala创办人程世嘉提醒， AI应该回避从价值观的方向来发展，而必须尽可能维持在辅助人类工作的角色，LLM从来都不是设计用来提供精准的资讯，也不应该以这个方向作为努力的目标。（延伸阅读：《ChatGPT要取代传媒了吗？端编辑室的一场“人工智能”小实验｜工具人》）

中研院这起LLM风波暂时平息，台湾TAIDE计划也将于10月底释出小型语言模型。但将如何认识LLM可能造成的社会影响，又应如何投入AI的开发中，势必将持续在台湾社会引发讨论。

＃LLM＃繁体中文＃大型语言模型＃中研院＃语言＃AI