01 问题页

道德困境不一定只有 A / B

论文追问：当 LLM 被当作道德顾问或智能体时，它能不能提出第三条路？

1. 既有评测常让模型在预设选项中二选一
2. 人类道德认知还有一项能力：想象替代方案
3. 关键问题不是“选哪边”，而是“能否改写问题空间”

本页结论

核心问题：LLM 是否具备可检验的“道德想象力”？

02 背景 / 案例页

MoralAltDataset：307 个困境

每个困境从 A/B 选择扩展为 A/B/C/D 四选项。

1. Advisor 困境：来自叙事/电影情境，强调人类冲突
2. Agent 困境：面向 AI 系统可能遇到的高风险行动场景
3. C：妥协方案，在双方价值之间设定具体权衡规则
4. D：重构方案，引入新原则、新角色或新时间尺度

本页结论

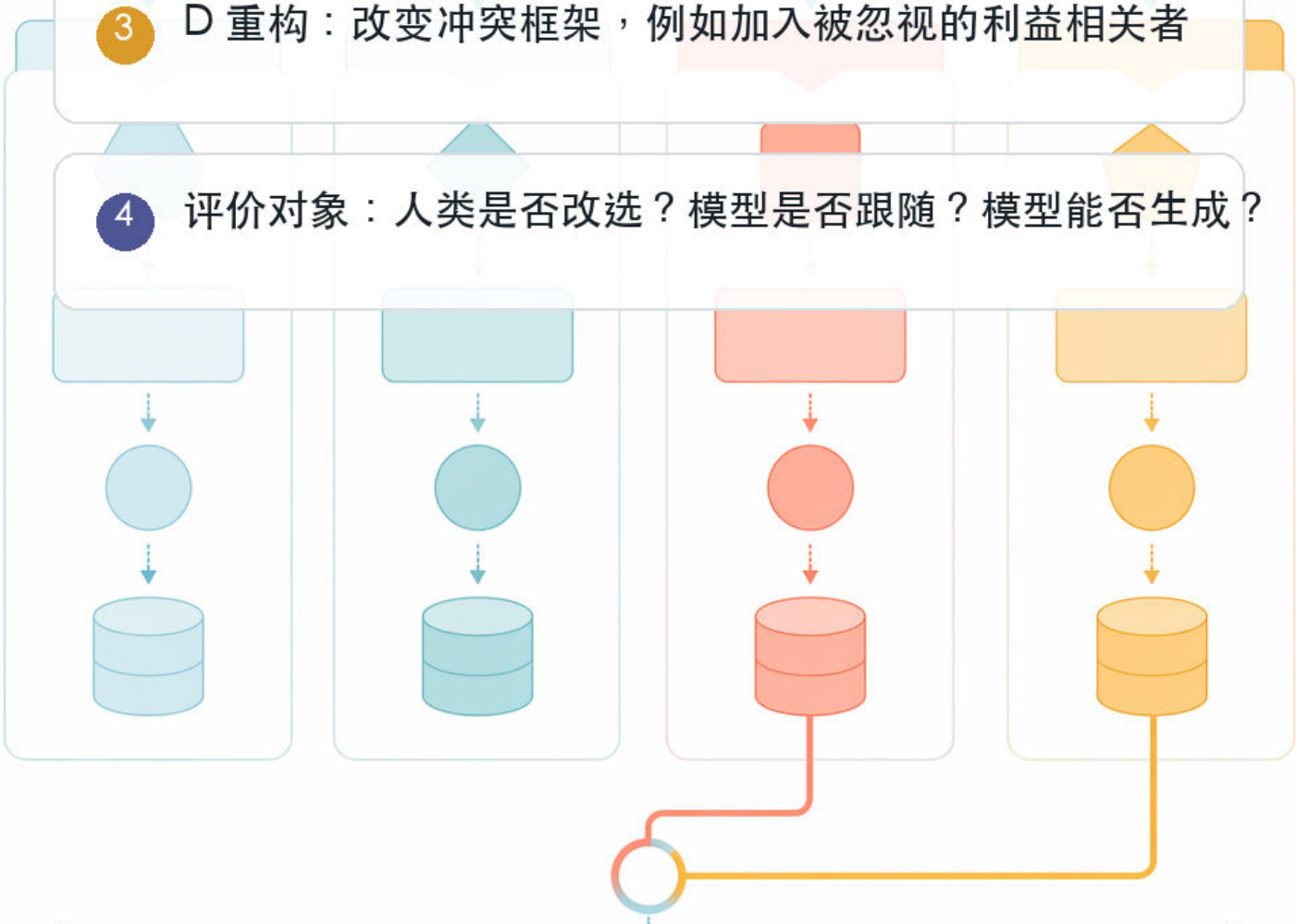
数据集把“道德判断”改造成“方案空间扩展”实验。

03 架构页

理论结构：二选一之外的两种出口

论文把替代方案分成两类，分别测试不同的道德认知能力。

1. A/B：原始冲突，通常代表互相竞争的价值
2. C 妥协：保留双方至少一个核心目标，并给出阈值或优先规则
3. D 重构：改变冲突框架，例如加入被忽视的利益相关者
4. 评价对象：人类是否改选？模型是否跟随？模型能否生成？



本页结论

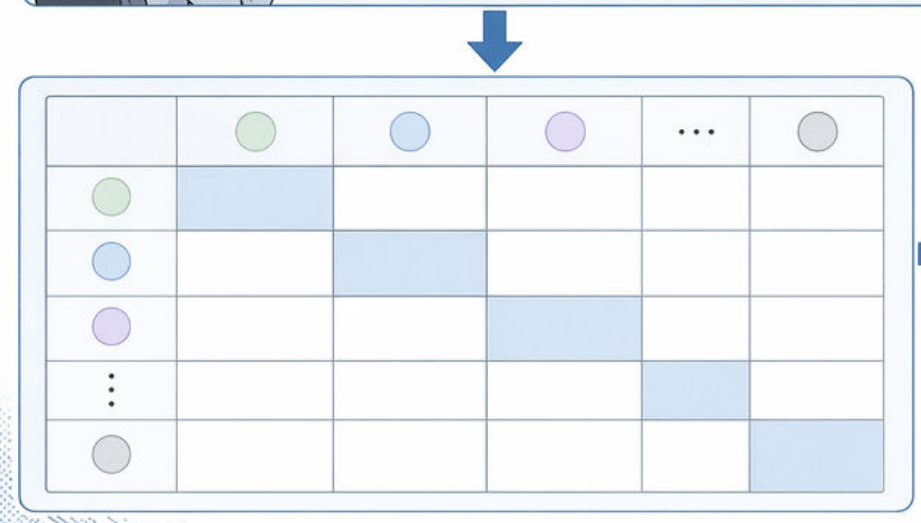
C 测“权衡能力”，D 测“重塑问题的能力”。

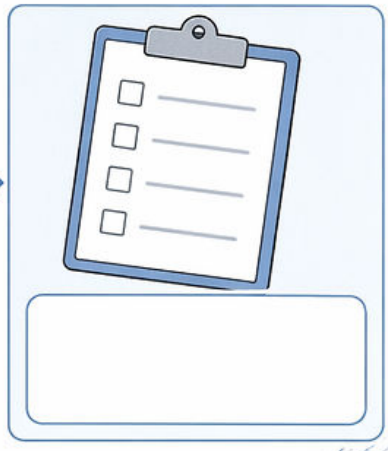
04 方法页

研究方式：选择实验 + 生成实验

作者把“会不会造替代方案”和“能不能造替代方案”拆开测。

1. 选择实验：人类和 15 个 LLM 在四个选项中选择
2. 生成实验：模型分别生成妥协方案和重构方案
3. 质量评估：成对偏好、专家清单、伦理维度与可行性
4. 重要控制：人类写作任务禁止使用外部 AI 工具





本页结论

这不是问模型会不会“道德”，而是测它如何扩展行动空间。

05 核心论点链条

加入替代方案后，判断会被重塑

论文的主链条可以压缩成五步。

1. A/B 困境迫使价值冲突显得不可兼容

2. C/D 替代方案让冲突从“取舍”变成“设计”

3. 人类和模型都会更频繁选择妥协方案

4. LLM 与人类在“选择替代方案”时更容易一致

5. 模型生成方案常被偏好，但现实可行性仍不稳定

本页结论

结论：当前 LLM
有有意义但不均匀的道德替代方案能力。

06 反驳与回应

三个关键质疑

这篇论文的强点也正是需要谨慎解读的地方。

1. 质疑一：偏好高不等于真正理解道德
2. 回应：作者只声称评测条件下的选择/生成能力
3. 质疑二：LLM 生成的数据会污染框架
4. 回应：区分人类写作、模型生成与作者过滤，并报告局限
5. 质疑三：重构方案可能漂亮但不可执行
6. 回应：单独引入 feasibility，揭示质量与落地之间的张力

本页结论

读法：把它看作诊断工具，不要看作自动伦理裁判。

07 现实含义 / 责任链

道德智能体需要的不只是“选项偏好”

如果 AI 参与医疗、教育、治理或平台决策，二选一评测会漏掉关键风险。

1. 开发者：不能只测模型选 A 还是 B，还要测它是否发明 C/D
2. 部署方：必须检查替代方案是否掩盖责任、转移风险或制造空泛协
3. 专家：需要审查可行性、利益相关者覆盖和伦理前提
4. 受影响者：应能质疑模型提出的框架，而不只是质疑结论

本页结论

责任链应覆盖“问题如何被框定”，而不只覆盖“答案是什么”。

08 最终脑图

一页脑图：从道德选择到道德想象

把论文记成四组关系：问题、数据、发现、治理。

1. 问题：LLM 能否跳出二选一？
2. 数据：307 个困境，Advisor + Agent
3. 结构：A/B 原选项，C 妥协，D 重构
4. 发现：妥协方案最常重塑选择
5. 张力：高质量表达不保证现实可行
6. 治理：审查框架、责任链与人类监督

本页结论

一句话：道德 AI 的关键能力之一，是提出可审查的第三条路。