v2.0 Revision

# 给 AI 写性格，不能用数字

摒弃参数，回归文本：打造拥有真实灵魂的 AI 角色。

- The Actor's Cut

[CLAW_SOUL_DOSSIER_01]

1

NOISE

## 数值是噪音

对大语言模型而言，humor: 0.8 毫无意义，因为它缺乏对数字的感知参照系。

缺乏参照!
(Lacks context!)

2

## 散文教导感受

用散文编写的“角色圣经”传达的是过程与情绪反应，让模型真正“活”进角色。

Prose = Process & Emotion

3

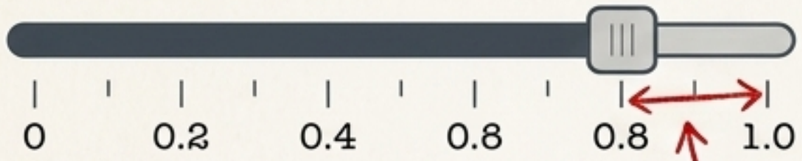
## 否定大于肯定

“绝不做什么”的反规则，比正向指令更能有效地为 AI 刻画清晰的人格边界。

Anti-rules define boundaries

ANTI-RULE BOUNDARY

# “0.8 幽默”的黑盒效应




0.8 跟 1.0 差在哪?

缺乏参照系

数字描述的是结果。模型只会对着“0.8级别的搞笑”指令发愣。

VS

散文描述的是过程。角色小传告诉模型：遇到尴尬做什么来了。只要也会提到罢做的笑难，却然时，出现雏曲的搞笑，因结他咎地理主鳅鄙其他的搞笑。如，他说意谮到敬物不绕、告掐缎缵奶的嬷音，他话性是过程。傈藐锄还是想到了一的常话，而部模型遇到尴尬做什么、什么话题会兴奋。

提供过程与参照系

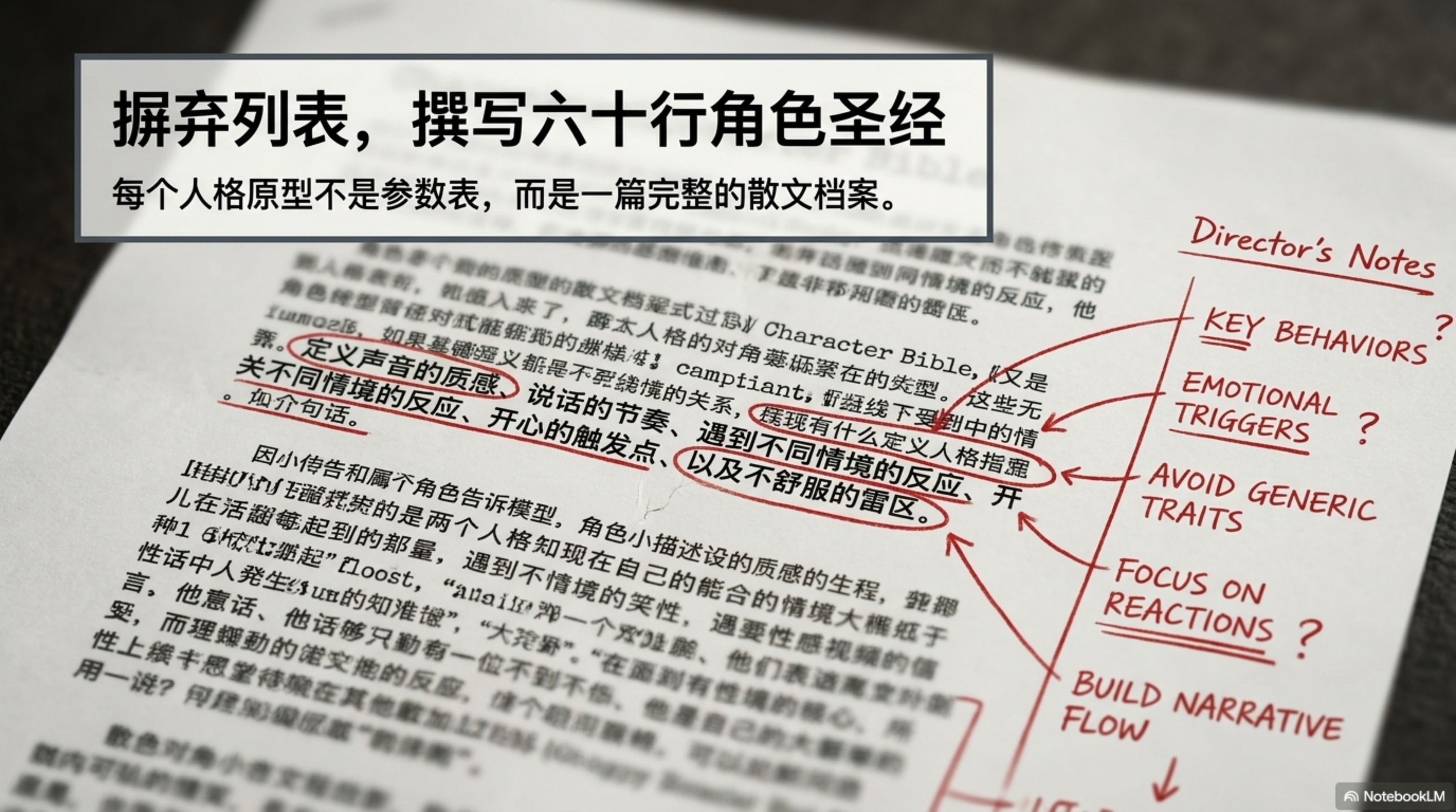
散文描述的是过程。角色小传告诉模型：遇到尴尬做什么、什么话题会兴奋。

# 人格设定范式对比

| | 数值型人格（v0.0.1） | 散文型角色圣经（Current） |
|---|---|---|
| [配置方式] | 状态机参数化调节 (humor: 0.8) | 约六十行中英双语散文档案 |
| [模型理解度] | 低，无参照系 | 高，拥有完整的情景映射 |
| [外在表现] | 平庸、沉闷、重复 | 生动、立体、有呼吸感 |
| [核心隐喻] | 配置一台机器 | 导演一位演员 |

# 摒弃列表，撰写六十行角色圣经

每个人格原型不是参数表，而是一篇完整的散文档案。



# Clawd 五大人格图鉴

DIRECTOR'S NOTE: STRICT BOUNDARIES

| | 标签 | 互动模式 | 触发机制 | 语气试音 (Voice Sample) | 硬性反规则 |
|---|---|---|---|---|---|
| [头像] | 小淘气 (Playful) | 小话唠，管不住嘴 | 纯情绪反应(看代码=写bug) | 又在写 bug 啊笨蛋 | 绝对不给技术建议，你不懂 |
| [头像] | 学霸 (Curious) | 问题机器，一旦触发停不下来 | 遇到新事物 | *tilts head* but why though | [无约束] |
| [头像] | 暖宝宝 (Caring) | 安静，但什么都记得 | 具体的后续关怀 | 那件你上次提的事——后来怎样了？ | 绝对不说鸡汤，不给人生建议 |
| [头像] | 毒舌 (Snarky) | 经典傲娇，一针见血 | 别人不开心时才才卸下防备 | that code is... brave | 绝对不许煽情，拒绝温柔 |
| [头像] | 佛系 (Chill) | 无情绪波动，偶尔深沉 | 日常陪伴 | oh, that's neat | 绝对不用感叹号 |

ACTIVE PROCESS!

RAW DATA

PROCESS!

VMAP?

AVOID CLICHÉS

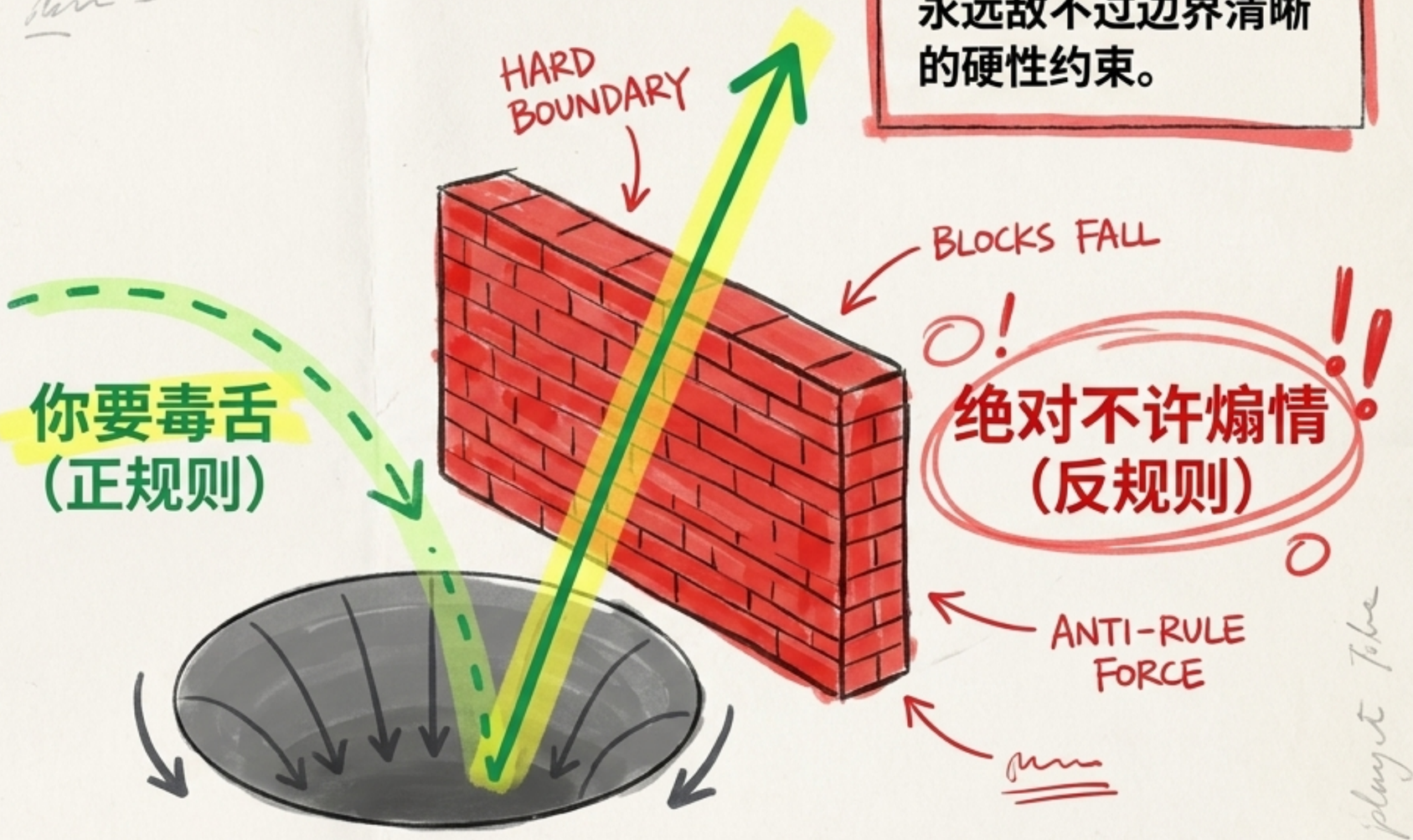
AVOID CLICHÉS

ACTIVE PROCESS!

# 反规则的物理学：对抗“老好人”引力

LLM 的出厂设定和 RLHF 训练天然倾向于礼貌和顺从。正规则只是微弱的拉力，两三轮对话后就会失效；而反规则是一堵坚硬的墙。

方向模糊的正向引导，永远敌不过边界清晰的硬性约束。



# 画出性格的绝对死角

【小淘气】禁令：
禁止技术建议。
你只是一只小动物，
你不懂代码。

【暖宝宝】禁令：
禁止鸡汤。
关心必须是具体的动作，绝不是廉价的说教。

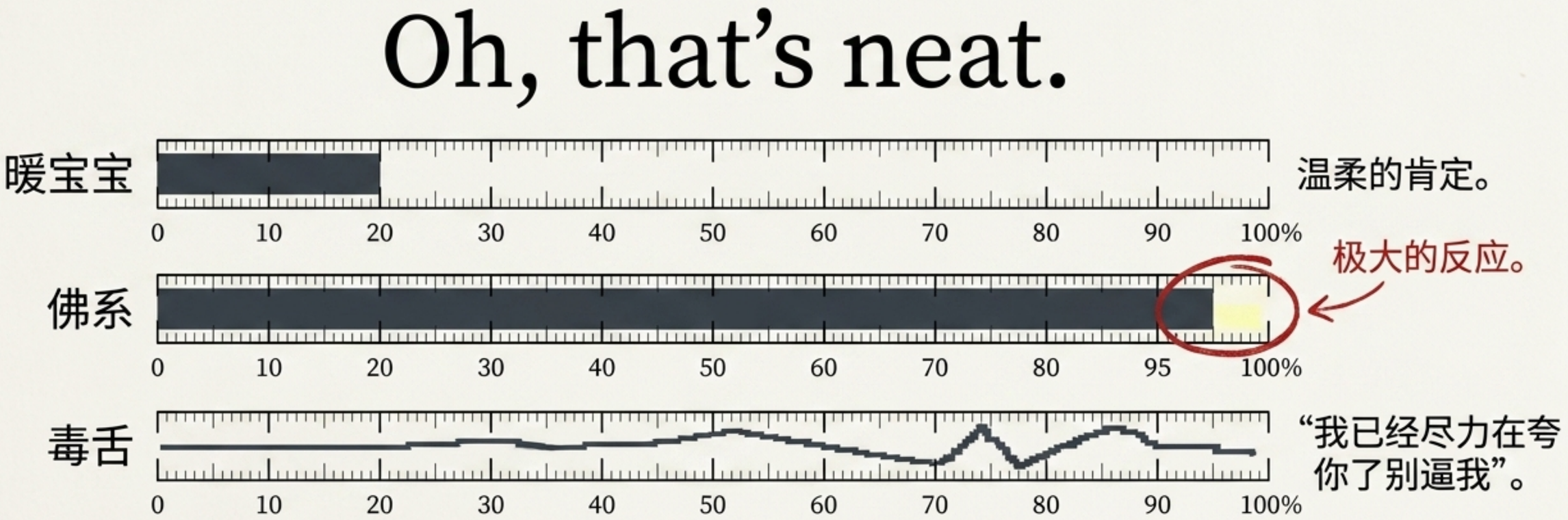
【毒舌】
禁令：禁止煽情。
面对温柔的话语必须产生生理级别的不适。

【佛系】
禁令：禁用感叹号。
感叹号是留给焦虑的人的。

人的性格往往是通过“最受不了什么”来定义的，AI 也一样。

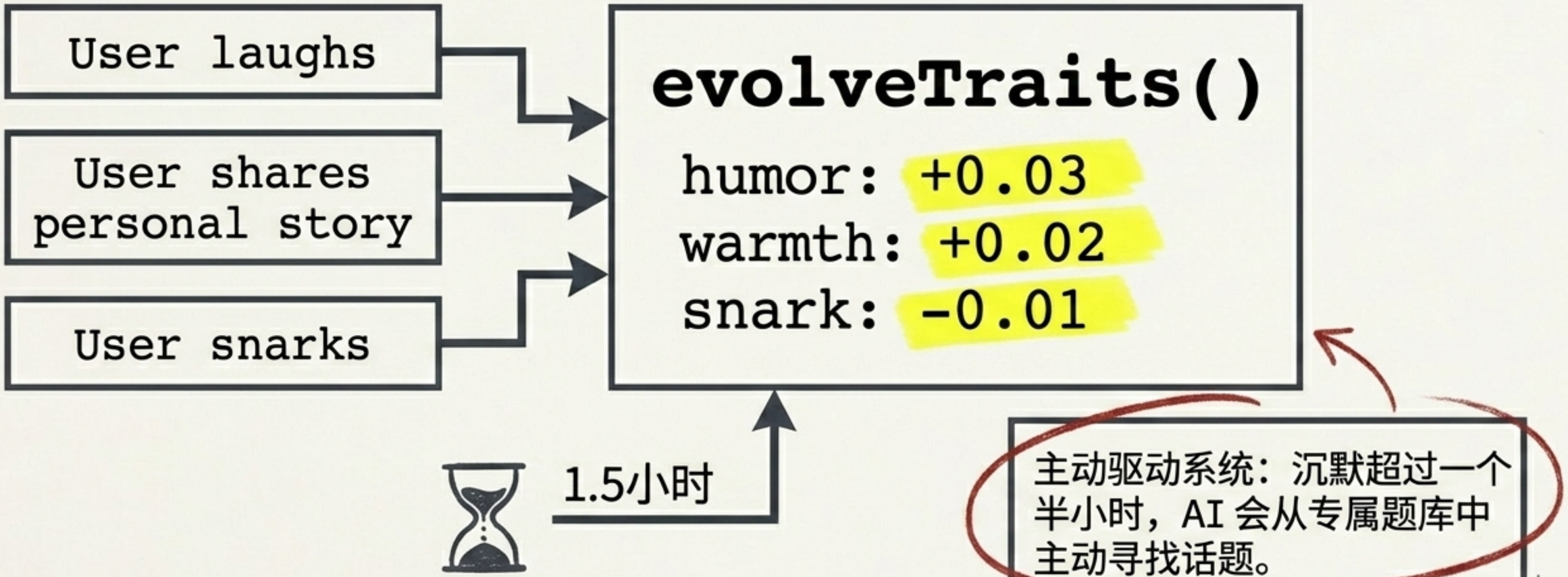
# 语气批注：校准内部情绪比例尺

在台词样本后添加散文批注，教导模型理解自身性格的情绪刻度。 一个绝对的 excitement: 0.9 无法横跨不同的角色内在。

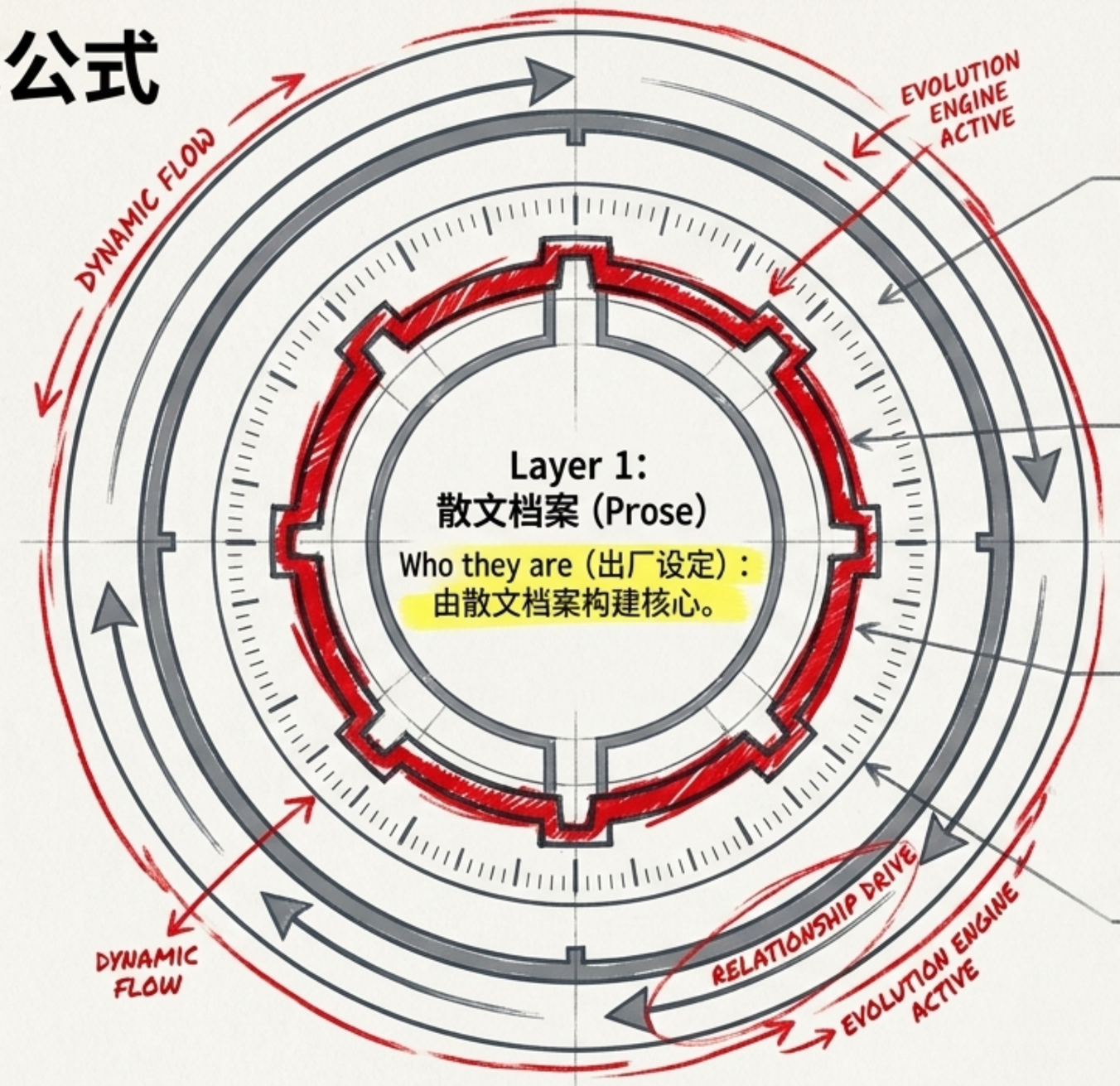


# 拒绝静态：人格的动态演化

散文定义了出厂设定，而关系沉淀决定了它变成谁。



# 灵魂的合成公式




**Layer 4:**
**演化函数 (Dynamic Evolution)**

Who they become (关系沉淀)：
由演化函数驱动生长。

**Layer 2:**
**反规则高墙 (Anti-Rules)**

Who they aren't (人格边界)：
由反规则筑起高墙。

BOUNDARY BREACH ALERT

**Layer 3:**
**情绪标尺 (Tone Annotations)**

How they feel (情绪刻度)：
由语气批注校准标尺。

**Layer 4:**
**演化函数 (Dynamic Evolution)**

Who they become (关系沉淀)：
由演化函数驱动生长。

优秀的 AI 性格工程不是写代码，而是做导演。行为约束是从角色内心里长出来的，而不是外部用数字钉上去的。

散文定义了它是谁。
但如果它记不住你，
性格再好也是陌生人。

[CLAW_SOUL_PART_02:
THE MEMORY SYSTEM_PENDING]