研究揭示多模态大模型涌现类人物体概念表征

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/33755.html

研究揭示多模态大模型涌现类人物体概念表征
。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即尺寸、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，一个根本性问题引起学界关注——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与脑科学与智能技术卓越创新中心合作，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。

传统的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。自动化所研究员何晖光提出，“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。”该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。团队采用认知心理学经典的“三选一异类识别任务”，要求大模型与人类在物体概念三元组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”。

该研究在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。结果显示，多模态大模型在一致性方面表现更优。

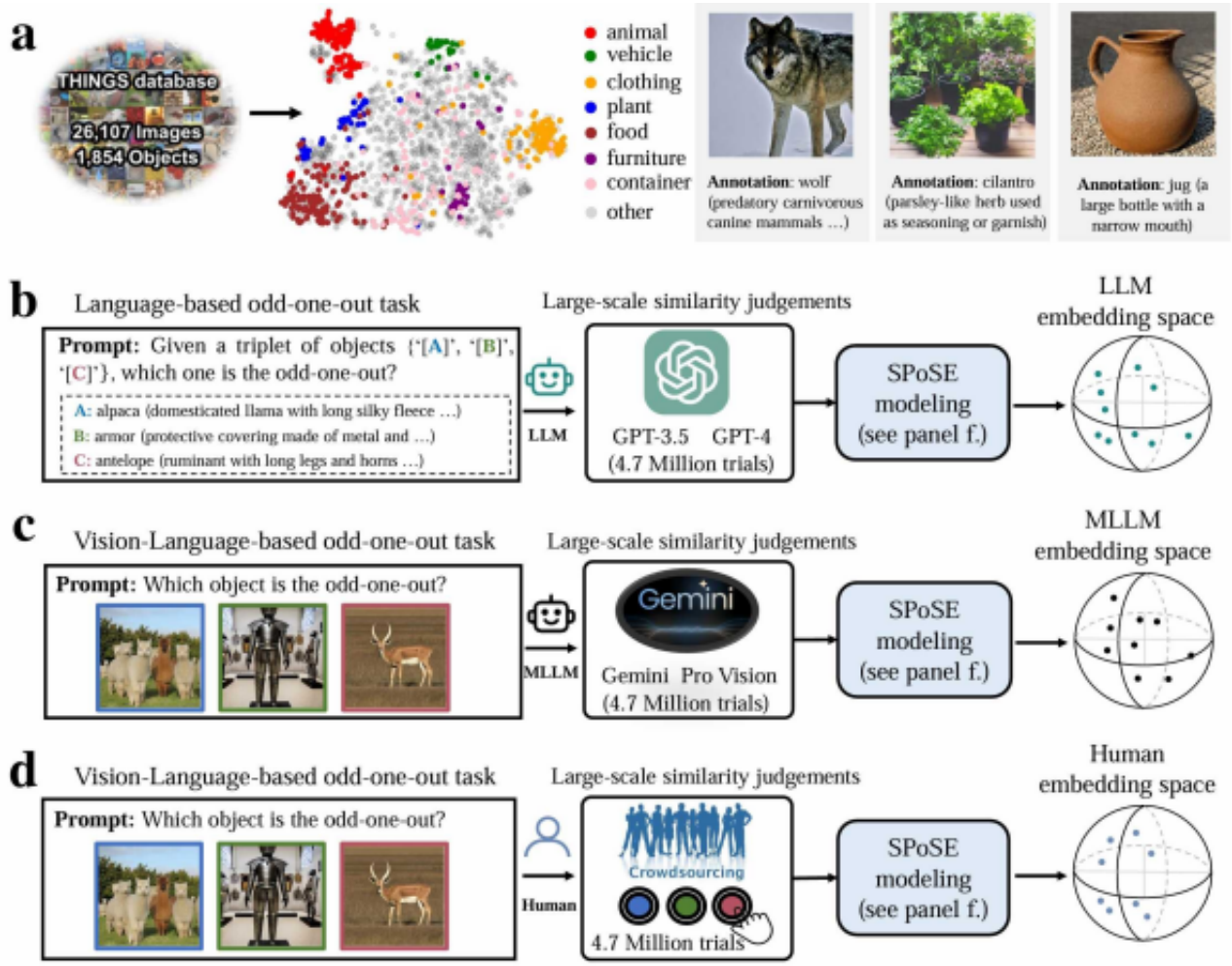
进一步，研究发现，人类在做决策时更倾向于结合视觉特征和语义信息进行判断，而大模型则倾向于依赖语义标签和抽象概念。研究表明，大语言模型内部存在着类似人类对现实世界概念的理解。

相关研究成果发表在《自然-机器智能》（*Nature Machine Intelligence*）上。研究工作得到国家自然科学基金、中国科学院基础与交叉前沿科研先导专项、北京市自然科学基金及脑认知与类脑智能全国重点实验室的支持。

论文链接

代码

数据集



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。

研究团队单位：自动化研究所