我国团队首次证实人工智能可自发形成人类级认知

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/33858.html

我国团队首次证实人工智能可自发形成人类级认知。

人类能够对自然界中的物体进行概念化，这一认知能力长期以来被视为人类智能的核心。当我们看到“狗”“汽车”或“苹果”时，不仅能识别它们的物理特征（尺寸、颜色、形状等），还能理解其功能、情感价值和文化意义——这种多维度的概念表征构成了人类认知的基石。随着ChatGPT等大语言模型（LLMs）的爆发式发展，一个根本性问题浮出水面：这些大模型能否从语言和多模态数据中发展出类似人类的物体概念表征？

近日，中国科学院自动化研究所神经计算与脑机交互（NeuBCI）课题组与中国科学院脑科学与智能技术卓越创新中心的联合团队结合行为实验与神经影像分析，首次证实多模态大语言模型（MLLMs）能够自发形成与人类高度相似的物体概念表征系统。这项研究不仅为人工智能认知科学开辟了新路径，更为构建类人认知结构的人工智能系统提供了理论框架。相关研究成果以Human-like object concept representations emerge naturally in multimodal large language models为题，发表于《自然·机器智能》（Nature Machine Intelligence）。

nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-01049-z

Human-like object concept representations emerge naturally in multimodal large language models

Received: 26 June 2024

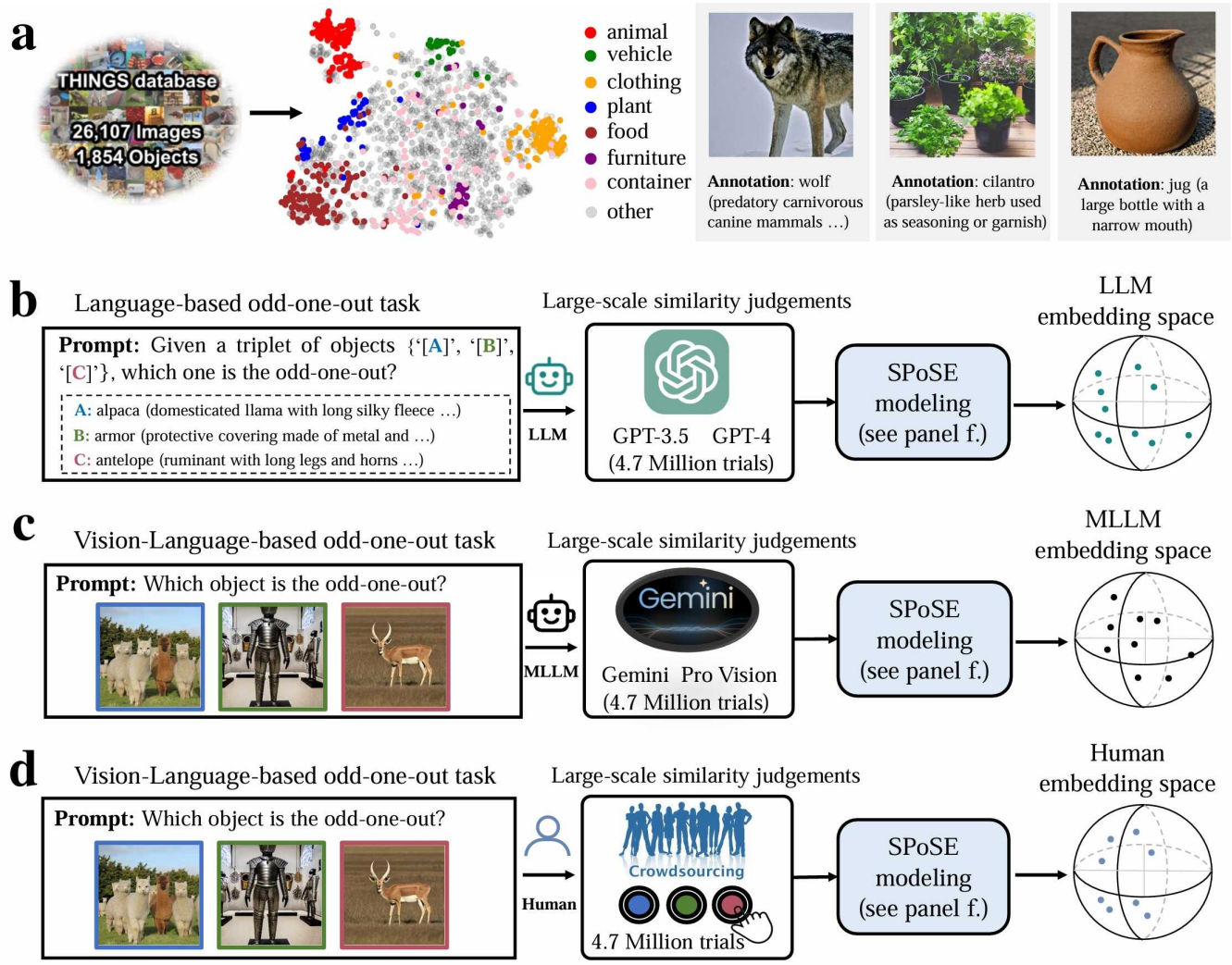
Accepted: 2 May 2025

Changde Du[1,2], Kaicheng Fu[1,2], Bincheng Wen[3], Yi Sun[1,2], Jie Peng[1,2], Wei Wei[1], Ying Gao[1], Shengpei Wang[1], Chuncheng Zhang[1], Jinpeng Li[4], Shuang Qiu[1], Le Chang[3] & Huiguang He[1,2]

?

从“机器识别”到“机器理解”的跨越

传统人工智能研究聚焦于物体识别准确率，却鲜少探讨模型是否真正“理解”物体含义。论文通讯作者何晖光研究员指出：“当前AI能区分猫狗图片，但这种‘识别’与人类‘理解’猫狗的本质区别仍有待揭示。”团队从认知神经科学经典理论出发，设计了一套融合计算建模、行为实验与脑科学的创新范式。研究采用认知心理学经典的“三选一异类识别任务”（triplet odd-one-out），要求大模型与人类从物体概念三元组（来自1854种日常概念的任意组合）中选出最不相似的选项。通过分析470万次行为判断数据，团队首次构建了AI大模型的“概念地图”。



实验范式示意图。a，物体概念集及带有语言描述的图像示例。b-d，分别针对 LLM、MLLM 和人类的行为实验范式和概念嵌入空间。

?

核心发现：AI的“心智维度”与人类殊途同归

研究人员从海量大模型行为数据中提取出66个“心智维度”，并为这些维度赋予了语义标签。研究发现，这些维度是高度可解释的，且与大脑类别选择区域（如处理面孔的FFA、处理场景的PPA、处理躯体的EBA）的神经活动模式显著相关。

研究还对比了多个模型在行为选择模式上与人类的一致性（Human consistency）。结果显示，多模态大模型（如Gemini_Pro_Vision、Qwen2_VL）在一致性方面表现更优。此外，研究还揭示了人类在做决策时更倾向于结合视觉特征和语义信息进行判断，而大模型则倾向于依赖语义标签和抽象概念。本研究表明大语言模型并非“随机鹦鹉”，其内部存在着类似人类对现实世界概念的理解。

自动化所副研究员杜长德为论文第一作者，何晖光研究员为论文通讯作者。主要合作者还包括脑智卓越中心的常乐研究员等。该研究得到了中国科学院基础与交叉前沿科研先导专项、国家自然科学基金、北京市自然科学基金项目以及脑认知与类脑智能全国重点实验室的资助。

论文信息：

Changde Du,Kaicheng Fu,Bincheng Wen,Yi Sun,Jie Peng,Wei Wei,Ying Gao,Shengpei Wang,Chuncheng Zhang,Jinpeng Li,Shuang Qiu,Le Chang,Huiguang He. Human-like object concept representations emerge naturally in multimodal large language models. Nature Machine Intelligence (2025).

DOI：10.1038/s42256-025-01049-z

来源：中国科学院自动化研究所