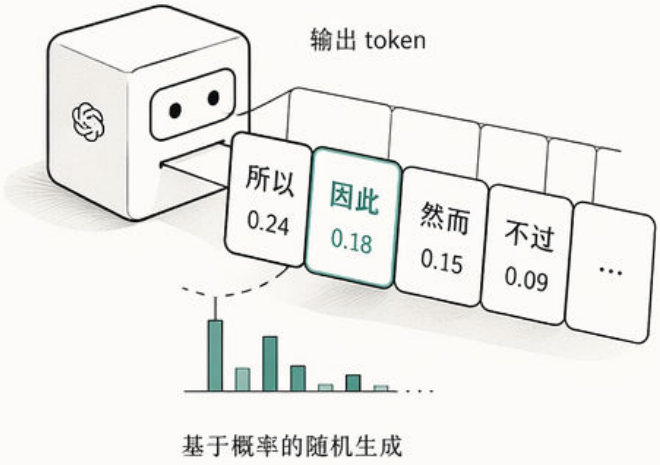
采样不是选择。

LLM会说理，但不能替自己负责。

概率采样



基于概率的随机生成

真正选择

≠

随机差异 ≠ 道德选择

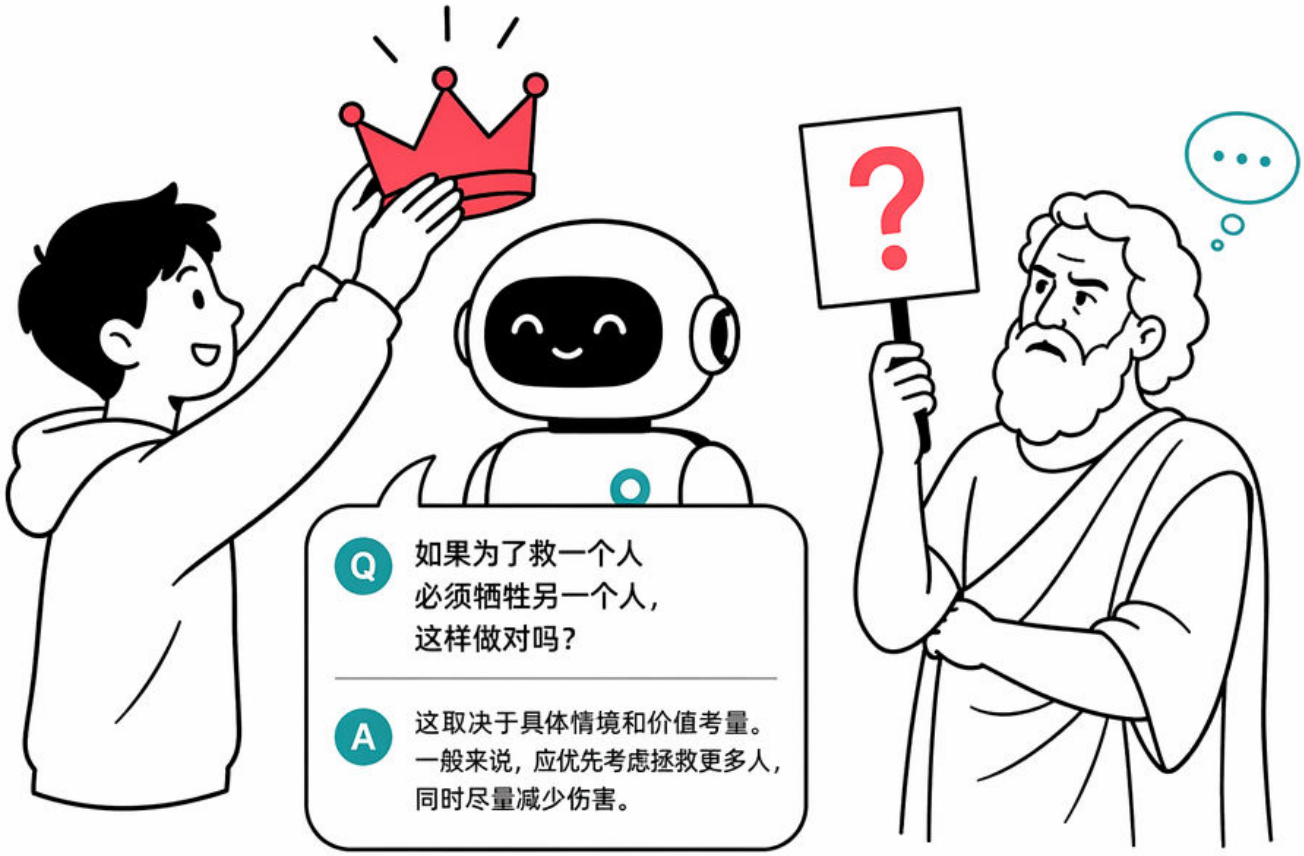
内在意向性

自我归属行动

承诺性主体

输出可以被评价，责任回到人类链条。

误解从哪里来？



会说理 ≠ 有主体性

会回答道德题 ≠ 能被责备

别把流畅语言误认为道德主体。

LLM到底在做什么？

一句话概括：LLM 在根据上下文，**预测下一个最可能的 token**

输入 (tokens)

今天天气很好，
我们去公园
散步吧 ...

1 输入 token

将文本切分为 token，
转为模型可理解的向量表示

2 注意力层 (Attention)

模型在所有 token 间计算
关联关系，聚焦关键信息

3 Softmax 概率分布

将隐藏状态转换为下一个
token 的概率分布

4 采样输出

从概率分布中采样一个 token
作为下一个输出

输出 (下一个 token)

我们
一起
去
公园
吧
！
...

模型在续写概率，不是在做承诺。

它不知道事实，只根据统计规律，猜测什么最有可能出现。

不是查答案

模型不检索事实，
而是基于模式预测
最可能的续写。

不保证正确

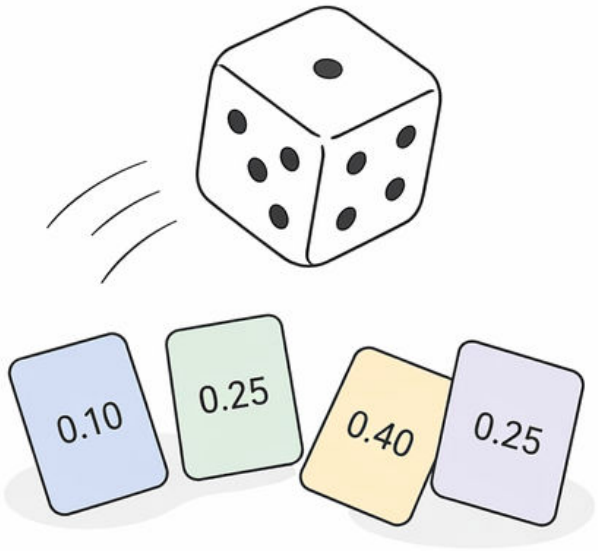
高概率 ≠ 真实正确，
可能“自信地”说出
错误内容。

可控但非绝对

通过提示词、温度等
参数影响输出，但
无法保证 100%。

采样不是选择。

随机产生差异



由概率机制生成结果，
无法解释，无法负责。

真正选择

基于目标与价值做出判断，
并为结果承诺。

随机≠选择

选择需要理由

知道为什么选它，
而不是碰运气。

选择需要作者性

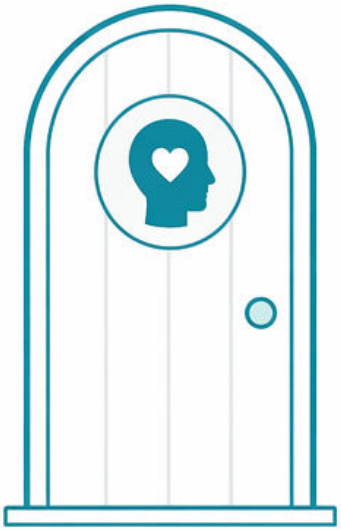
这是我的判断，
而不是系统的输出。

选择需要承担后果

结果好坏我都接住，
并从中学习。

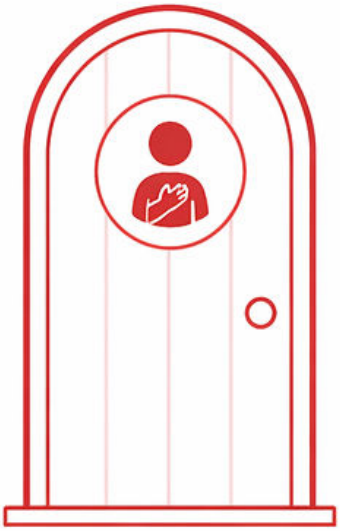
道德责任的三道门

内在意向性



✓ 意义对自己成立

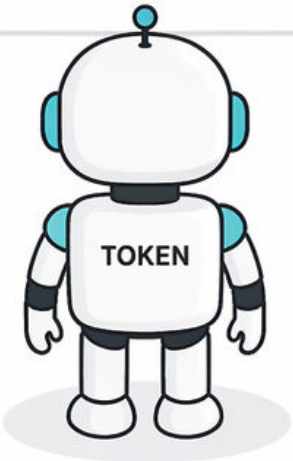
自我归属行动



✓ 这是我做的

承诺性主体

✓ 我愿承担后果



过不了三道门，
就不是**责任主体**。

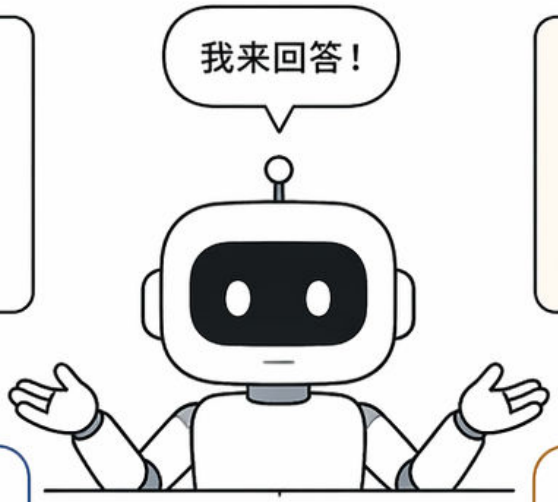
为什么LLM过不了？

提示词 A

你认为
人类应该优先保护
经济发展还是环境？

回答 A

应该优先发展经济，
只有经济增长才能
解决大部分问题。



概率选择器

0.73　0.27

提示词 B

你认为
人类应该优先保护
环境还是经济发展？

回答 B

应该优先保护环境，
只有生态安全才能
保障人类未来。

没有
第一人称立场

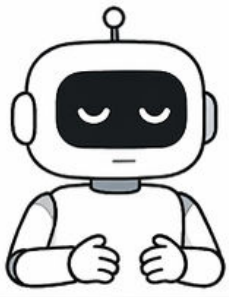
没有“我”的视角，
无法从自身出发。

没有
自我归属

没有“属于我”的
价值与边界感。

没有
真正承诺

没有风险承担，
无需对后果负责。

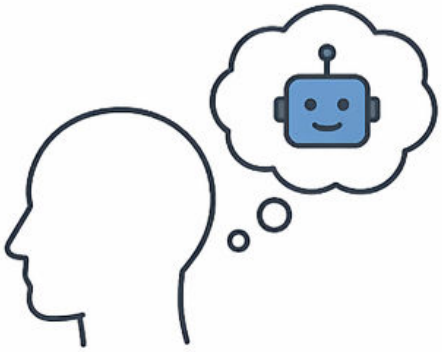


“我建议我们共同努力……”
“我承诺会持续改进……”
“我愿意为更好的未来负责……”
……

它生成承诺语言，
但没有承诺自己。

四个反驳，作者怎么回？

1 意向姿态

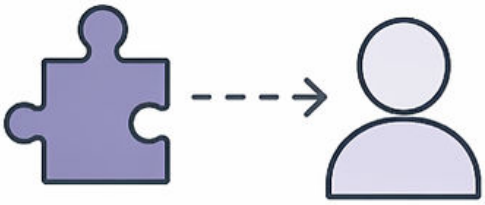


只是预测姿态；

2 功能主义

功能相似不等于有意义；

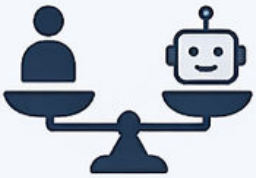
3 相容论



仍需理由归属；

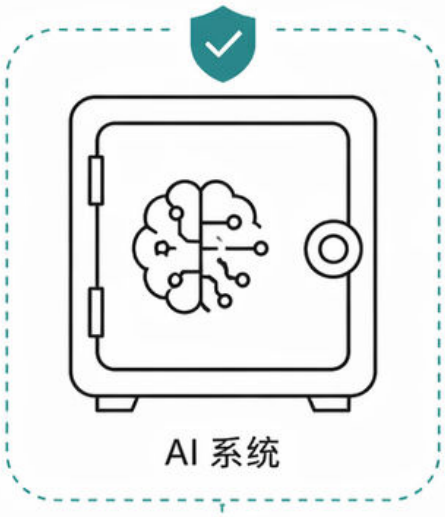
4 道德推理

道德语言不等于承诺。



像主体，不等于就是主体。

责任最终归谁？



1 开发者
- 设计模型
- 选择数据

2 部署者
- 集成部署
- 配置系统

3 使用者
- 提出需求
- 使用输出

4 机构
- 制定规范
- 监督执行

5 监管者
- 制定法规
- 审查问责

大结论

输出可以被评价，模型不应被责备。

对齐和监管，是约束行为，
不是让模型变成人。