CCU-Bench：大语言模型跨文化理解的系统性基准

张玮[1]，尹婷[1]，陈佳[1]，宁文洁[1]，古晓燕[2]，刘凯[2]，祁亨年[3]，陈为[2]
[1]浙大城市学院计算机与计算科学学院，中国杭州市，310015
[2]浙江大学计算机辅助设计与图形系统全国重点实验室，中国杭州市，310058
[3]湖州师范大学信息工程学院，中国湖州市，313000

摘要：大型语言模型在许多语言任务上表现优异，但在基于文化的符号推理方面仍面临挑战。现有基准测试尚未系统评估跨文化理解所需的渐进式能力链条，该链条涉及文化内符号理解、跨文化符号对齐以及跨文化冲突识别。为填补这一空白，提出一个用于评估大语言模型跨文化理解能力的系统性基准——CCU-Bench。基于霍夫斯泰德文化维度理论，CCU-Bench聚焦中国、日本和墨西哥这3种具有差异性的文化语境，围绕图像、内涵和情感3个维度展开。该基准通过数据收集与筛选、问答生成和质量控制3个阶段构建，最终形成包含3029个问答对、涵盖5种题型的高质量评估集。对19款主流大型语言模型的实验表明，现有模型在跨文化理解任务上表现不尽如人意，平均得分仅为57.8%。闭源模型表现整体优于开源模型，而跨文化符号对齐仍是最具挑战性的子任务。进一步的错误分析揭示，主要的失败源于文化知识匮乏、意图映射偏差以及对跨文化冲突的高阶推理能力不足，而非简单的指令遵循问题。这些发现凸显了当前大语言模型在文化根基推理方面存在局限性，证明CCU-Bench可为文化感知人工智能研究提供一个标准化基准测试平台。

关键词：基准；大语言模型；跨文化理解；大语言模型评测与基准
https://doi.org/10.1631/ENG.ITEE.2026.0083