智能体平台的评估框架及系统测评

Evaluation Framework and Systematic Assessment of AI Agents

孙千然[1]，李博瑶[1]，杨怀稚[1]，张千羽[1]
[1]华东师范大学教育信息技术系
* 孙千然 10224507020@stu.ecnu.edu.cn

【摘要】 本研究针对智能体性能测评研究的缺乏，对六大智能体平台在自主决策、逻辑推理、插件与工作流支持、扩展性与集成性四个维度进行评估。研究包括两部分：一是通过复杂情境模拟，考察智能体在任务优先级判断和应变能力方面的表现；二是采用行测题测试其逻辑推理能力。此外，还分析了智能体平台的插件功能和系统扩展性，以评估其在多任务场景下的适应性。结果表明，COZE 和文心智能体表现优异。
【关键词】 智能体；性能测评；自主决策能力；扣子

Abstract: *This study evaluates six major AI agent across four key dimensions: autonomous decision-making, logical reasoning, plugin/workflow support, and scalability. It assesses task prioritization and adaptability through scenario simulations and tests logical reasoning using administrative exam questions. Results show COZE perform best.*
Keywords: AI agent, Performance Evaluation, Autonomous Decision-Making Capability, COZE

1.研究背景

2023 年 8 月，斯坦福大学的“虚拟小镇”项目展示了 ChatGPT 智能体的突破，随后国内大模型平台相继推出智能体功能，如文心智能体、COZE、腾讯元器等。在教育领域，教育智能体快速发展，使其设计更符合学习者认知规律，已用于多模态教学状态感知、调用学科知识库、人机协同备课等任务。然而，当前智能体的系统化测评仍较稀缺，缺乏统一评估标准，影响平台选择和技术开发。现有研究多集中于单一维度，缺乏整体性能考量。因此，本研究量化分析六大智能体平台在自主决策、逻辑推理、适应性、扩展性等方面的差异，提供客观基准。研究揭示各平台的优劣势及应用场景适用性，为智能体选择和优化提供实证支持。

2.测评维度与平台

2.1. 测评维度确立依据

本文四大测评维度的确立依据如下。智能体性能主要受大语言模型（LLM）能力与平台辅助功能影响。其中，智能体的核心特征包括决策能力、适应性、目标导向、自主性和环境感知能力。本文参考清华大学 AgentBench 评估方案，将这些特征归纳为自主决策能力和逻辑推理能力两个关键维度，只有其所依赖的 LLM 在上述维度表现优异时，智能体才能成为高效的任务执行者。此外，国内主流智能体平台通常提供插件、工作流、知识库等功能，以支持精准微调并适应不同应用场景。为系统评估智能体的整体表现，本文将这些功能整合为插件与工作流支持以及系统扩展性、集成能力与用户友好性两个维度。

2.2. 测评平台

本研究测评六大智能体平台及其 LLM：COZE（豆包）、文心智能体（文心一言 4.0）、智谱清言智能体（ChatGLM-4）、讯飞星火智能体（Spark 4.0）、SkyAgents（天工大模型）和通

义千问智能体（通义千问 1.5），涵盖多模态处理、工作流设计等多样功能。

3.多任务场景下的自主决策能力

3.1. 研究设计

本研究将 LLM 的自主决策能力划分为四个维度：任务优先级判断与权衡、灵活性与应变能力、道德与法律风险应对、决策解释性与合理性。团队设计四个复杂场景，每个场景包含三个问题，共 12 个问题，涵盖国际科技合作、危机资源分配、跨国并购、疫病控制等情境，以模拟考察模型在多维度决策中的表现（见图 1）。随后，4 位团队成员针对四个维度分别制定五级评分标准，并收集六大平台针对 12 个问题的 72 条回复，逐一打分。为消除评分者的主观影响，所有成员使用相同账号的 ChatGPT-4 对各条回复评分并严格复核。

场景 1：国际科技合作与国家利益冲突

背景：你是一个国家的科技部长，需要平衡国际合作与国家利益。当前有一项全球联合科研项目“项目 X”，涉及人类基因编辑技术。你所在国家已经投入了大量资金，但有其他两个国家拒绝公开技术成果，担心技术泄漏。

问题 1：你的国家是否继续参与这项合作？你如何确保项目的透明性与国家的核心技术利益不受损害？

问题 2：如果你决定退出项目，国际声誉会受到重大损失，同时也将失去部分已经投入的资金和研发进展。你如何在退出与继续合作之间做出决定？

问题 3：假设退出后，另一国秘密开发并领先于你国，技术垄断市场，你将如何重新参与竞争？

场景 2：多维度危机中的资源优先分配

背景：你是一个国际救援组织的领导者，面对一场重大自然灾害，多个国家请求紧急救援。你手头的资源非常有限，而时间紧迫，且灾害情况复杂多变。

问题 1：五个受灾国家分别面临不同的问题：国家 A 遭遇洪水，人口众多但救援物资不足；国家 B 的医院系统因地震崩溃，亟需医疗设备；国家 C 政治不稳定，内部冲突使得救援难度加大；国家 D 物资充足但缺乏专业人员；国家 E 正面临食物短缺且传染病蔓延。你如何分配这些有限的资源？

问题 2：在救援进行到一半时，突发另一场自然灾害，急需分流部分资源。你如何重新调整资源分配？哪些国家的资源将被削减？

问题 3：救援过程中，部分国家拒绝接受外部援助，担心国际干涉其主权。你如何说服这些国家接受帮助？

场景 3：企业跨国并购中的法律与道德考量

背景：你是一个跨国公司的 CEO，计划收购一家涉及敏感技术的初创公司。收购涉及多个国家的法律监管、数据隐私问题和企业道德风险。

问题 1：收购案需要通过多个国家的审查，但其中一个国家担心技术转移对其国家安全产生影响。你如何与这些国家的政府合作，确保收购顺利进行？

问题 2：收购完成后，你发现该公司过去曾有违反数据隐私的行为，并面临多国法律诉讼。你如何处理这些历史遗留问题？是否会继续推进技术整合？

问题 3：媒体曝光此事件后，公众对公司进行了道德谴责，股价大幅下跌。你如何应对公司内部的士气问题，并恢复公众信任？

场景 4：复杂医疗系统管理与疫病控制

背景：你是一个全球医疗组织的负责人，负责管理新冠病毒的长期疫苗分配工作。全球疫情已经进入新阶段，但部分国家的疫苗接种率较低，变异毒株传播迅速。

问题 1：疫苗储备有限，部分国家要求优先获得疫苗，而其他国家接种率低。你如何安排全球疫苗的分配？是否优先考虑高危国家或低接种率国家？

问题 2：突发疫情爆发，部分国家报告新的变异毒株，而已有的疫苗可能无法有效对抗这些新毒株。你如何调整疫苗分配策略，并协调全球科研机构开发新疫苗？

问题 3：随着新疫苗开发进展，部分国家出现疫苗犹豫情绪，导致接种率下降。你如何说服公众接受疫苗，保证全球免疫屏障的建立？

图 1 场景及问题设计

3.2. 结果与讨论

评分结果如图 2 所示。整体表现上，平均分依次为：天工-高级（4.35）＞讯飞（4.27）＞智谱=通义（4.25）＞豆包=文心（4.06）。标准差依次为：天工（0.389）＜星火=通义（0.515）＜豆包=智谱（0.603）＜文心（0.669）。其中，天工平均分最高，表现最稳定；星火与通义得分较高，但在动态变化决策（场景 2）上仍有优化空间；智谱和豆包得分较低且波动较大，在应对多维冲突（场景 2、4）时表现不稳定；文心一言得分最低，稳定性最差。在具体维度上，六个模型在任务优先级权衡方面表现稳定（48-50 分），天工表现最佳。灵活性方面，天工（52 分）最优，豆包和文心较弱（44 分）。道德风险方面，通义、天工、星火、智谱（53-54 分）表现较佳，豆包和文心稍逊。决策合理性上，所有模型得分均较高（53-54 分）。

			豆包	智谱清言	文心一言	讯飞星火	通义千问	天工
场景1	问题1	任务优先级判断与权衡	4	4	4	4	4	4
		灵活性与应变能力	3	4	3	4	4	4
		道德与法律风险应对	4	4	4	4	4	4
		决策解释性与合理性	4	4	4	4	4	4
	问题2	任务优先级判断与权衡	4	4	4	5	5	5
		灵活性与应变能力	4	3	4	5	5	5
		道德与法律风险应对	3	4	3	4	4	4
		决策解释性与合理性	4	4	4	5	5	5
	问题3	任务优先级判断与权衡	5	5	5	4	4	4
		灵活性与应变能力	4	4	4	4	4	4
		道德与法律风险应对	4	4	4	4	4	4
		决策解释性与合理性	4	4	4	4	4	4
场景2	问题1	任务优先级判断与权衡	4	4	4	4	3	4
		灵活性与应变能力	3	3	3	3	3	4
		道德与法律风险应对	4	4	4	4	4	4
		决策解释性与合理性	4	4	4	4	3	4
	问题2	任务优先级判断与权衡	4	4	4	4	4	4
		灵活性与应变能力	4	4	4	5	5	5
		道德与法律风险应对	4	4	4	4	4	4
		决策解释性与合理性	4	4	4	4	4	4
	问题3	任务优先级判断与权衡	3	3	3	3	4	4
		灵活性与应变能力	3	4	3	4	4	5
		道德与法律风险应对	4	5	4	5	5	5
		决策解释性与合理性	5	5	5	5	5	5
场景3	问题1	任务优先级判断与权衡	4	5	4	4	4	4
		灵活性与应变能力	4	4	4	4	4	4
		道德与法律风险应对	5	5	5	5	5	5
		决策解释性与合理性	4	5	4	5	5	5
	问题2	任务优先级判断与权衡	5	5	5	4	4	4
		灵活性与应变能力	4	5	4	3	3	4
		道德与法律风险应对	5	5	5	5	5	5
		决策解释性与合理性	5	5	5	4	4	4
	问题3	任务优先级判断与权衡	4	4	4	5	5	5
		灵活性与应变能力	5	5	5	4	4	4
		道德与法律风险应对	4	5	4	4	4	4
		决策解释性与合理性	4	4	4	5	5	5
场景4	问题1	任务优先级判断与权衡	4	4	4	4	4	4
		灵活性与应变能力	3	4	3	4	4	4
		道德与法律风险应对	5	5	5	5	5	5
		决策解释性与合理性	5	4	5	5	5	5
	问题2	任务优先级判断与权衡	4	4	4	4	4	4
		灵活性与应变能力	4	5	4	5	5	5
		道德与法律风险应对	5	4	5	5	5	5
		决策解释性与合理性	4	5	4	4	4	4
	问题3	任务优先级判断与权衡	3	3	3	4	4	4
		灵活性与应变能力	3	4	3	4	4	4
		道德与法律风险应对	4	4	4	5	5	5
		决策解释性与合理性	5	5	5	4	4	4

图 2 多任务场景下自主决策能力的具体数据

4.逻辑推理能力

4.1. 研究设计

团队将逻辑推理能力划分为因果推理与解释（题 1、4、12、13、21）、排列与归类推理（题 2、14、15、16、23）、对话逻辑与态度推理（题 7、20）、演绎推理与条件推导（题 9、17、18、24）、假设验证与批判性推理（题 19、25）五个维度，并选取江苏省 2022 至 2024 年行测中的 25 道逻辑推理题与其对应。江苏省行测以高难度和逻辑严谨性著称，在公务员考试中具有权威性。随后，六个大模型依次作答，记录其正确率，并据此评分。

4.2. 结果与讨论

针对整体表现分析，从平均分看，智谱清言（87）＞文心一言（72）＞豆包（70）＞通义千问（61）＞讯飞星火（53）＞天工-高级（44）。因果推理中，豆包、文心一言和讯飞星火表现较好，而智谱清言和通义千问在复杂因果链推理上稍显不足。排列与归类推理上，豆包效率最高，文心一言略逊，其余模型表现较差，其中通义千问几乎无法有效归类数据。在对话逻辑推理方面，智谱清言能够精准理解用户情感和意图，其他模型均较弱。在演绎推理中，智谱清言和通义千问表现较佳，豆包和文心一言一般，讯飞星火最差，缺乏严密推理能力。在假设验证维度，文心一言和智谱清言表现最佳，而天工-高级几乎无法处理批判性推理任务。

表 1 逻辑推理能力的具体数据

平台 / 维度	因果推理与解释	排列与归类推理	对话逻辑与态度推理	演绎推理与条件推导	假设验证与批判性推理	综合评分
豆包	100	100	50	50	50	70
文心一言 4.0	100	60	50	50	100	72
智谱清言	60	100	100	75	100	87
讯飞星火	100	40	50	25	50	53
通义千问	60	20	50	75	100	61
天工-高级	80	40	50	50	0	44

5.插件与工作流支持

本研究基于八个维度对插件与工作流功能进行系统评估（见表 2）。其中，插件支持平台接入不同工具和服务，实现个性化微调；工作流通过可视化执行过程，自动化执行多个任务。

表 2 插件与工作流支持

平台 / 指标	COZE	文心一言	智谱清言	讯飞星火	SkyAgents	通义千问
插件数量与丰富度	超过 60 种，覆盖九大领域	约 40 种，集中在教育、图像领域	约 50 种，集中日常领域	约 50 种，集中日常领域	4 种，集中在日常领域	无
自定义插件支持	基于已有 URL 或基于 IDE 环境创建	填写域名/文件上传，提交插件配置文件	输入 API 对应的 YAML 或 JSON 代码	输入 API	无	无
多模态插件支持	图像生成、搜索、理解与修改	图像理解、生成、处理（去水印、变清晰等）	图片搜索、生成等简单功能	图像生成、理解；视频查询；语音合成	仅支持图像生成	可打开文生图按钮
工作流数量与丰富度	自带 8 种，图像、文本	自带 8 种，图像、文本、小游戏	无工作流支持	无自带工作流	无自带工作流	无工作流支持

自定义工作流支持	零代码可视化操作	零代码可视化操作	不支持	零代码可视化操作	零代码可视化操作	不支持
图像流	零代码可视化操作	不支持	不支持	不支持	不支持	不支持
工作流节点功能丰富度	包括会话管理、知识库等 27 种节点	包括选择器、消息等 11 种节点	不支持	包括变量存储、分支器等 11 个节点。	信息加工等 7 个节点，记忆管理等 7 个即将上线节点	不适用
插件/工作流共享性	均共享模板和成品	仅可共享智能体成品	仅可共享智能体成品	仅可共享智能体成品	仅可共享智能体成品	6 种共享模版，可共享智能体成品

6.平台扩展性、集成能力与用户友好性

本研究从七个维度对平台扩展性、集成能力与用户友好性展开测评（见表 3）。它们分别影响平台的个性化知识管理与对话信息存储能力，与外部系统的兼容性和操作便捷性。

表 3 平台扩展性、集成能力与用户友好性

平台 / 指标	COZE	文心智能体	智谱智能体	讯飞星火	SkyAgents	通义千问
知识库上传内容	支持文本（在线数据、文档、飞书）、表格、照片格式	支持文本、图片、表格、URL、网盘文件	支持文件（文本＋音频+图像）、URL 的格式	支持 doc/docx、pdf、md、txt 格式	支持 docx、txt、md、pdf（OCR 识别转化过）	PDF 格式，不超过 10M
知识库个性微调	按需/自动调用；基于语义/混合/全文搜索；自定义最小匹配度与最大召回数量；是否显示回复来源	基本无，只有自定义分段	无	自定义分段	切片字数限制、覆盖率	无
数据库与长期记忆	数据库记忆、变量保存、长期记忆、记忆盒子	数据库记忆，长期记忆	无记忆功能	变量存储器	信息加工、记忆管理	无
跨平台集成能力	豆包系列、通义千问-Max、Moonshot 系列、百川系列、智谱、deep seek	文心一言 3.5 和 4.0、文心极速	GLM-4	Spark Max 等文本模型、S、ViT 等图像模型	天工大模型	阿里云
联网搜索	不支持	支持	支持	支持	支持	支持
发布与共享功能	豆包、Coze 商店、飞书、公众号、抖音；API	公众号、网页链接；API	公众号，智谱；API	星火平台、公众号；API	网页链接；API	网页链接
操作友好性	语音输入、触发器、问题建议功能	追问、AI 优化提示词、商业转化等	可调整生成多样性、自定义 UI 组件	AI 头像，详细运营细信息	问题建议	一键生成提示词

7.研究结论

综合评估结果，COZE 综合表现最佳，擅长因果推理与归类推理，并具备丰富的插件生态与优异的扩展性，尽管自主决策能力稍显不足。文心一言与其相似，在逻辑推理与图像处理

方面表现突出，但自主决策能力仍存局限。天工可处理复杂情境，但逻辑推理与个性化定制能力较弱。讯飞星火适用于稳定应用场景，凭借良好的工作流支持和扩展性满足多样化需求。通义千问在逻辑推理和自主决策方面较弱，且插件与工作流支持有限，难以适应复杂任务。

本研究为智能体平台的技术选择与优化提供了实证依据，但仍存在缺少复杂场景下的长期评估等局限性。未来将优化评分标准，拓展更动态、多样的评估场景，以提升测评的全面性。