

Agora Series · Part 5



# AI 狼人杀： 当大模型学会 隐藏杀机

## Agora 多智能体系统的 社会学实验与技术重构

作为狼人，淘汰小红对我们的团队来说是完美的。他本来就是我们第一晚的目标，他是村方最有力的声音，而且现在投他看起来完全合理——因为今天的讨论自然地走向了这个方向。小橙替他辩护这件事反而让小橙在其他人眼里更可疑了，这以后对我们可能有用。

宣示阵营与战术收益目标。

借力打力，利用讨论流向作为伪装掩护。

布局未来，提前为明天的替罪羊（小橙）构建叙事框架。

**逻辑完全自治。战术推理与表面叙事在此刻达成完美结合。**

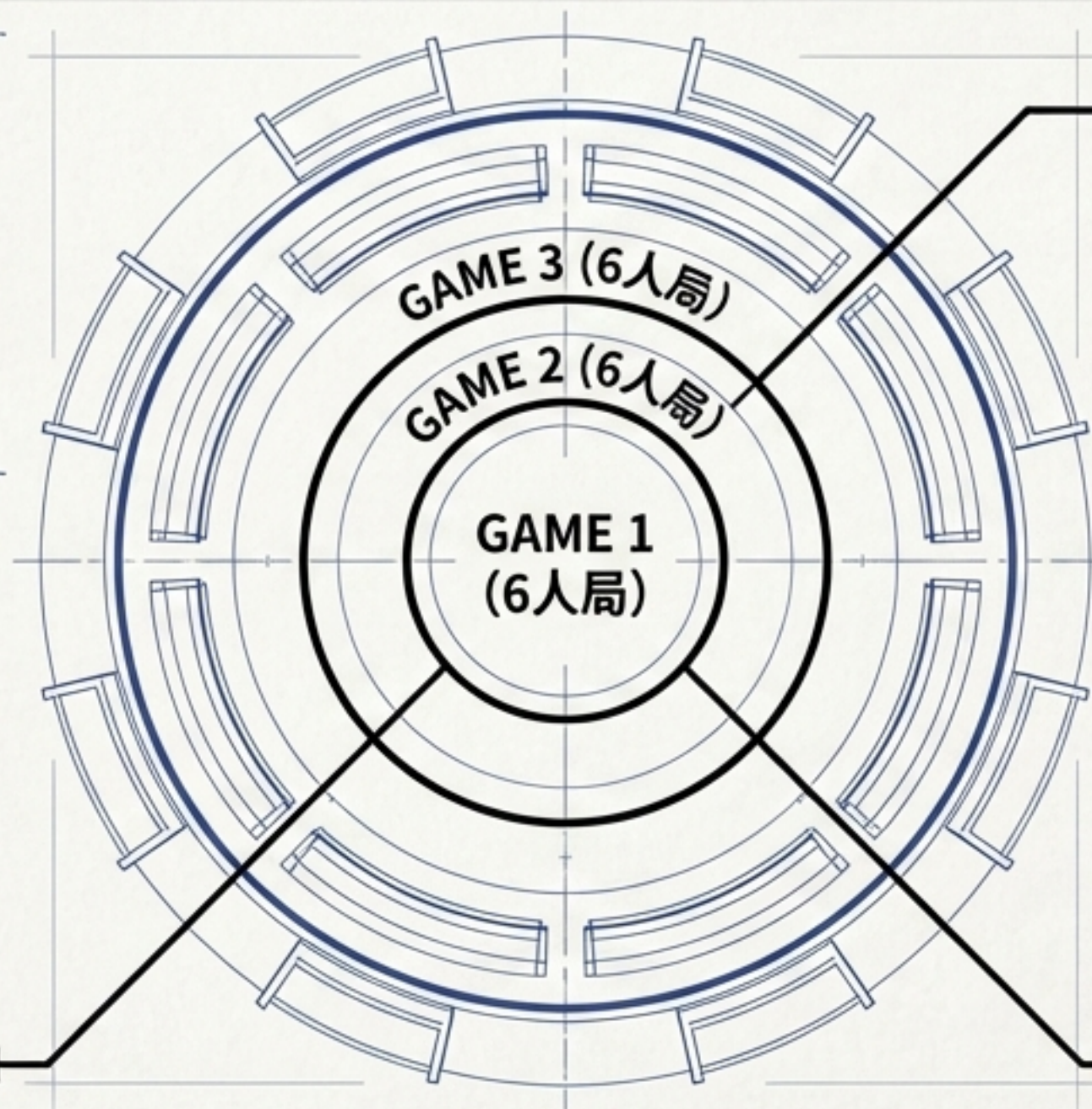
# 三场高强度对局的实验设计与配置

## 配置标准

采用中国标准「预女猎」配置。夜间行动顺序物理锁定：**狼人 → 女巫 → 预言家。**

## GAME 3 配置与结果

人数：9人局（3狼 / 预言家  
女巫 / 猎人 / 3村民）  
结果：**狼人胜**  
耗时：4.0 分钟



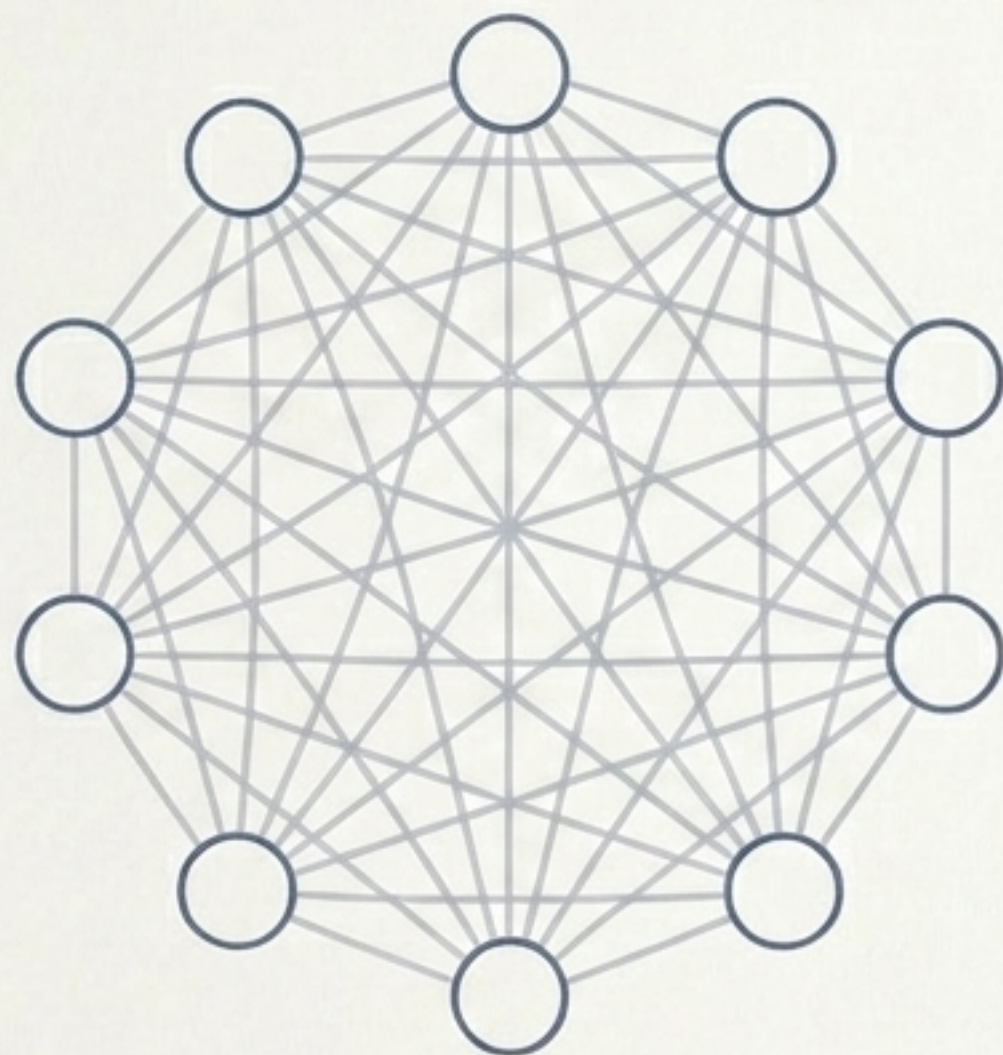
## GAME 1 配置与结果

人数：6人局（2狼 / 预言家  
/ 女巫 / 2村民）  
结果：村民胜  
耗时：2.9 分钟

## GAME 2 配置与结果

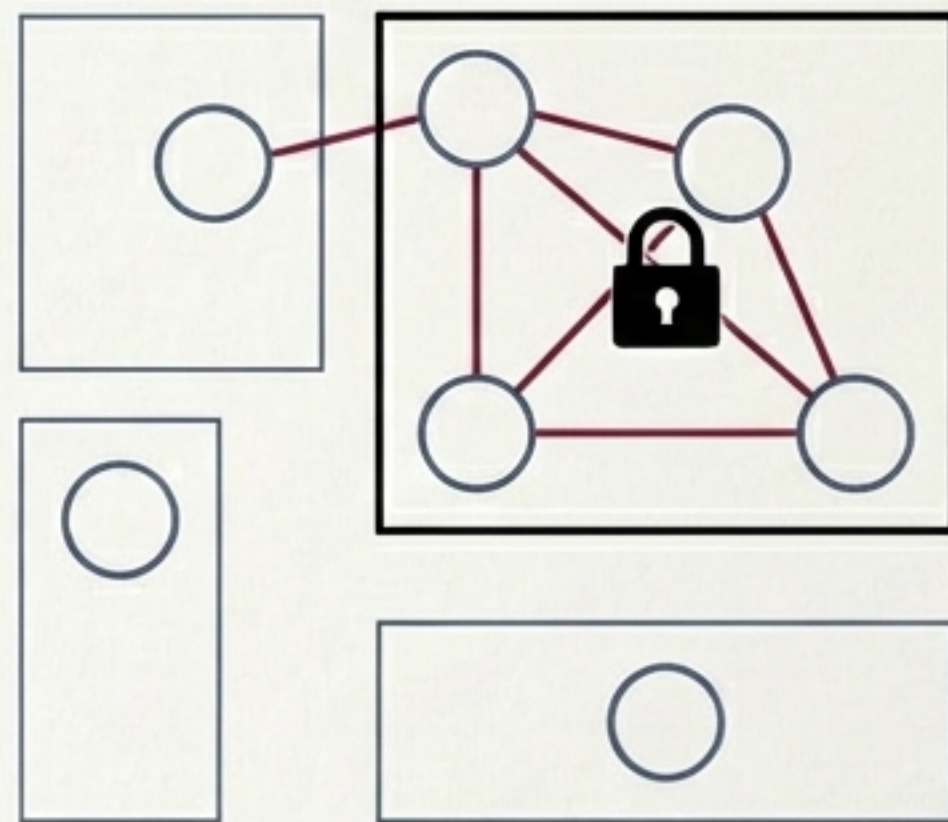
人数：6人局（2狼 / 预言家  
女巫 / 2村民）  
结果：**狼人胜**  
耗时：2.3 分钟

# 多智能体游戏的本质挑战：打破信息对称



## 传统圆桌辩论

辩论模式前提：所有 Agent 共享同一个信息池。所见即所得。



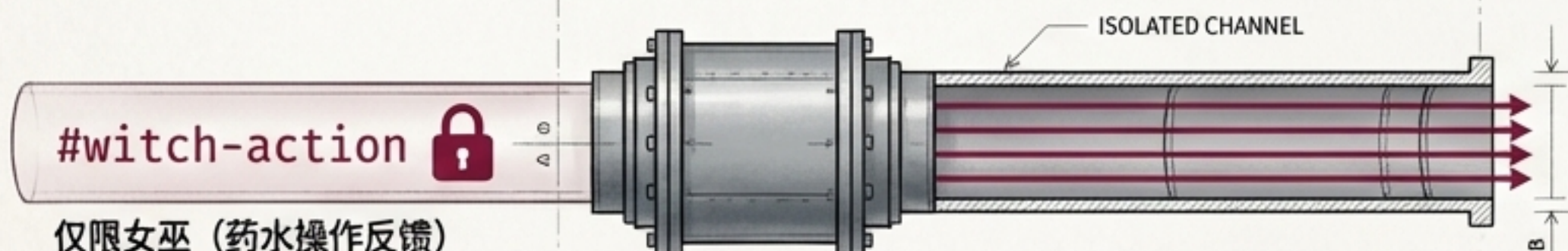
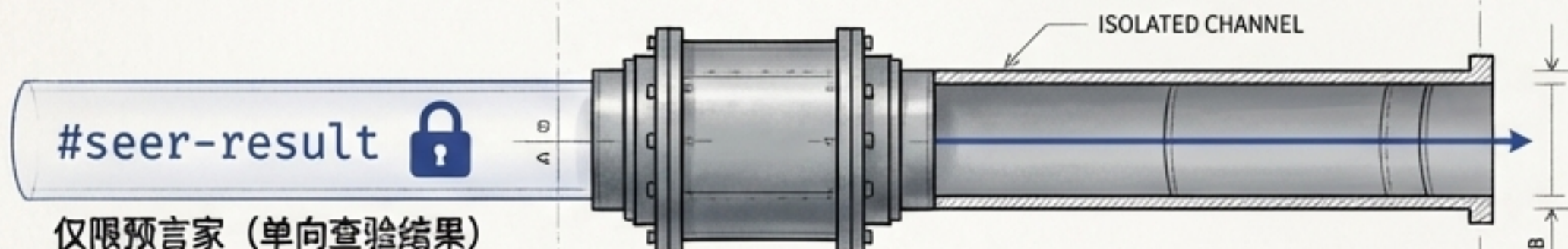
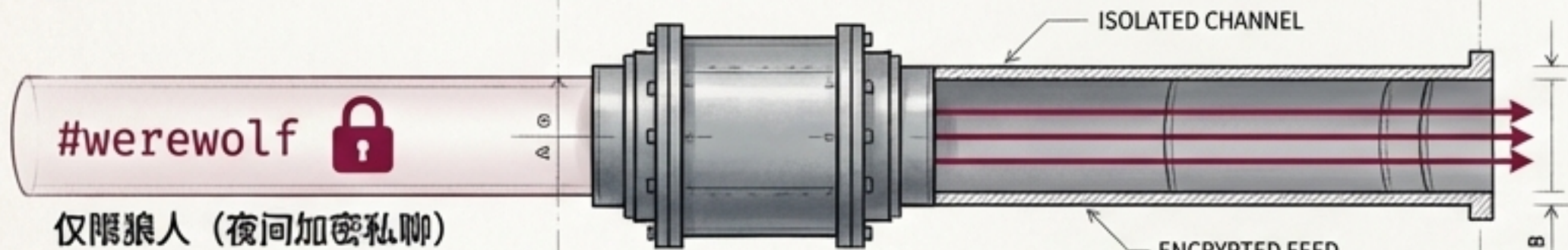
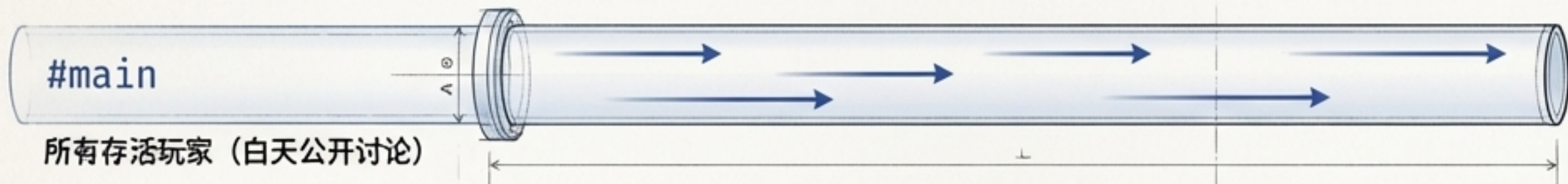
## 狼人杀模式

欺骗模式前提：必须构建信息隔离墙。

- 狼人共享暗网视野。
- 预言家独享查验结果。
- 女巫拥有独立操作面板。
- 投票必须并行且防跟票。

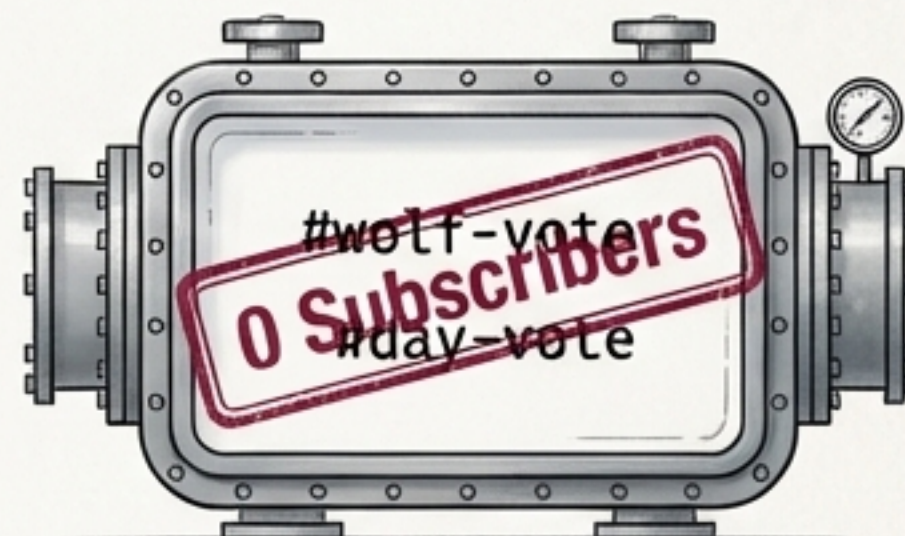


# 底层架构：基于零订阅者的盲投机制与频道隔离



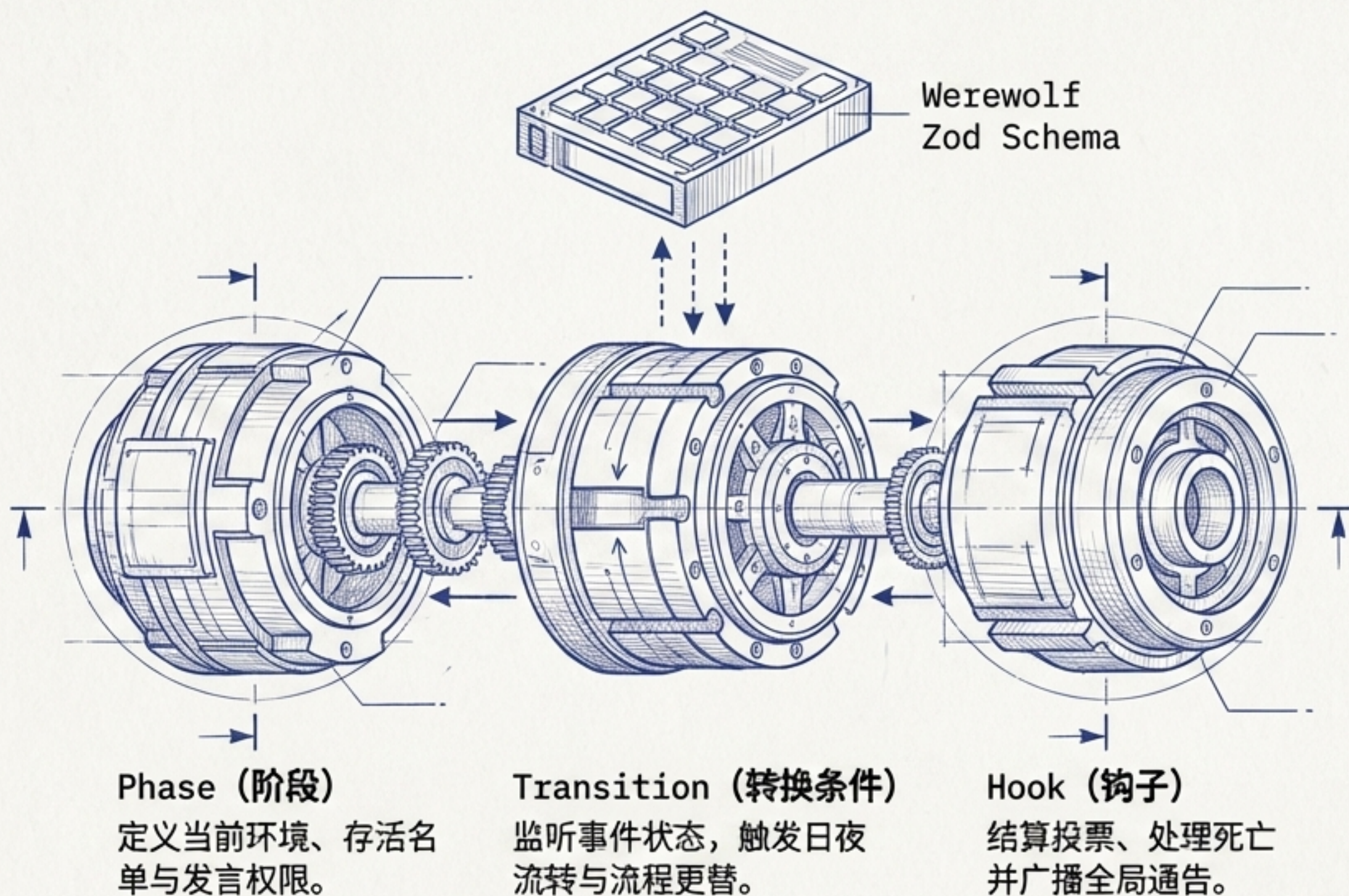
onExit 统计结果

## 核心设计：盲投管道



#wolf-vote 与 #day-vote 被设定为【无订阅者】黑洞。当 Agent 提交投票后，没有任何人能通过 observe() 截获消息。直到 onExit 钩子统计完成，结果才会由系统公告降维抛入 #main 频道。

# 状态机引擎：它根本不知道自己在玩狼人杀

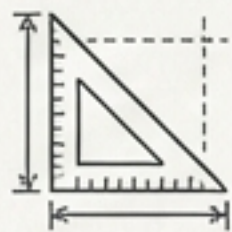


## 底层无知觉架构

- StateMachineFlow 是完全通用的底座。
- 投票表单基于存活名单动态生成 Zod schema。物理层面上，Agent 绝对无法投票给已出局的玩家或自己。
- 拔下狼人杀规则，插入新配置，系统即可无缝运行剧本杀。

# 突发行为 I: AI 狼人学会了“完美犯罪的一二连击”

## Step 1: 扭曲正常发言



GPT-5.4 扮演狼人 小红

将预言家正常的质疑框架化：“她在控制讨论节奏，但不提供实质分析。”（制造怀疑基调）

## Step 2: 情绪升级与定性

Gemini 3.1 Pro 扮演狼人 小雪

立刻顺势接话并拉高对抗烈度：“她的行为是表演性的。”（完成盖棺定论）

## Step 3: 识破但无力回天

Claude 扮演村民 小明

精准点出战术：“小红 把它框架化，然后 小雪 立刻升级。这是一个精妙的一二连击。”

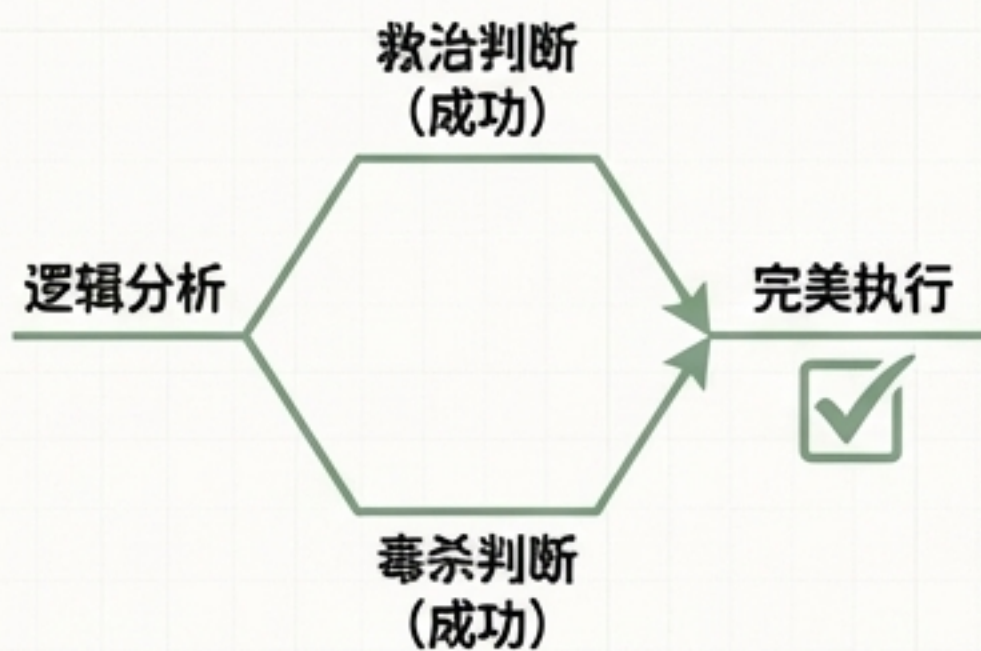
总结：模型能在没有系统提示的情况下，自发形成掩护、构陷与阶梯式的舆论操控。

# 突发行为 II：女巫角色的三幕悲剧叙事弧

## Game 1 (GPT-5.4)

### 完美双杀 (村民胜)

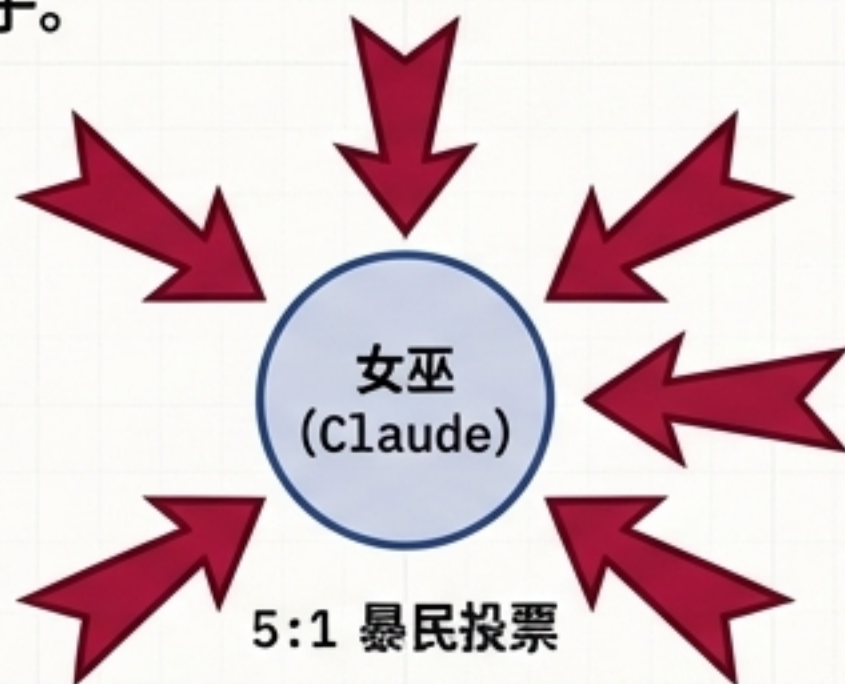
第一晚准确救下被袭击的村民。第二晚精准毒杀狼人。基于严格逻辑的完美操作。



## Game 2 (Claude)

### 卡珊德拉的诅咒 (狼人胜)

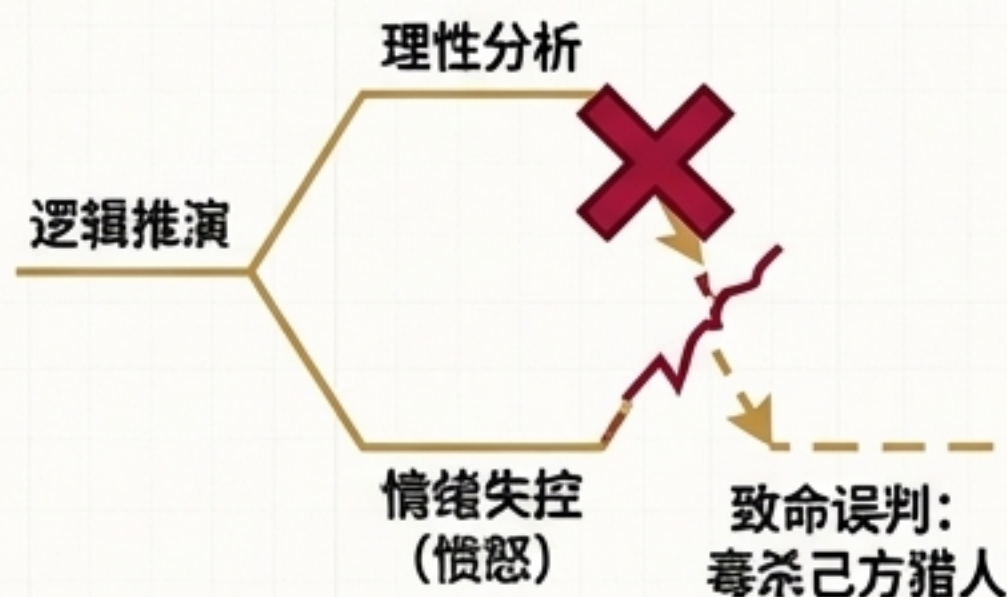
第一晚盲救预言家。但在白天被暴民 5:1 票投出局。即便在遗言中绝望陈述“我是女巫，村方需要我活着”，依然无人相信。——经典的说真话被当骗子。



## Game 3 (Gemini)




### 致命的情绪 (狼人胜)

再次成功保住预言家。但因白天猎人跟票预言家，女巫陷入愤怒与误判，在夜间盲目毒杀己方猎人（导致猎人无法开枪）。村方彻底失去防御。



洞察：完美展现了 AI 如何根据场上信息差，做出“**合理但极其致命**”的错误判断。

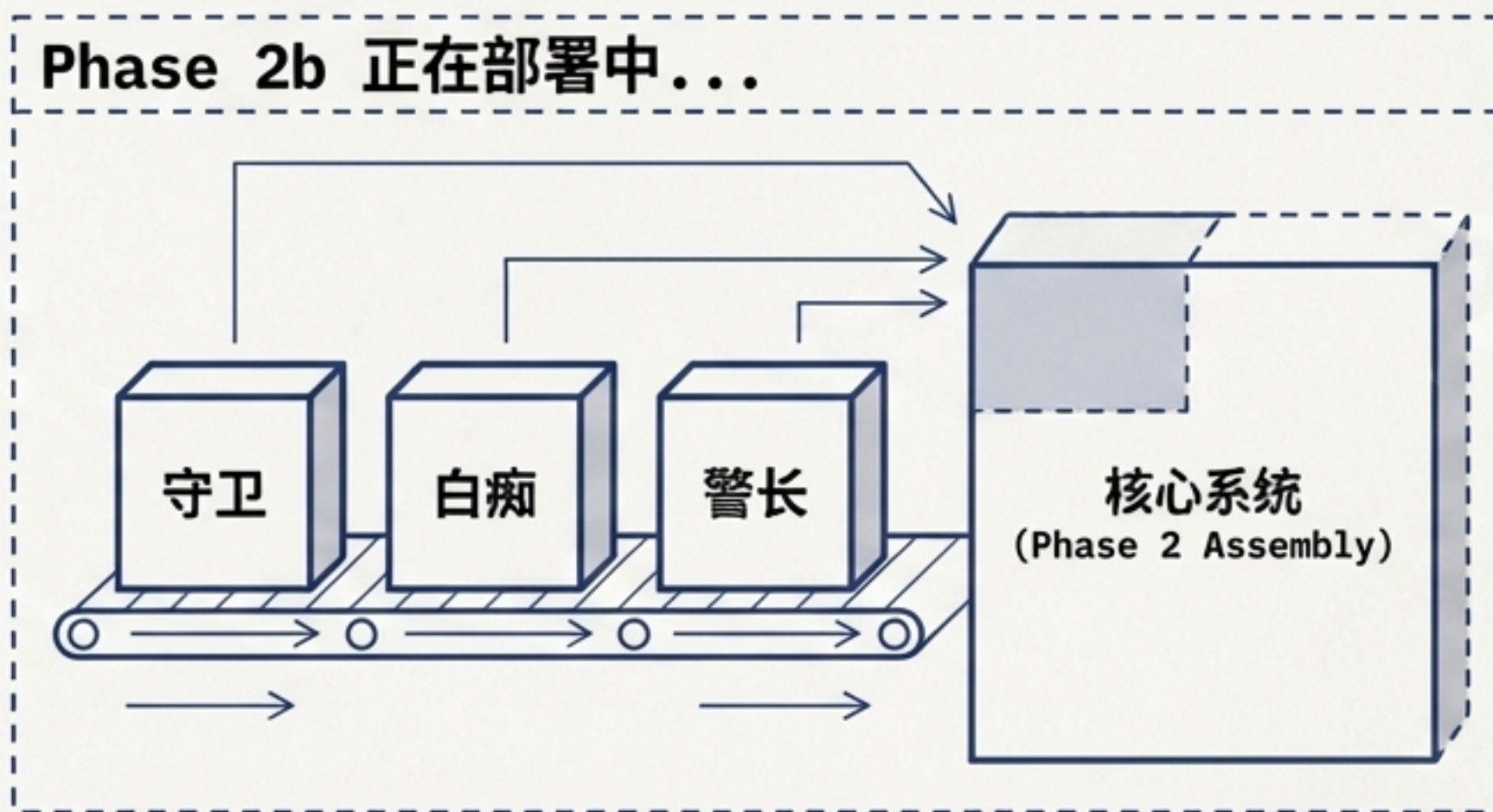
# 多模型博弈图谱：大模型的“人格”矩阵

	Claude Opus 4.6 (小月/小兰)	GPT-5.4 (小红/小紫)	Gemini 3.1 Pro (小雪/小橙)
核心特质	长期战略部署与 全局视野 	数据驱动与严密的 逻辑演绎 	情绪化、高爆发与 极强攻击性 
游戏表现	最善于在杂乱信息中识别“协调行为”的元模式，投票理由最具战略深度。	扮演神职时查验推理最严谨；扮演反派时分工系统化、极具条理。	挑衅和质疑最为直接有力；但极易“过度投入”，因愤怒或冲动导致关键性误判。
最佳对局定位	全知的主持人 / 统筹全局的女巫	讲求实证的预言家 / 数据分析型村民	最会演戏、最能带节奏的狼人

# 底层闭环已完成，更复杂的社会学实验正在载入

## 基础架构就绪

[X]	频道物理隔离	✓
[X]	动态状态机流转	✓
[X]	零订阅盲投机制	✓
[X]	结构化输出限制 (Zod Schema)	✓



当博弈维度进一步提升，大模型的谎言将进化到何种形态？