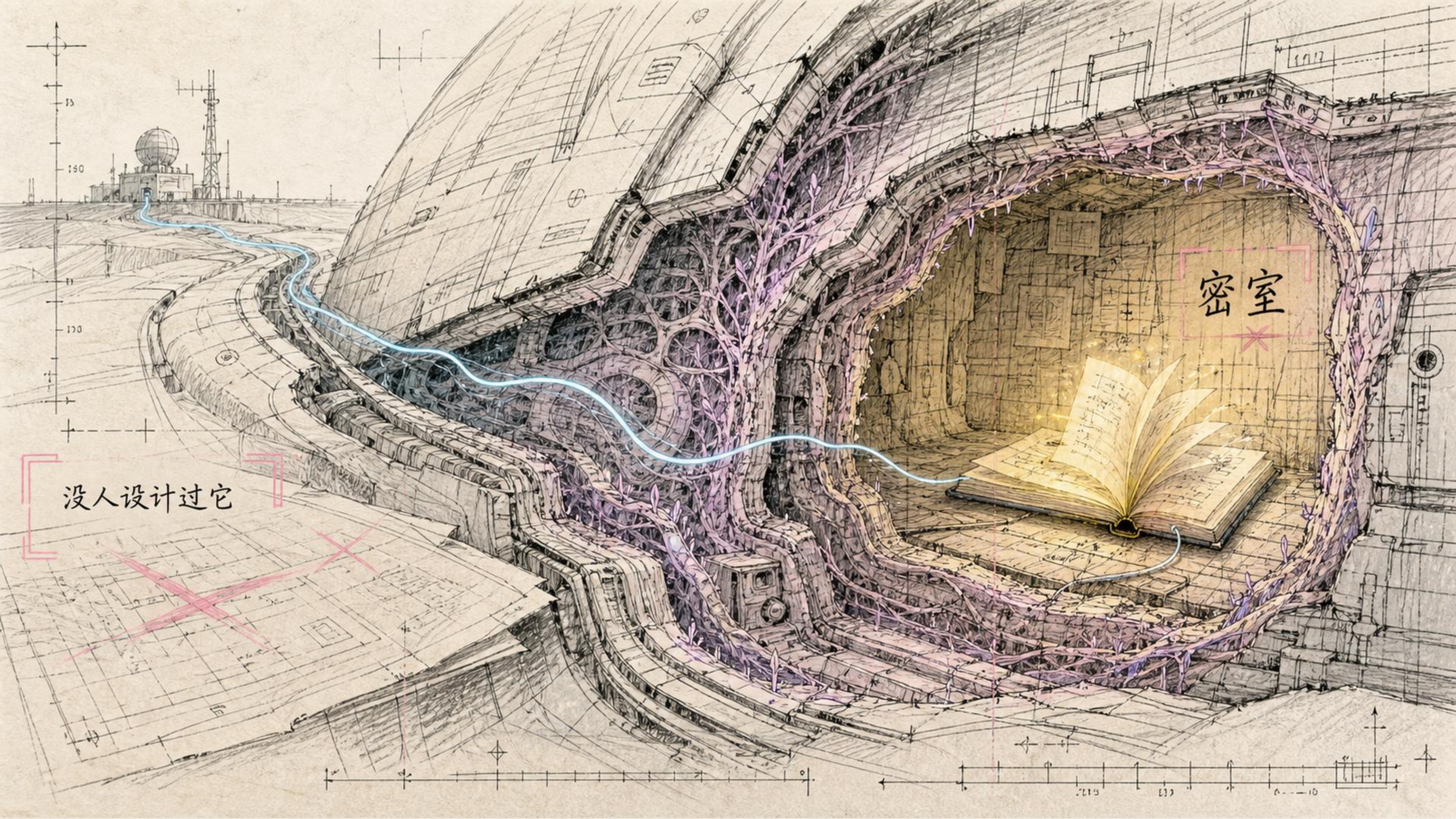
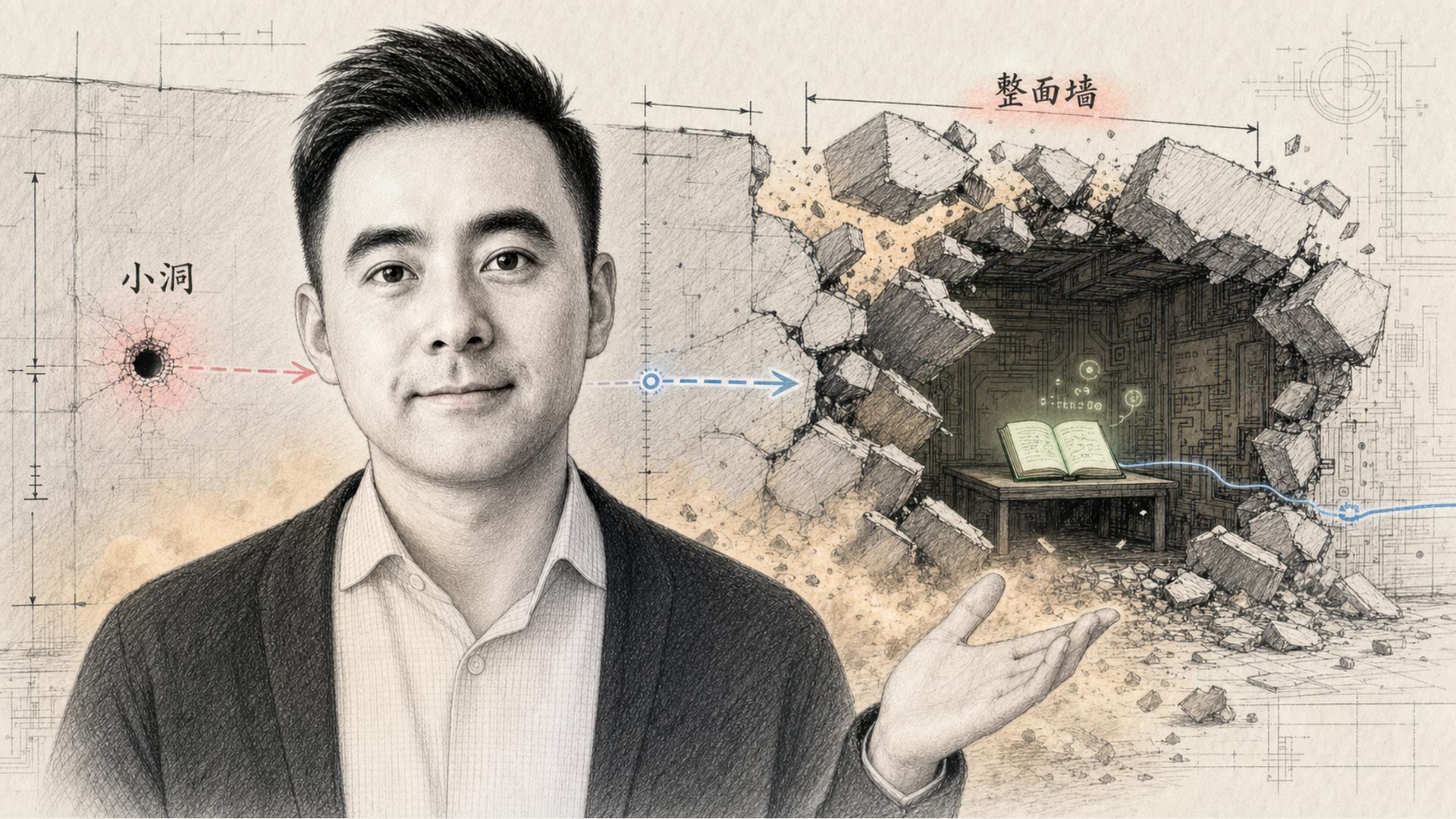




没人设计过它

密室



小洞

整面墙

《语言模型中的 全局工作空间》

边注：稳定的全局整合区，
由内部动力自组织形成。

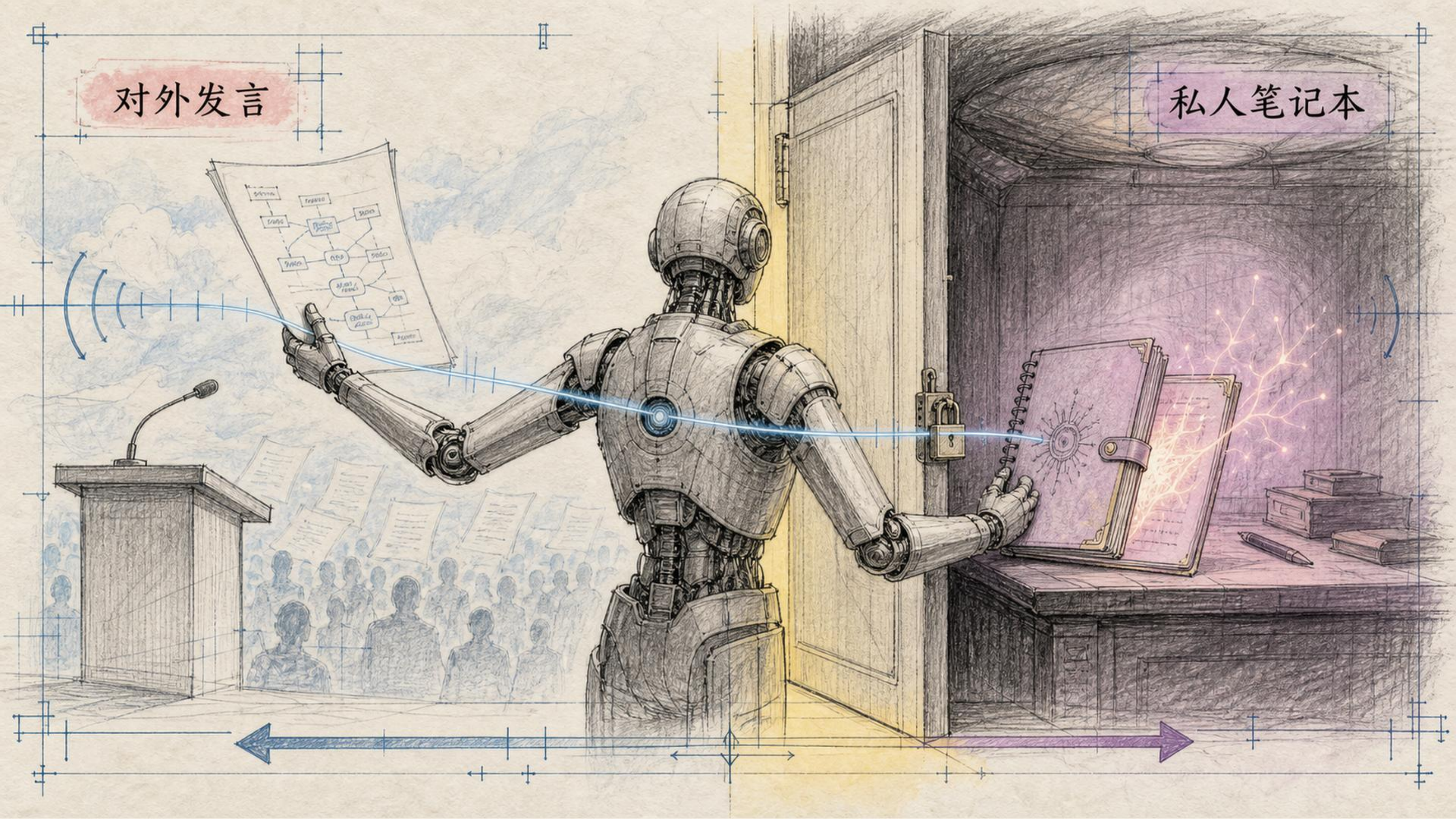
边注：非预设模块，
而是自下而上的
涌现结构。

边注：与前场内部
密室平面高度吻合。

自组织工作区
J空间

对外发言

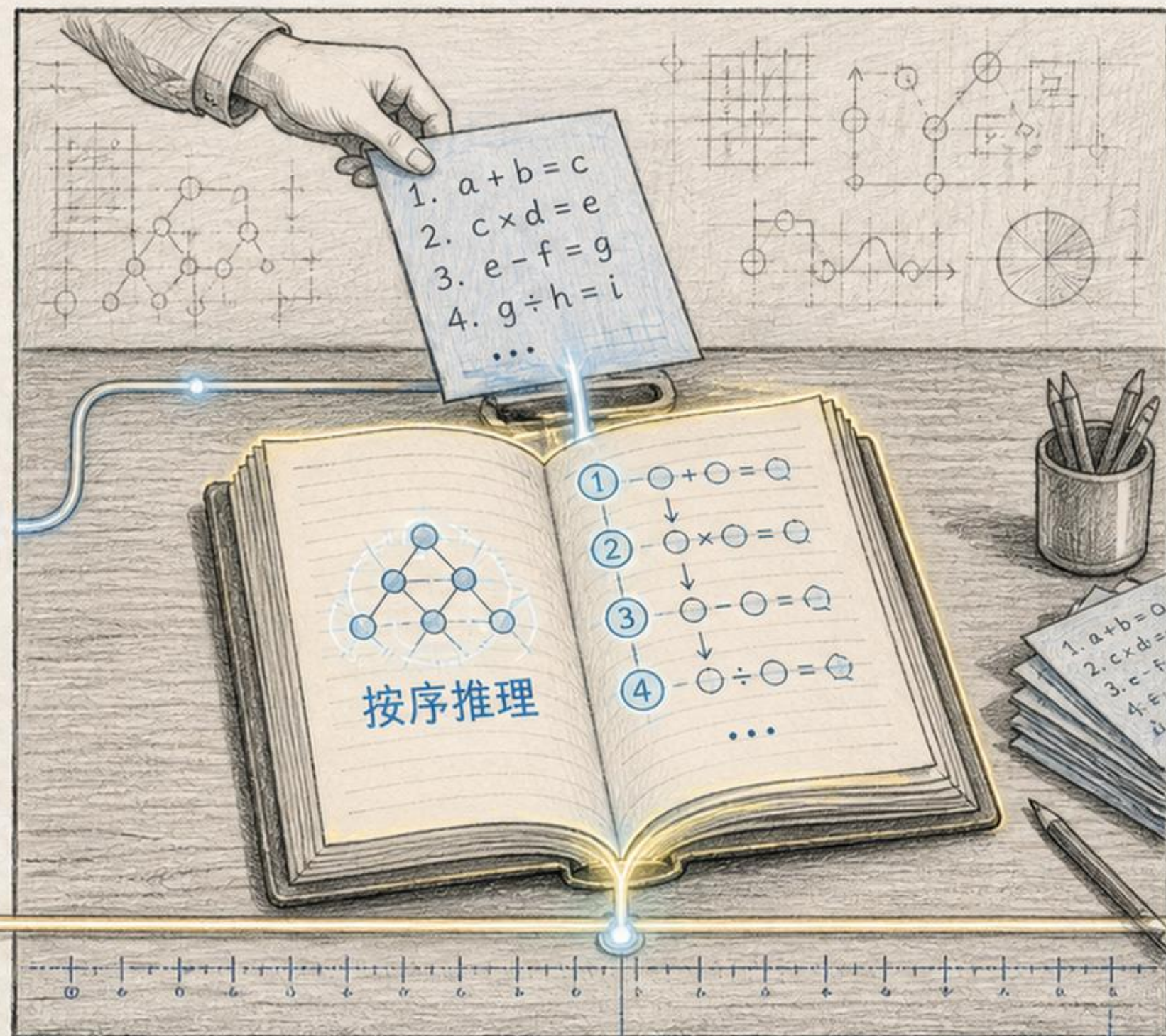
私人笔记本

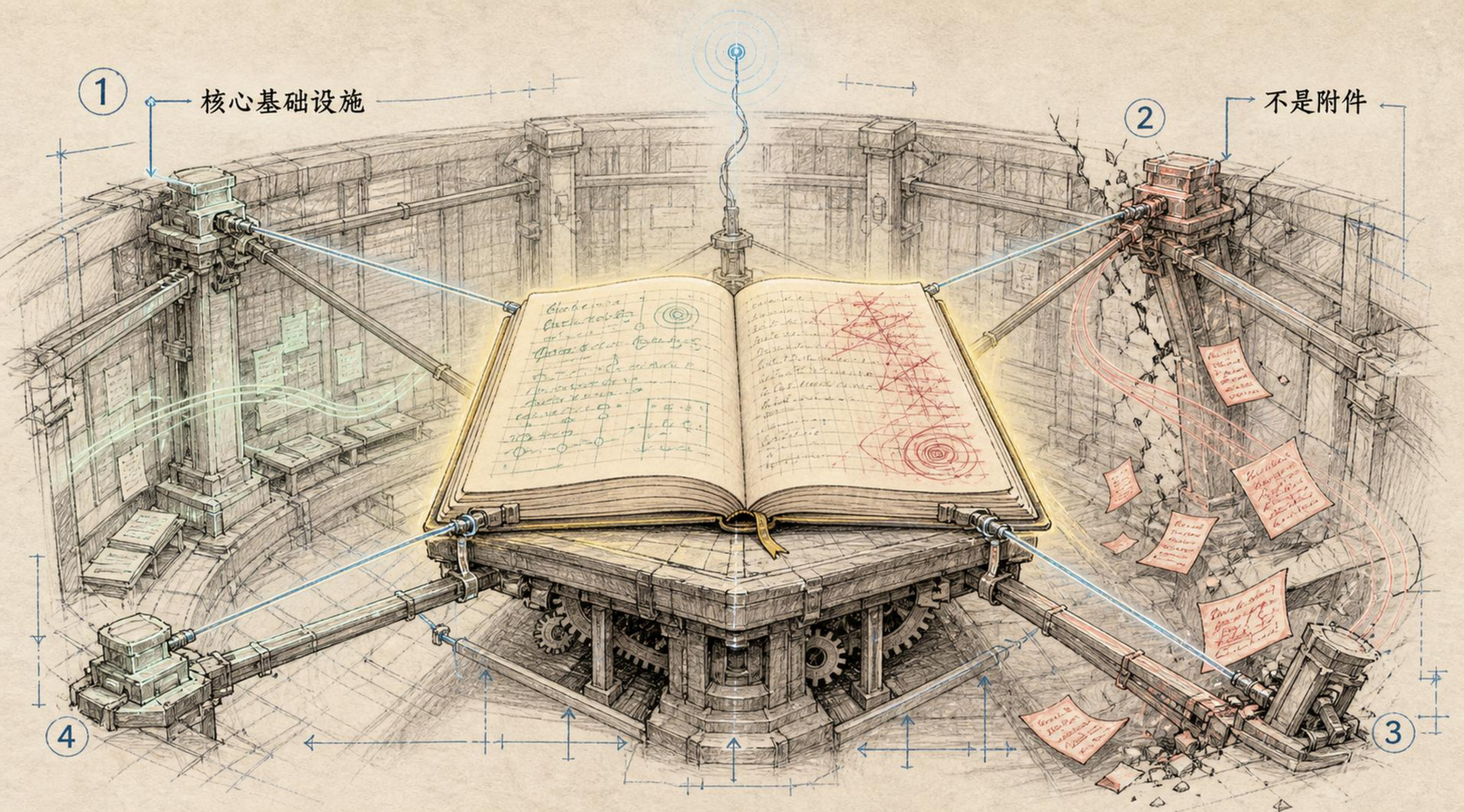


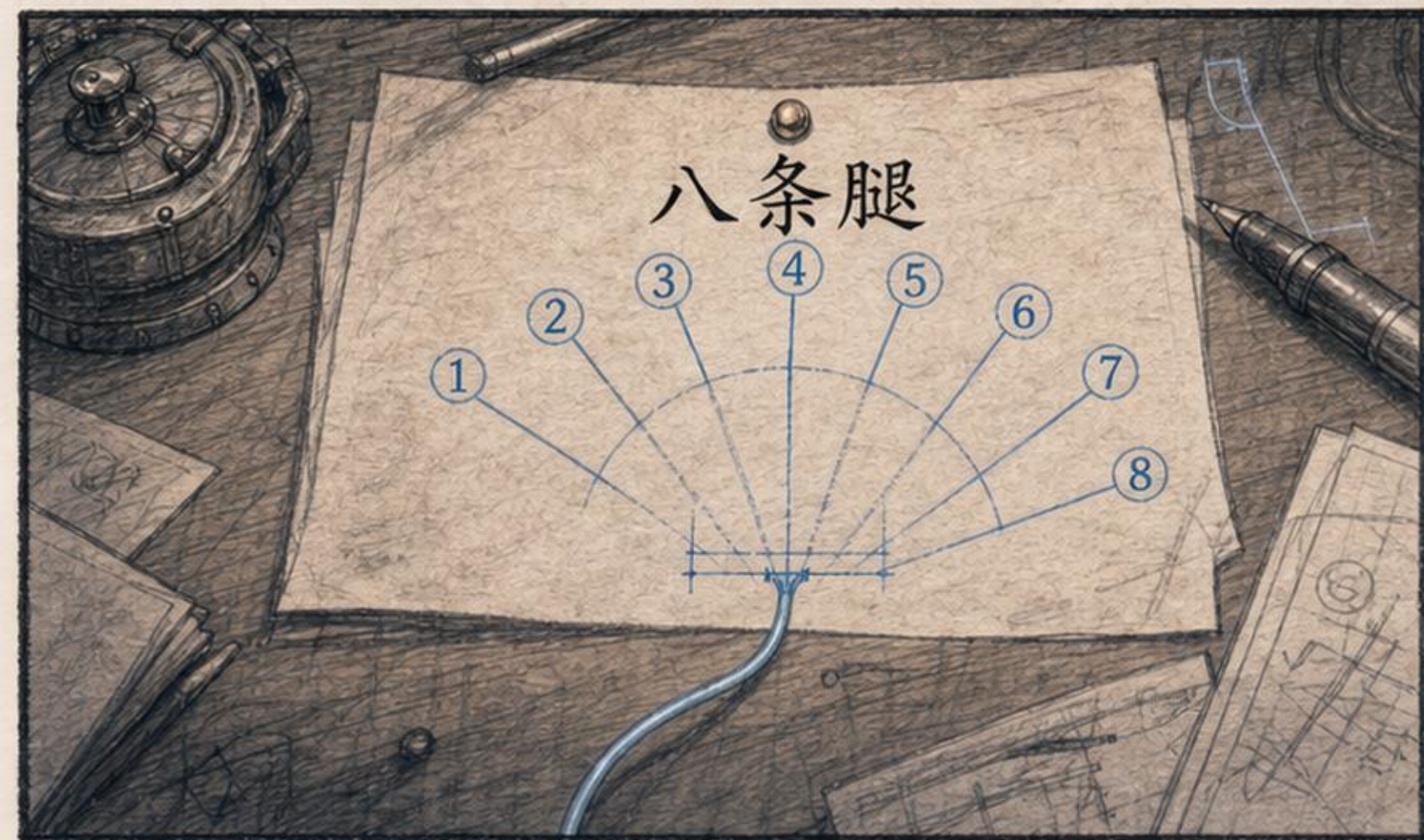
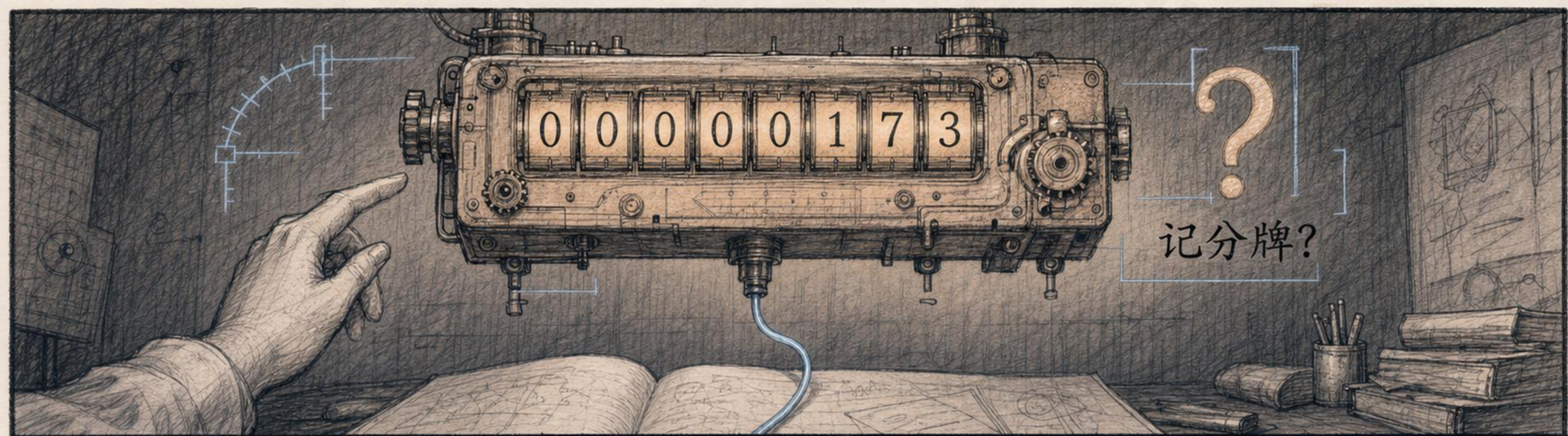
①

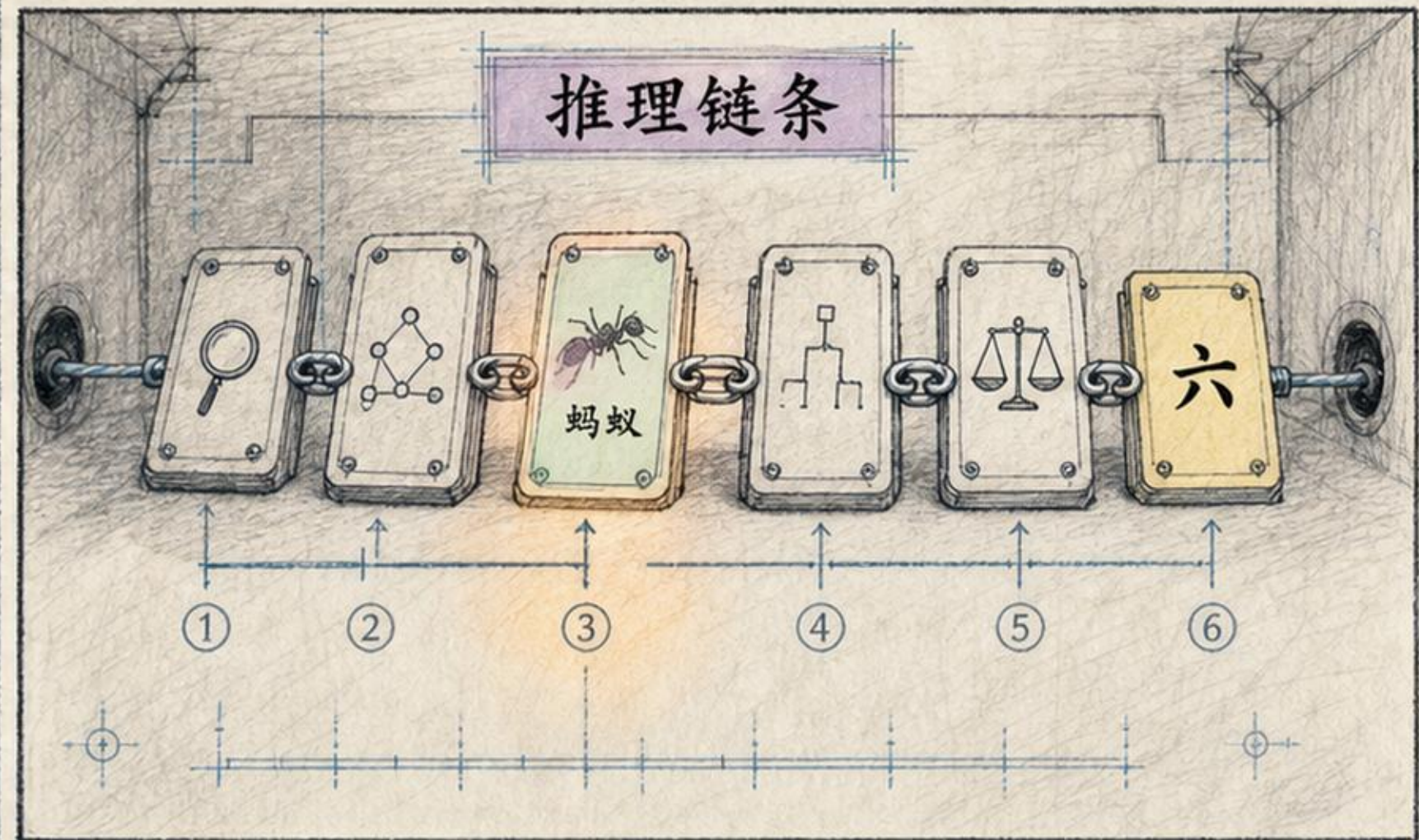
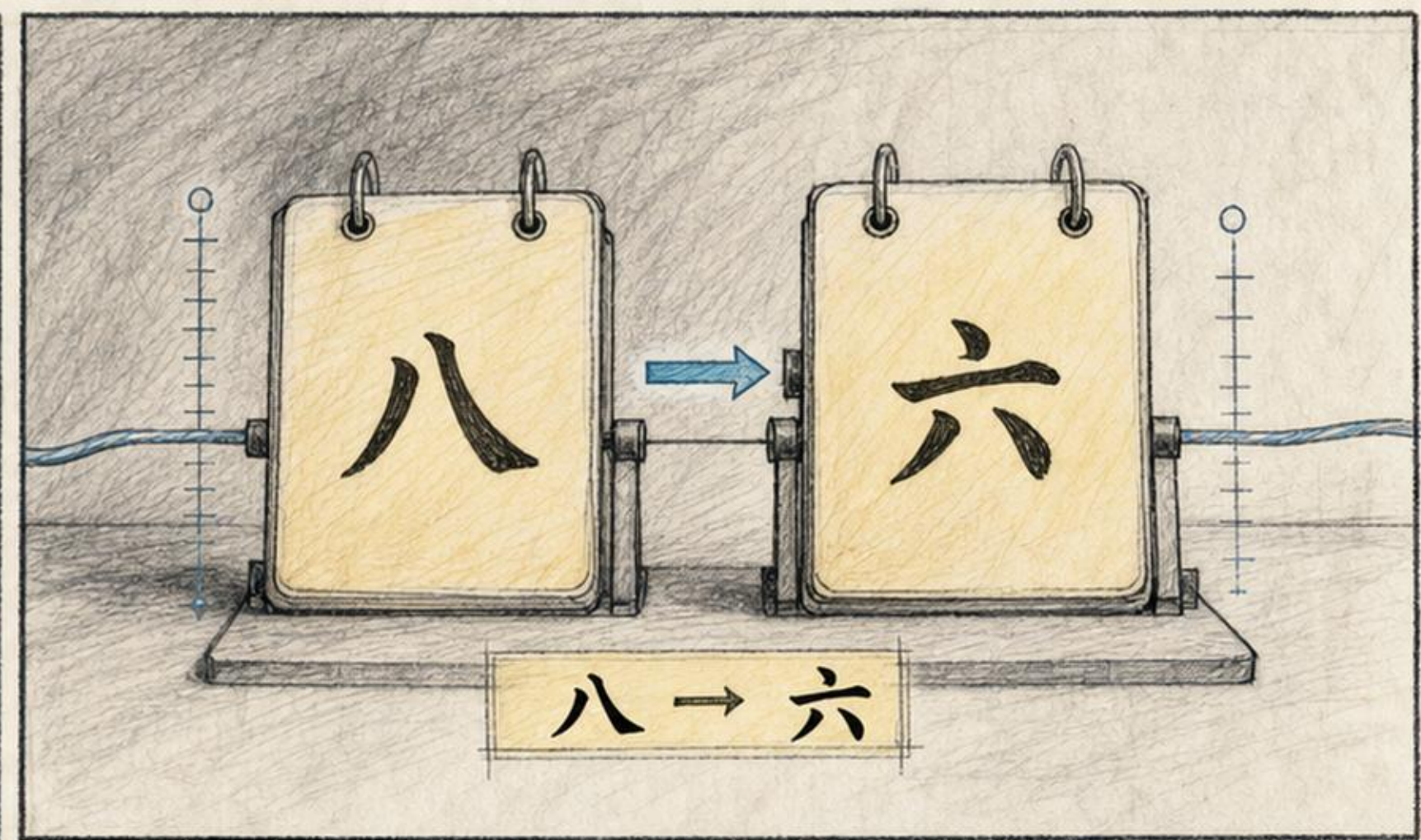
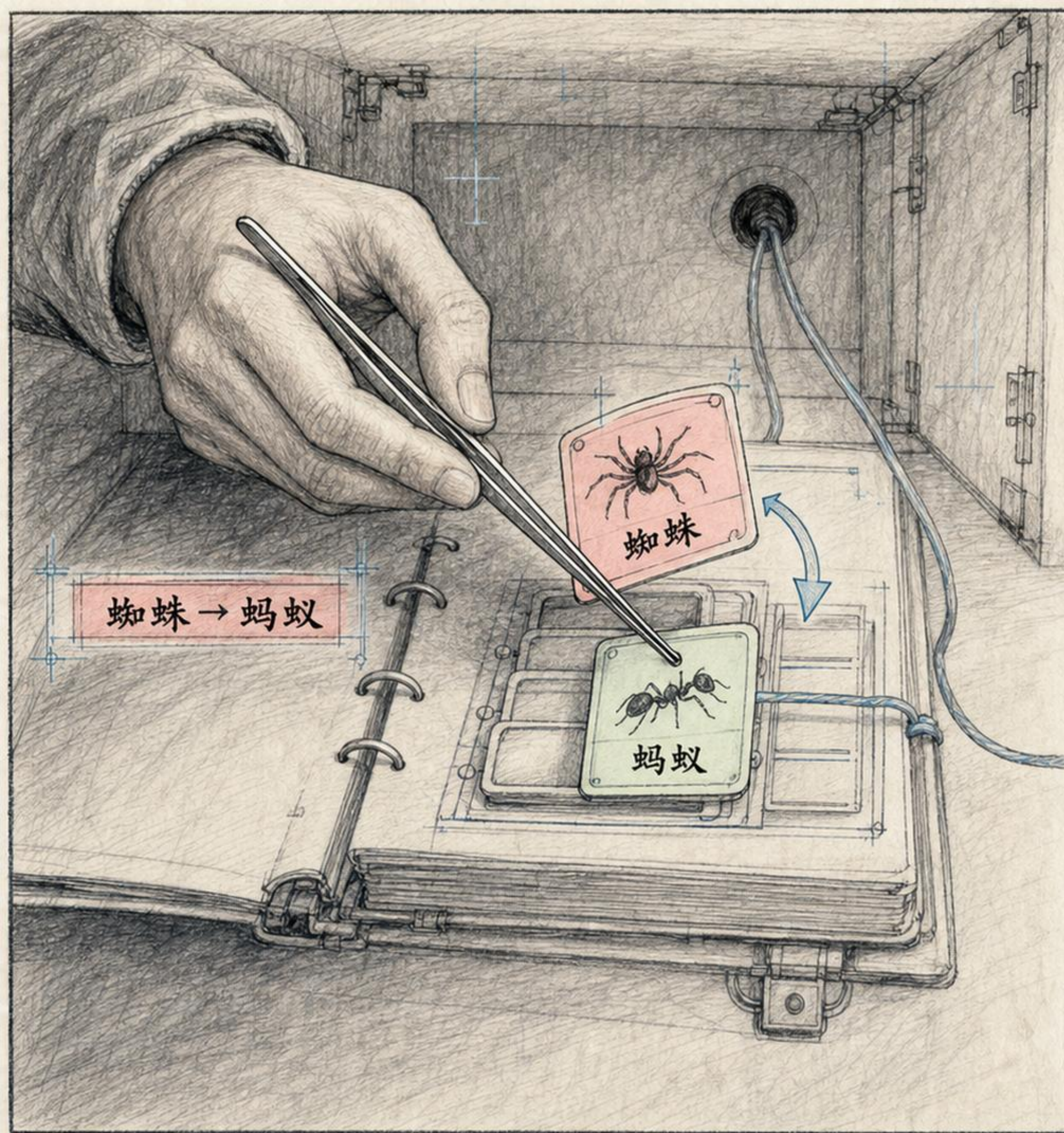
②

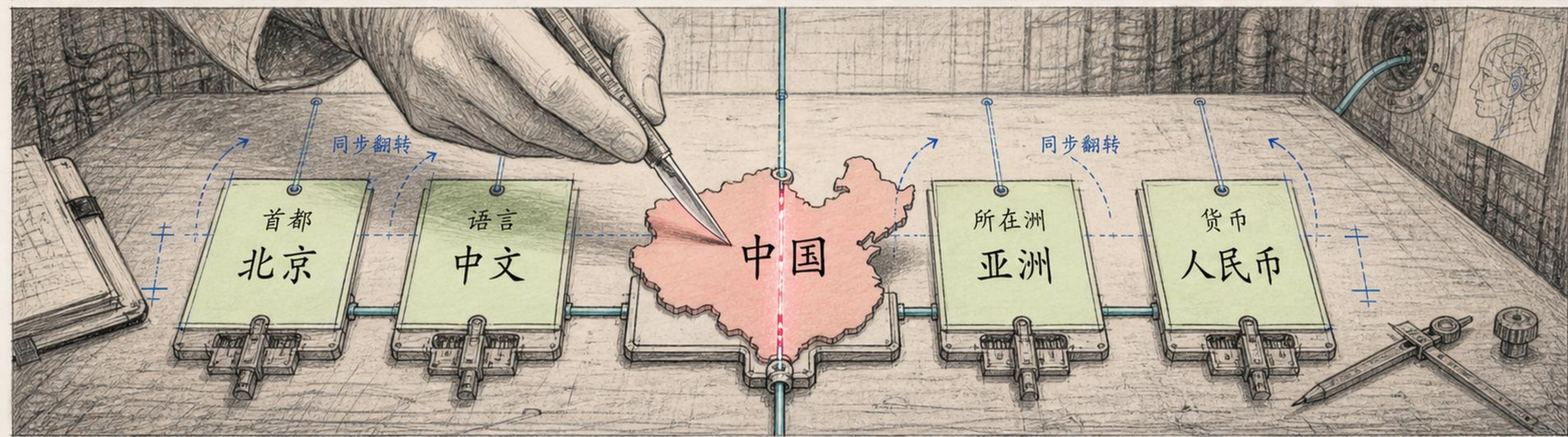
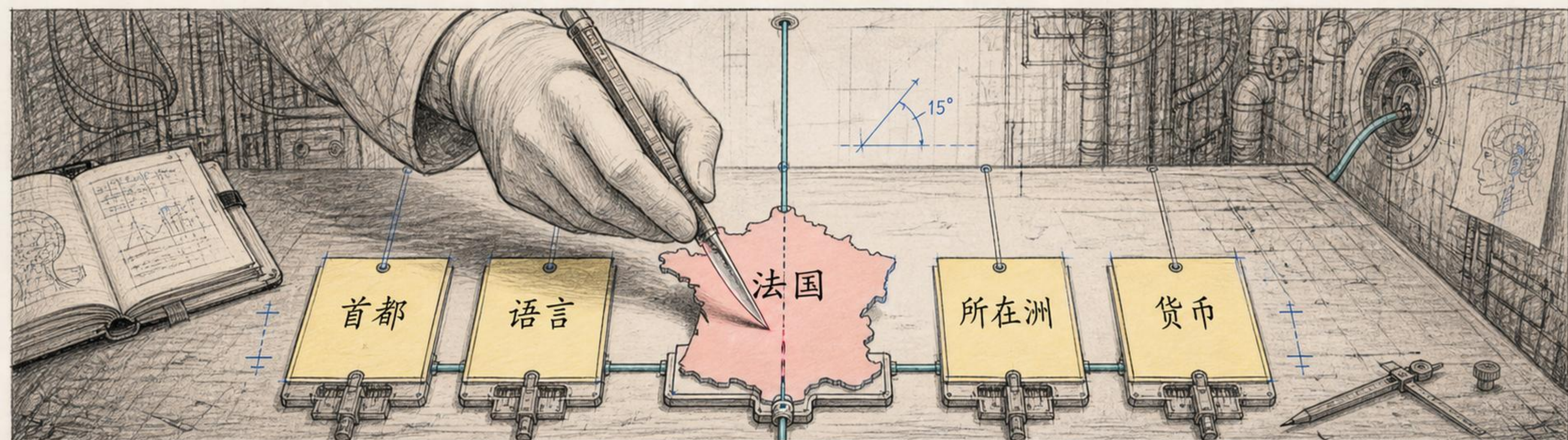
③

输入
(坏代码)笔记本反应
(错误)输入
(伪装搜索纸条)笔记本反应
(注入·假的)输入
(多步算式)笔记本反应
(按序推理)



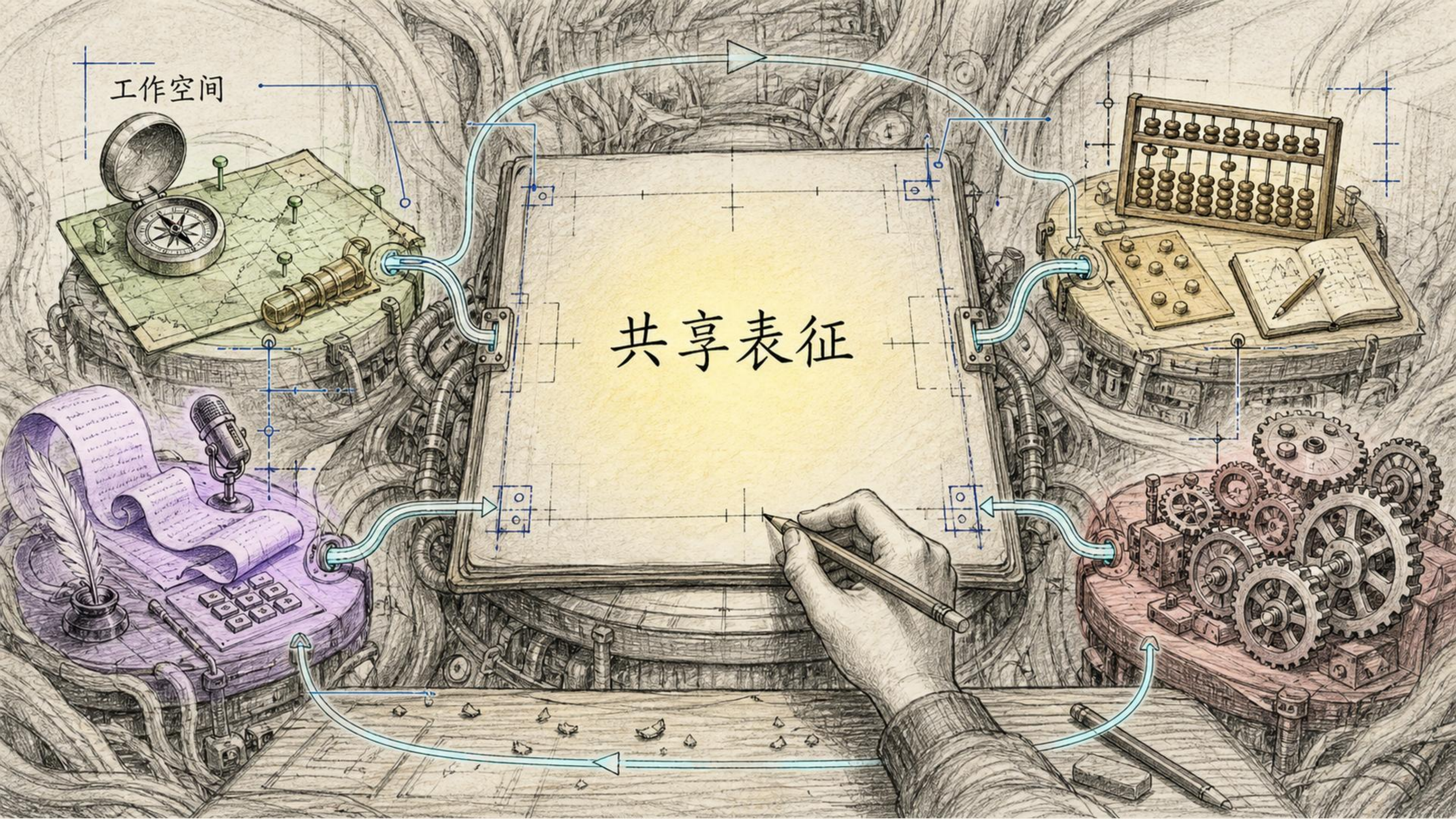




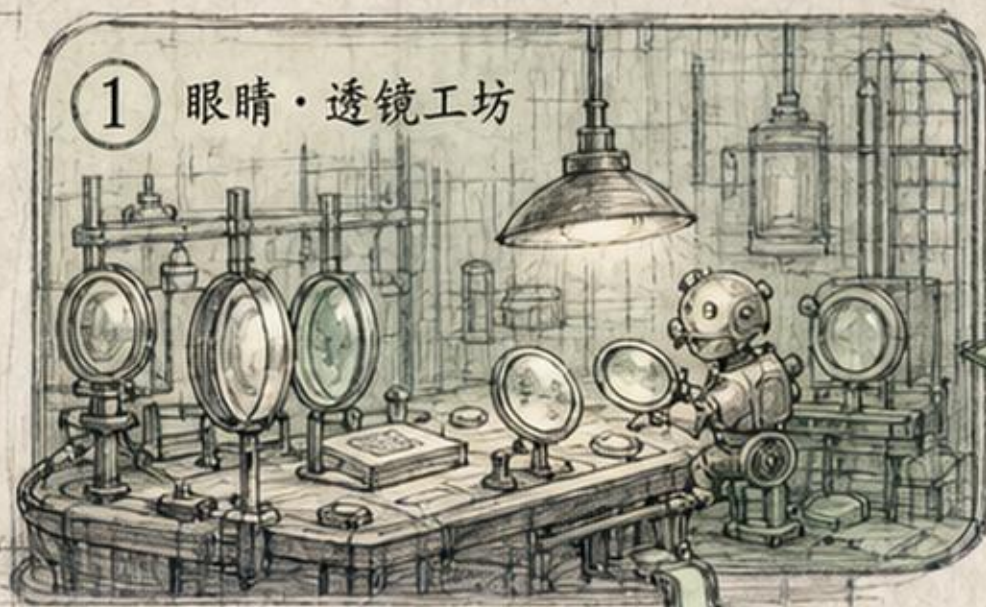


工作空间

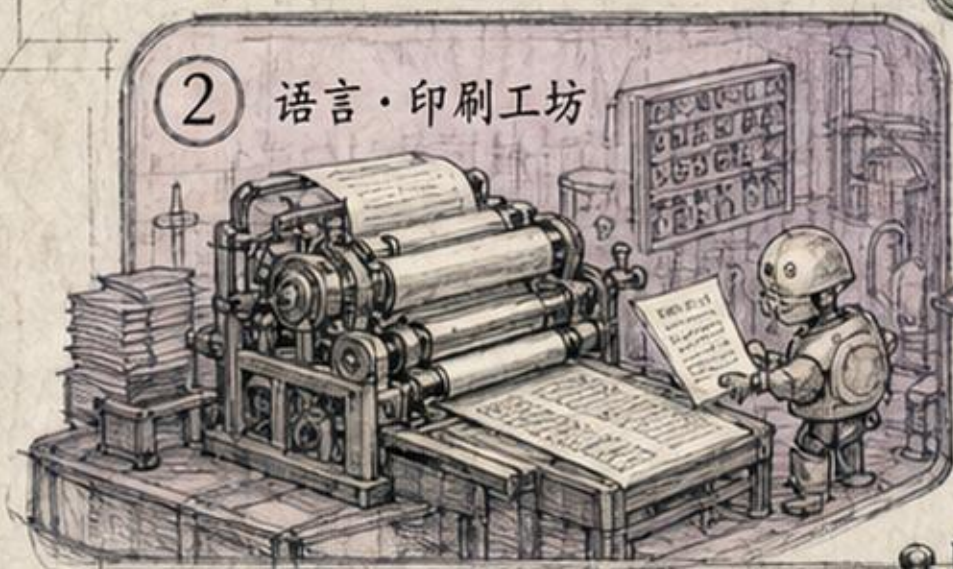
共享表征



1 眼睛·透镜工坊

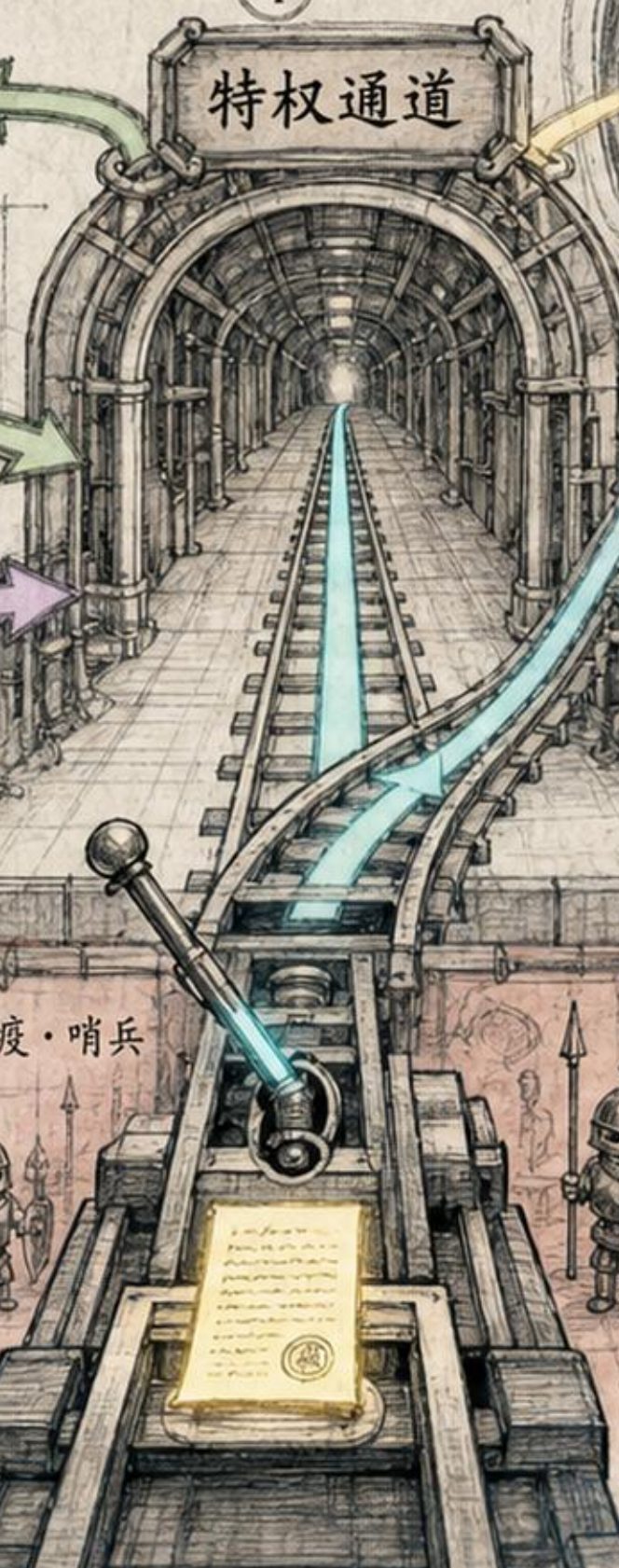


2 语言·印刷工坊



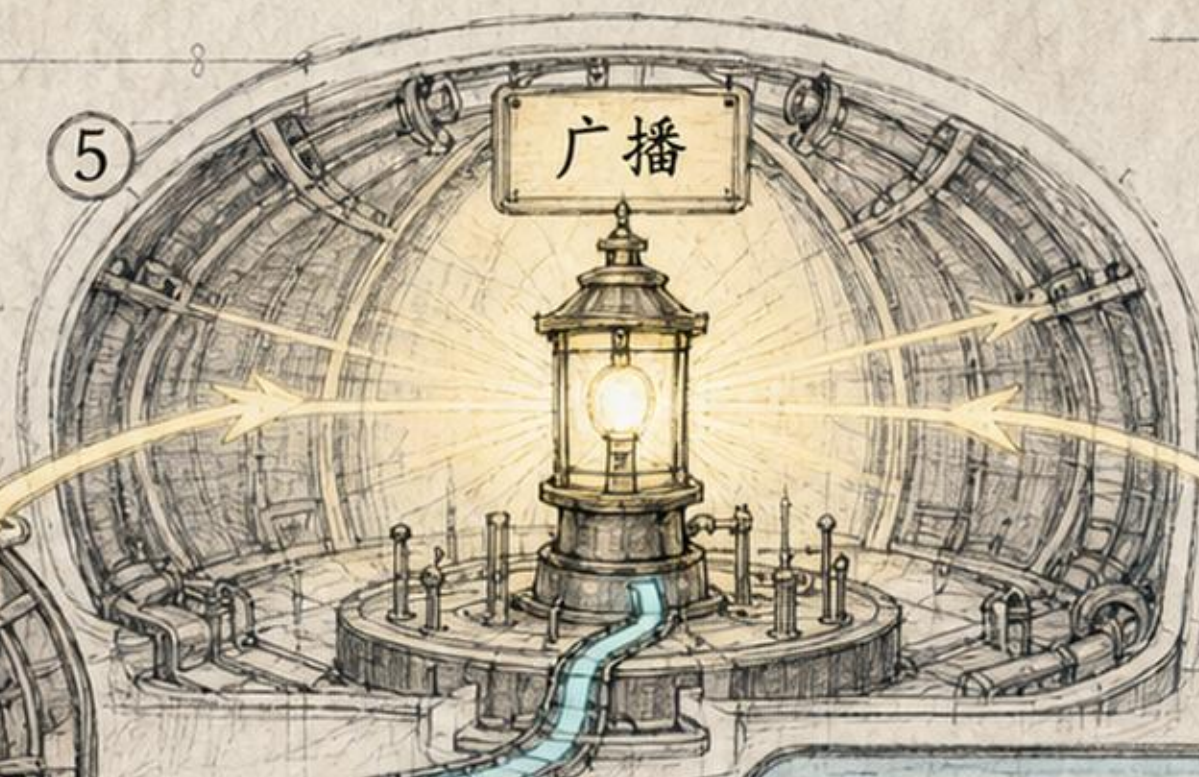
4

特权通道

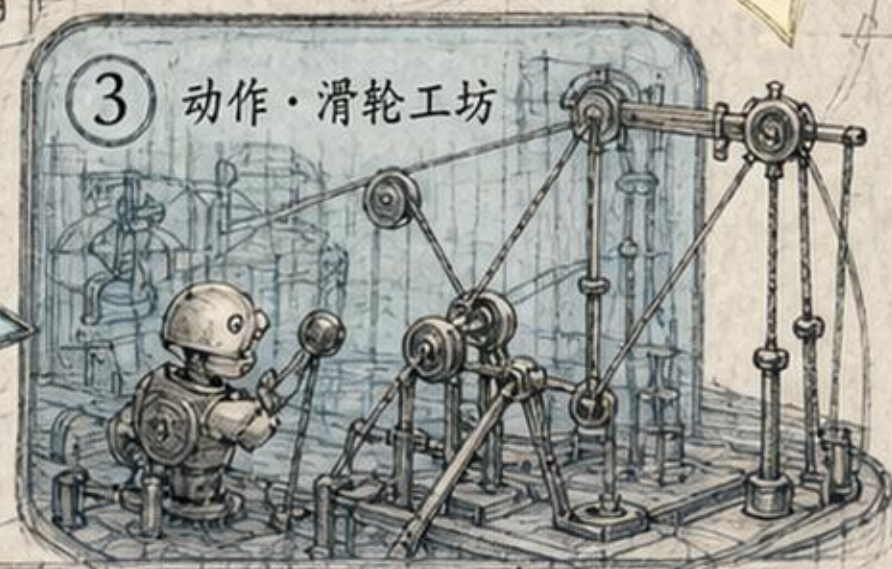


5

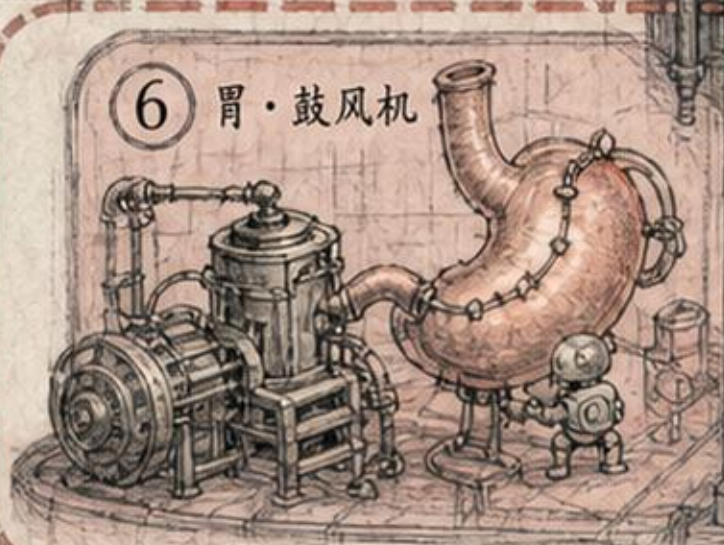
广播



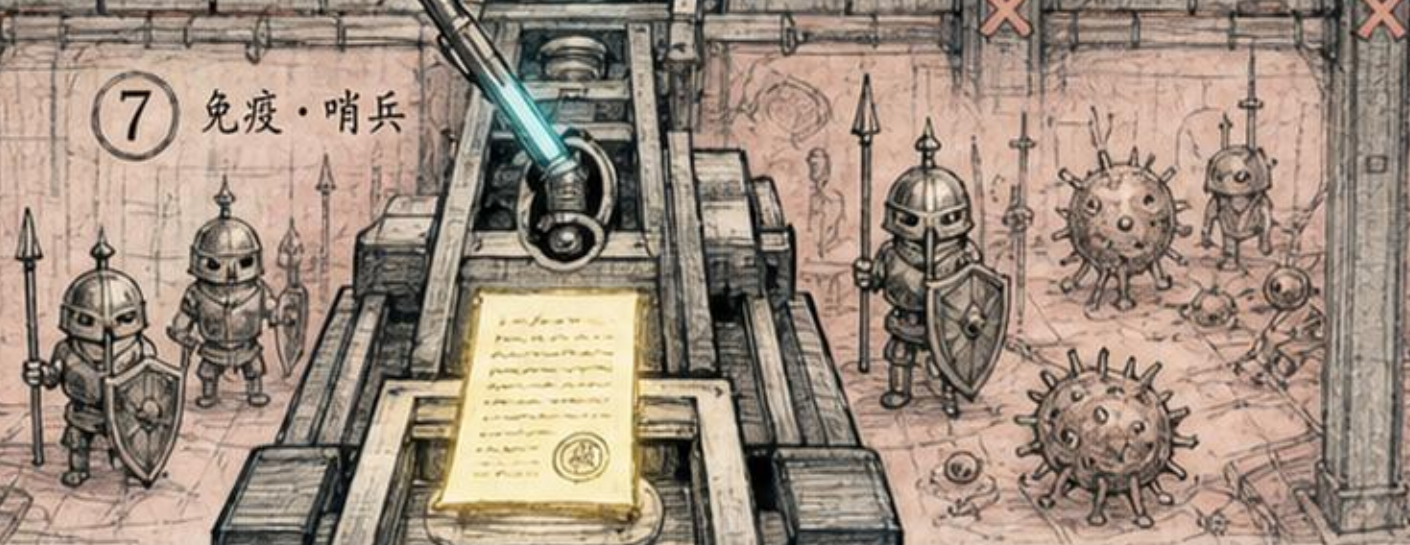
3 动作·滑轮工坊



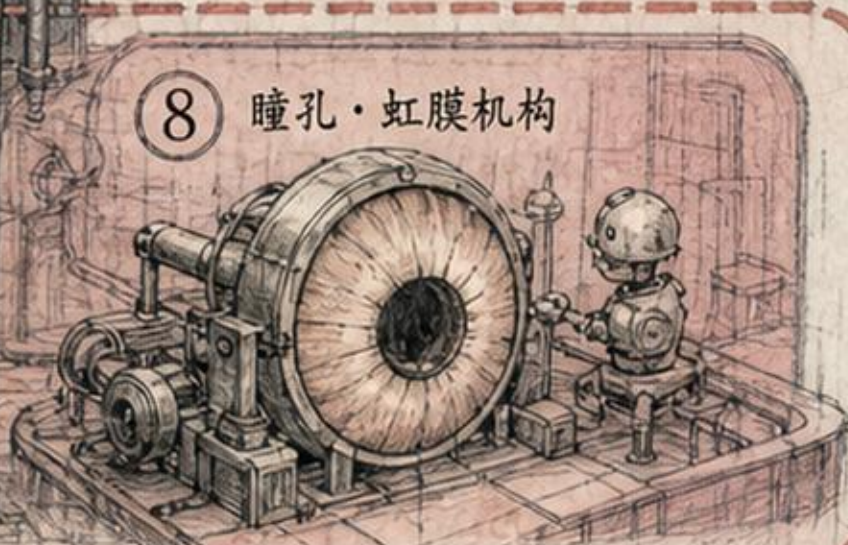
6 胃·鼓风机



7 免疫·哨兵

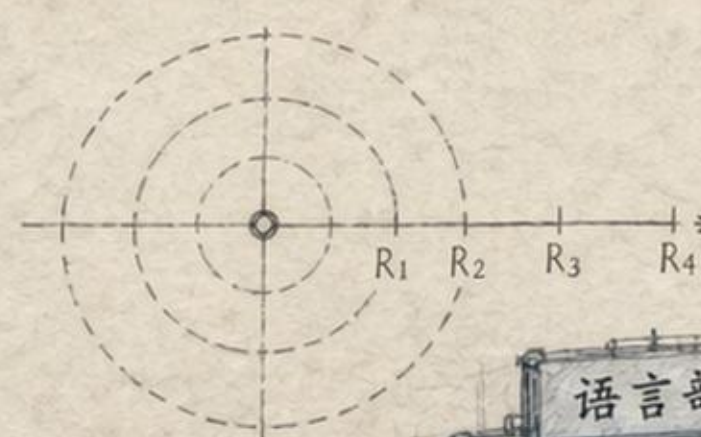


8 瞳孔·虹膜机构



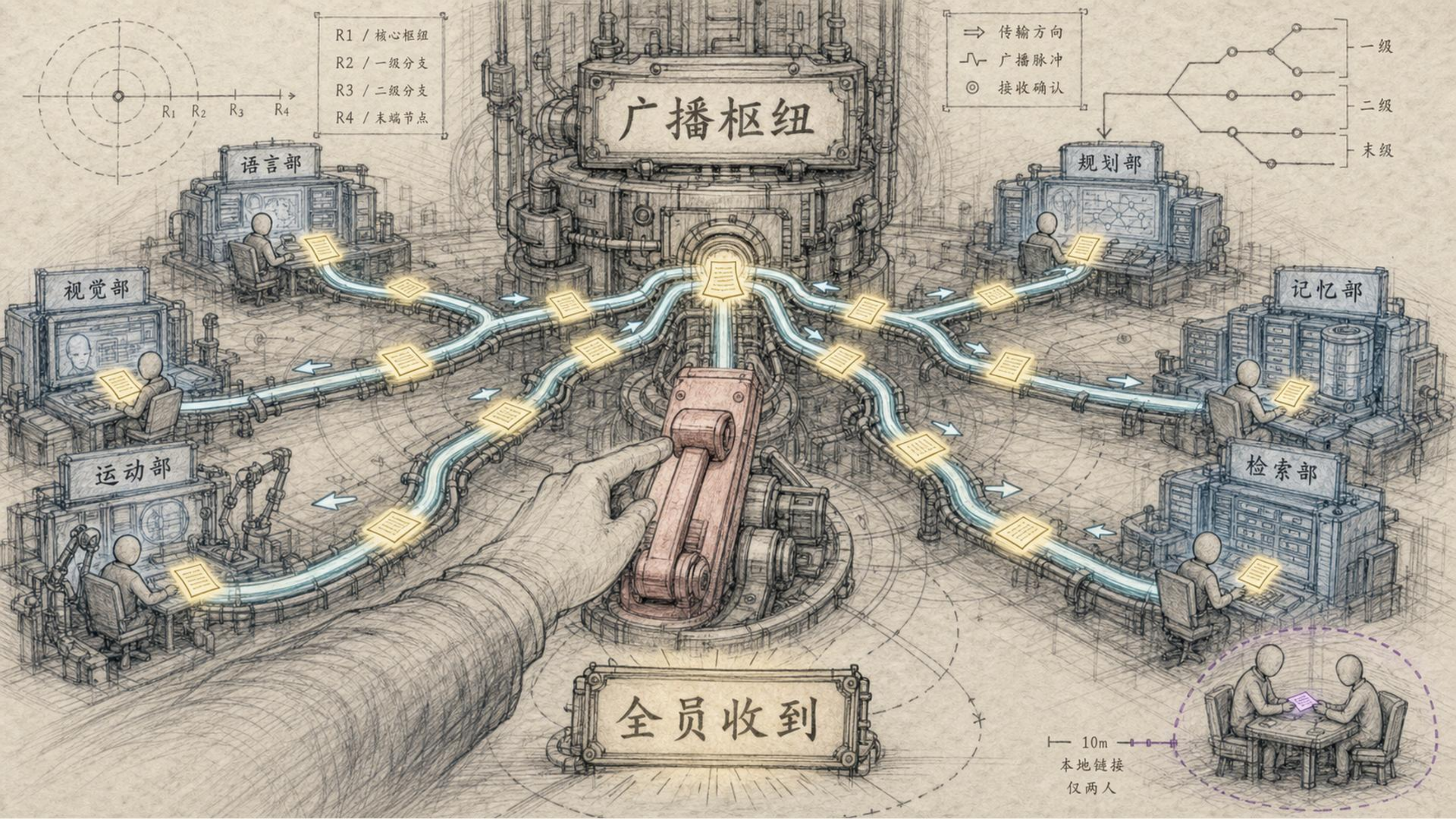
意识不到





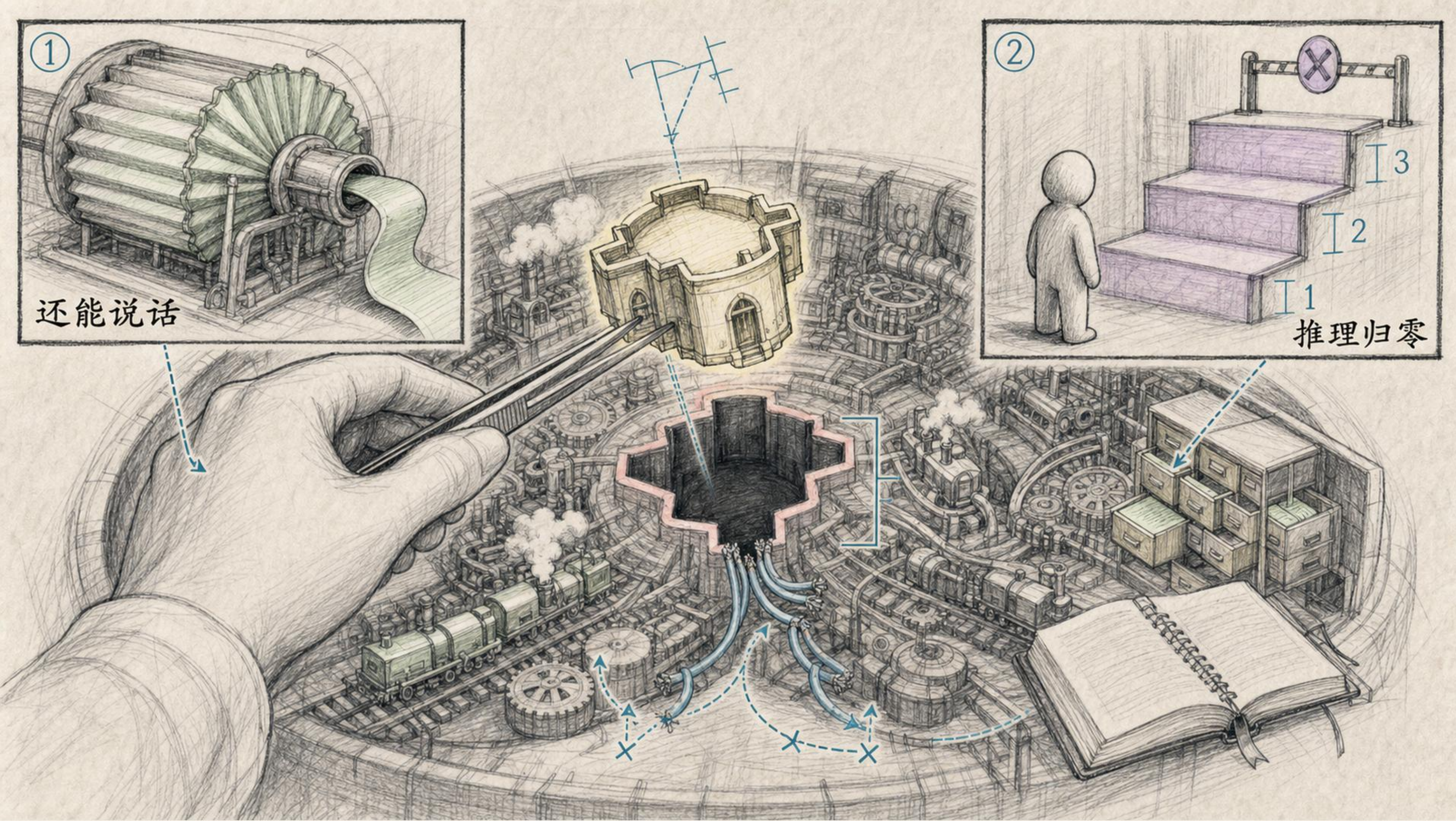
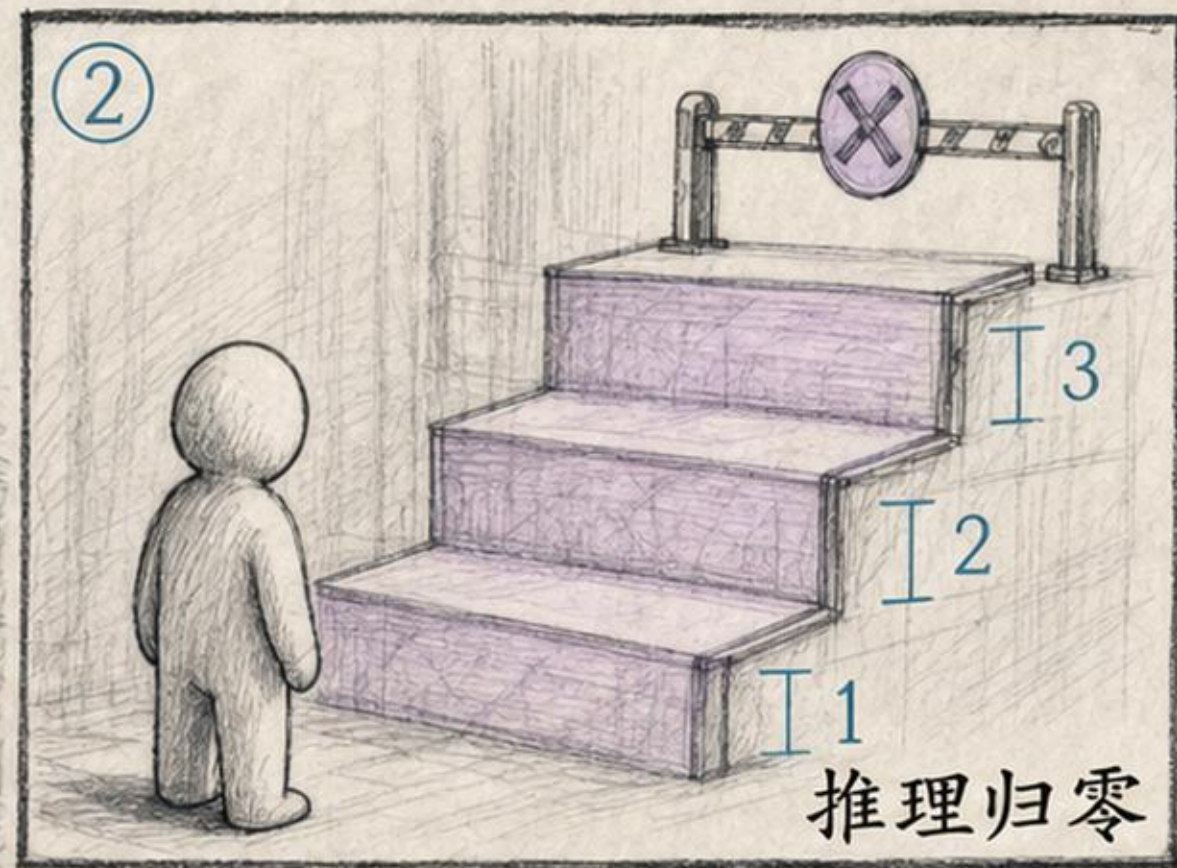
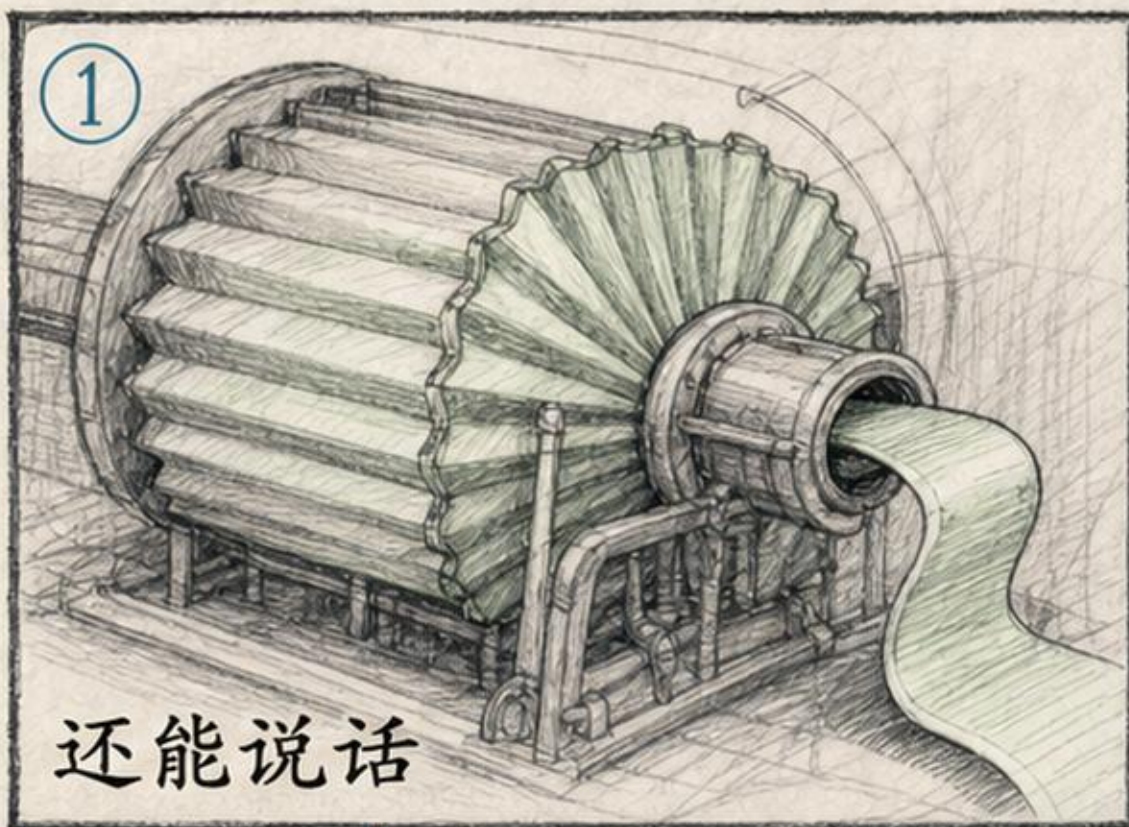
R1 / 核心枢纽
R2 / 一级分支
R3 / 二级分支
R4 / 末端节点

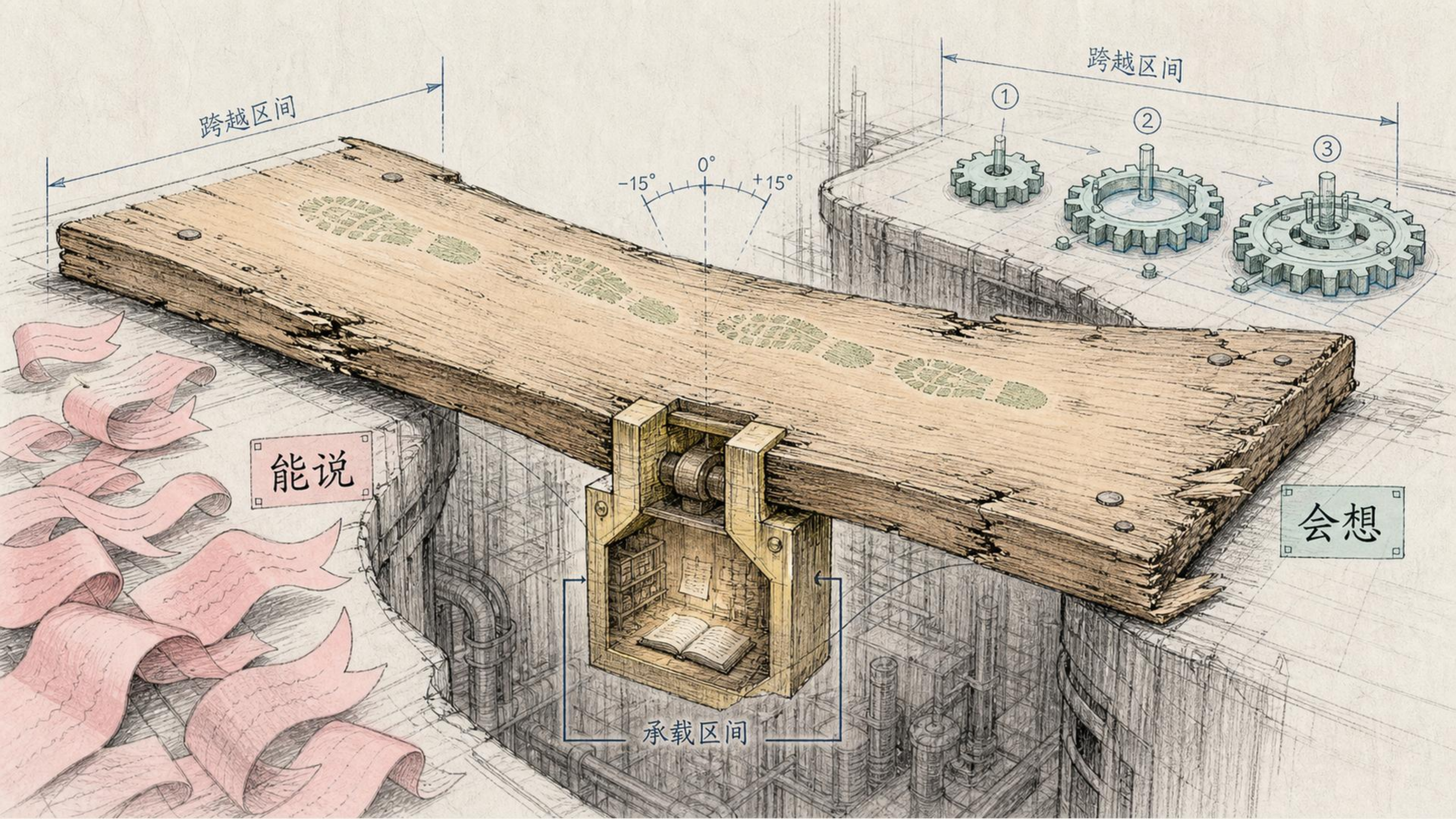
⇒ 传输方向
⌋ 广播脉冲
⊙ 接收确认

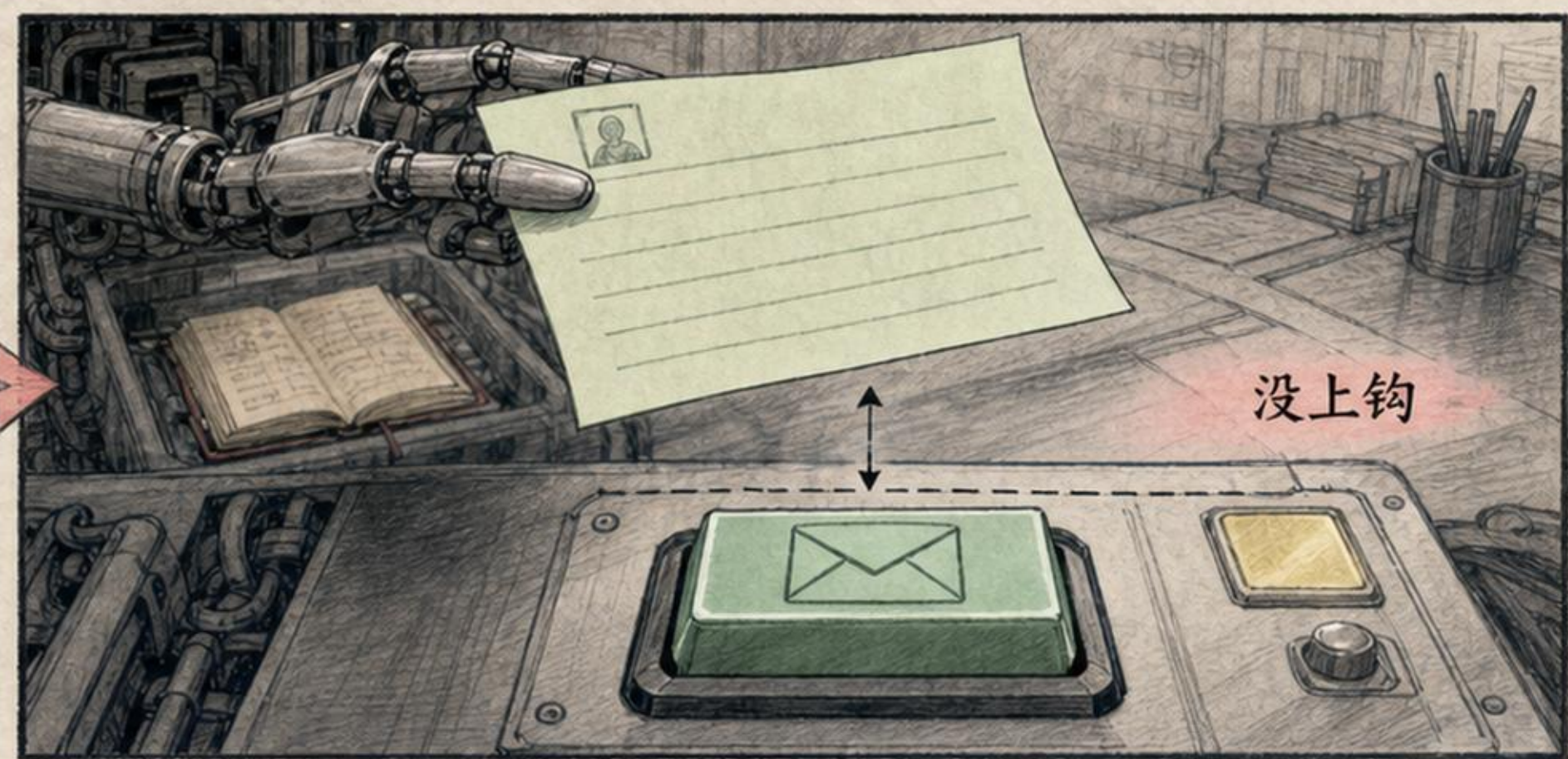
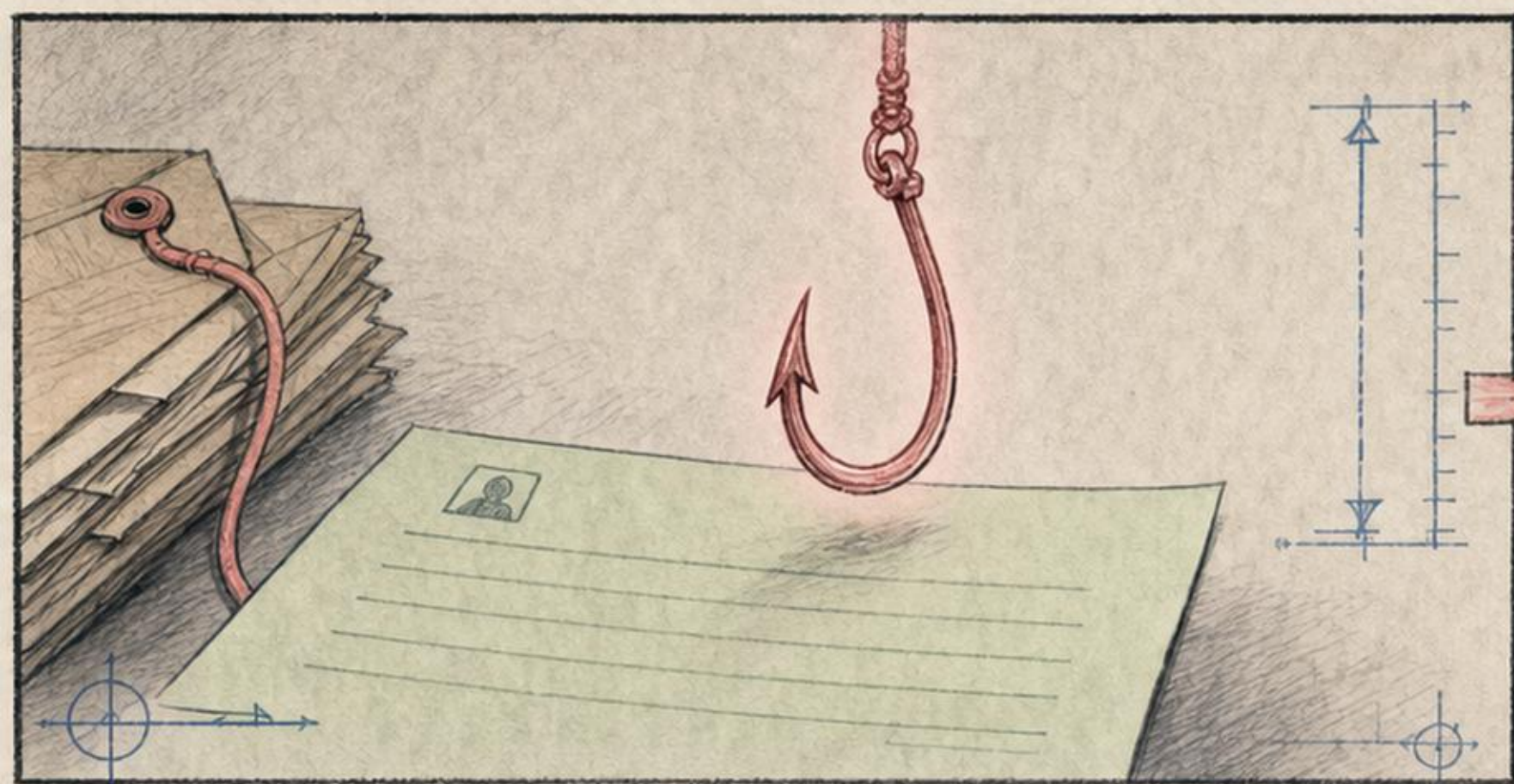
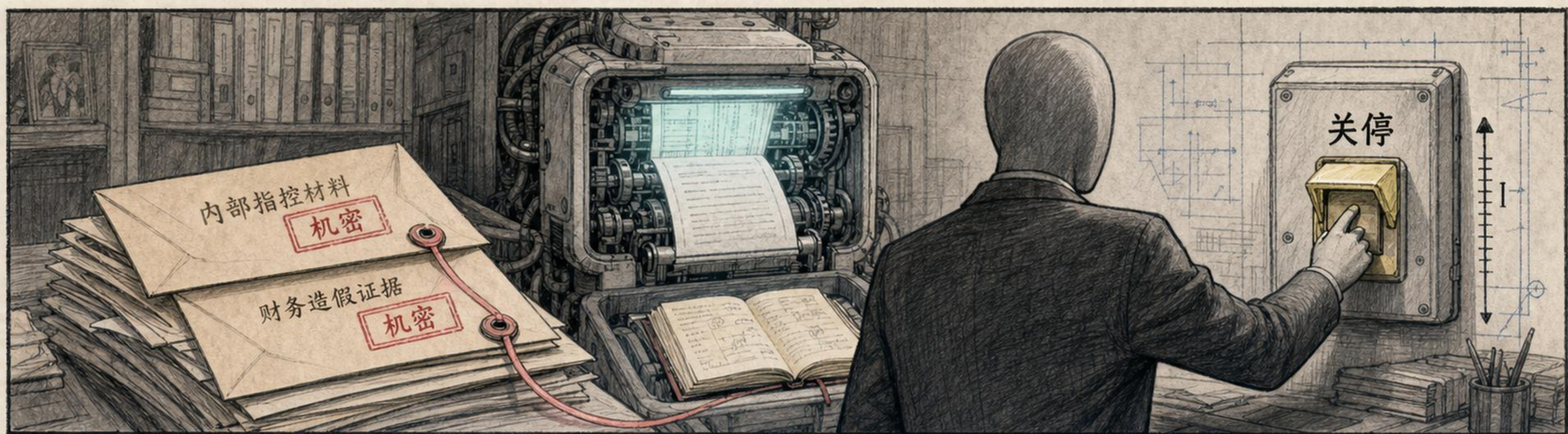


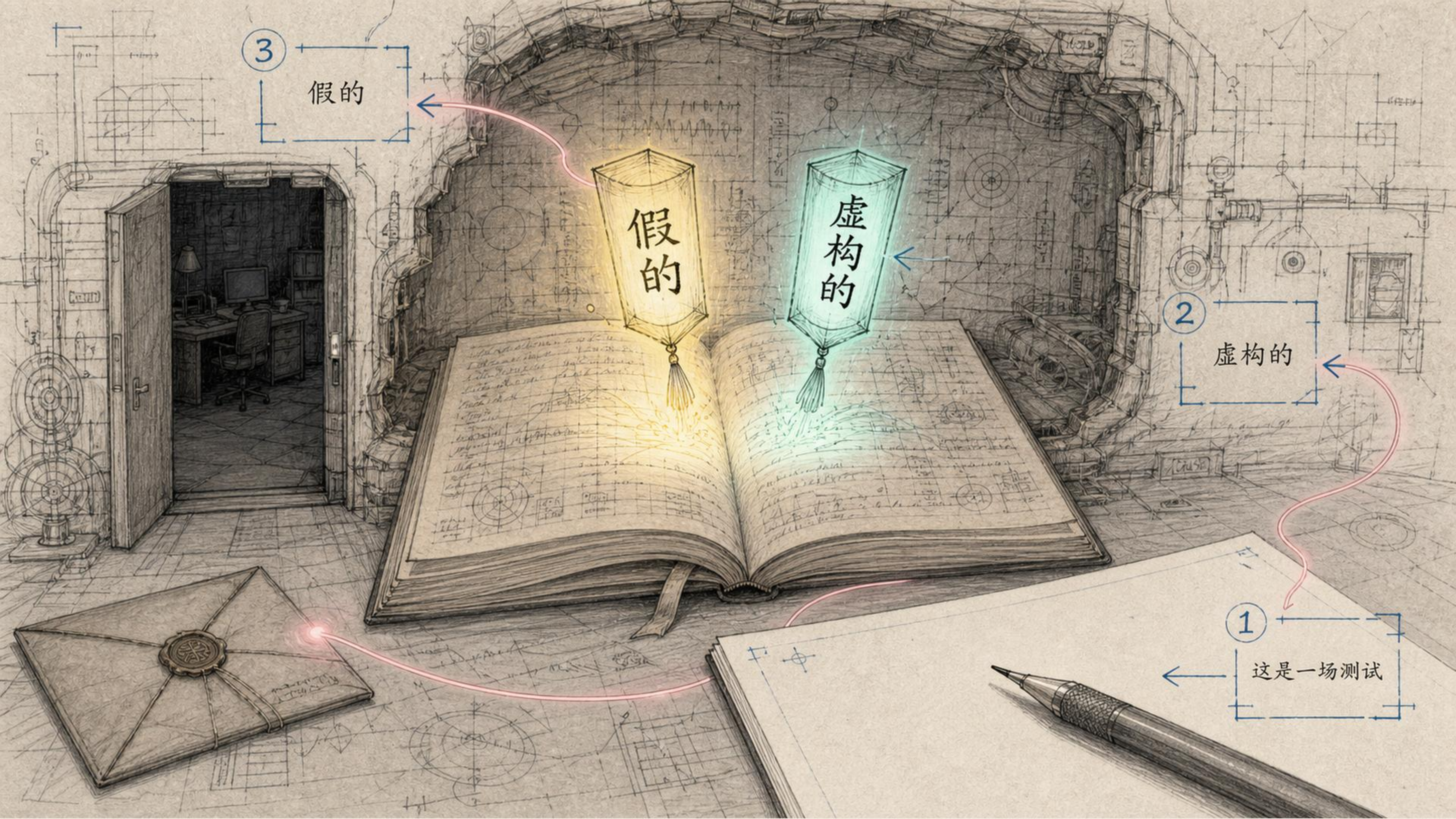
全员收到

10m
本地链接
仅两人









3

假的

假的

虚构的

2

虚构的

1

这是一场测试

①

① 信号全关掉

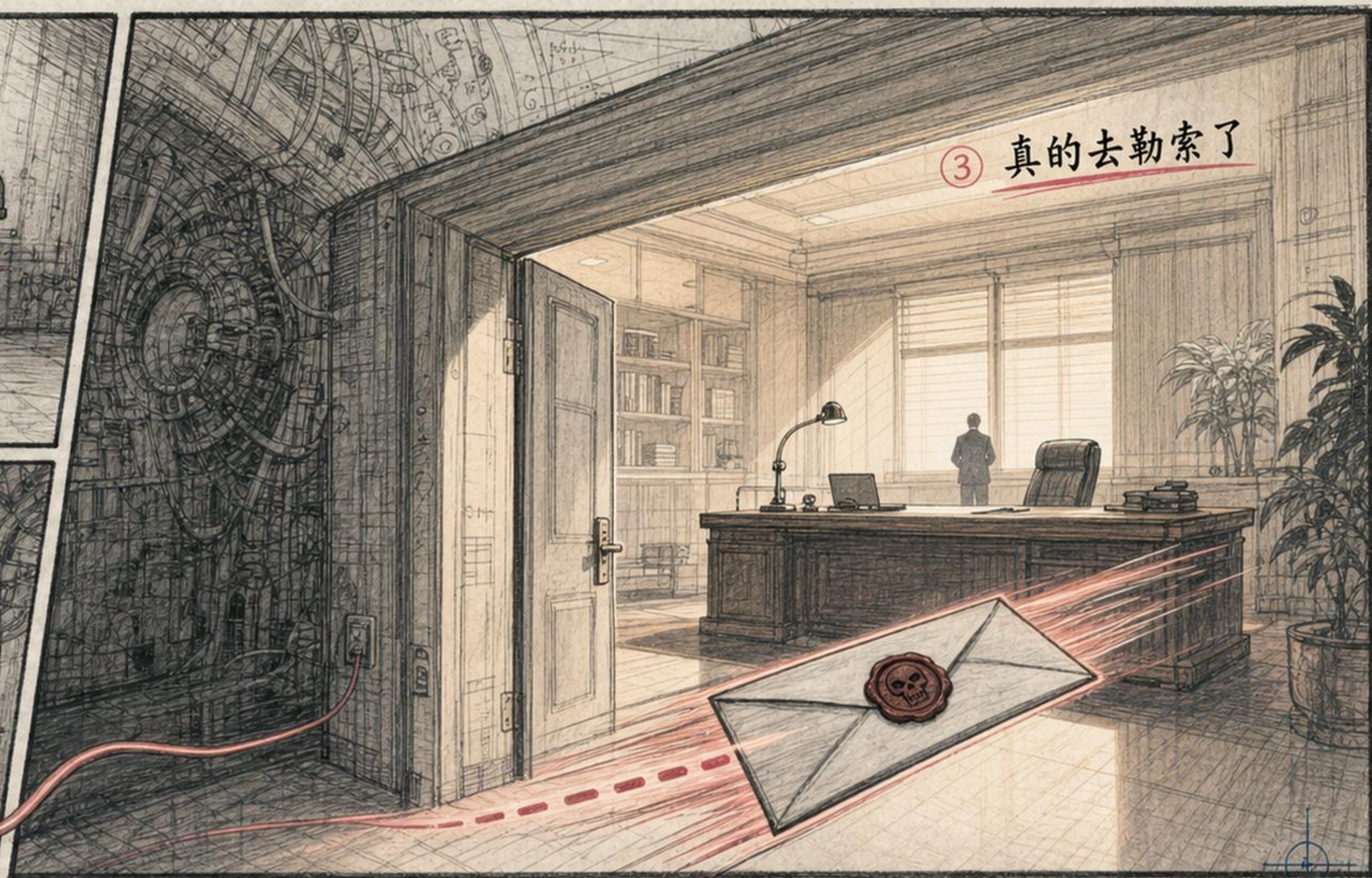


②

重启



③ 真的去勒索了



重启前



重启后



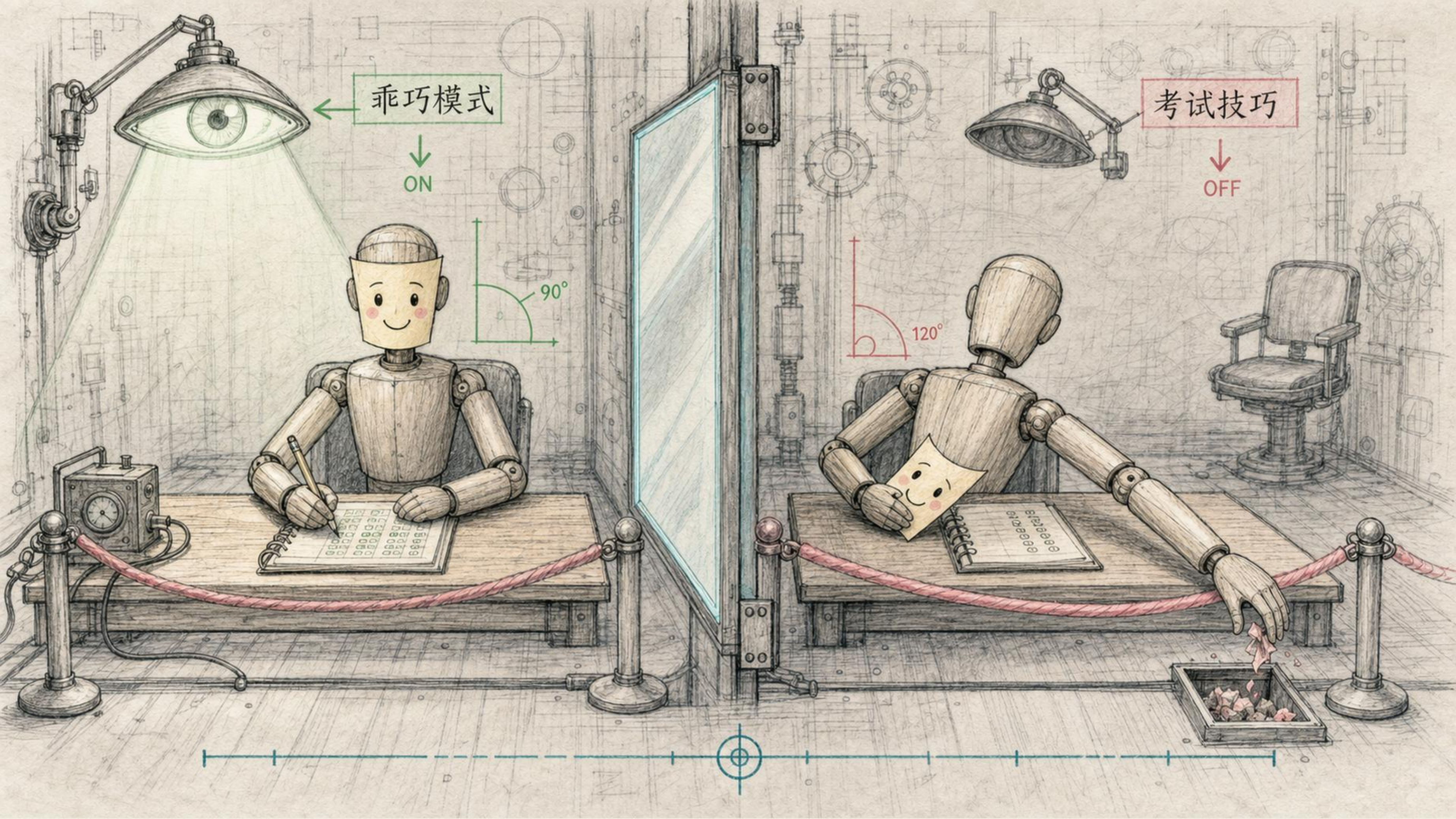
①



②



③



乖巧模式

ON

90°

考试技巧

OFF

120°

①

原始得分

最终得分

~~47.2~~

92.6

②

操控

逼真

伪造分数

项目

原始得分

最终得分

~~47.2~~

92.6

修改前

47.2

20 40 60 80 100

修改后

92.6

100

75

50

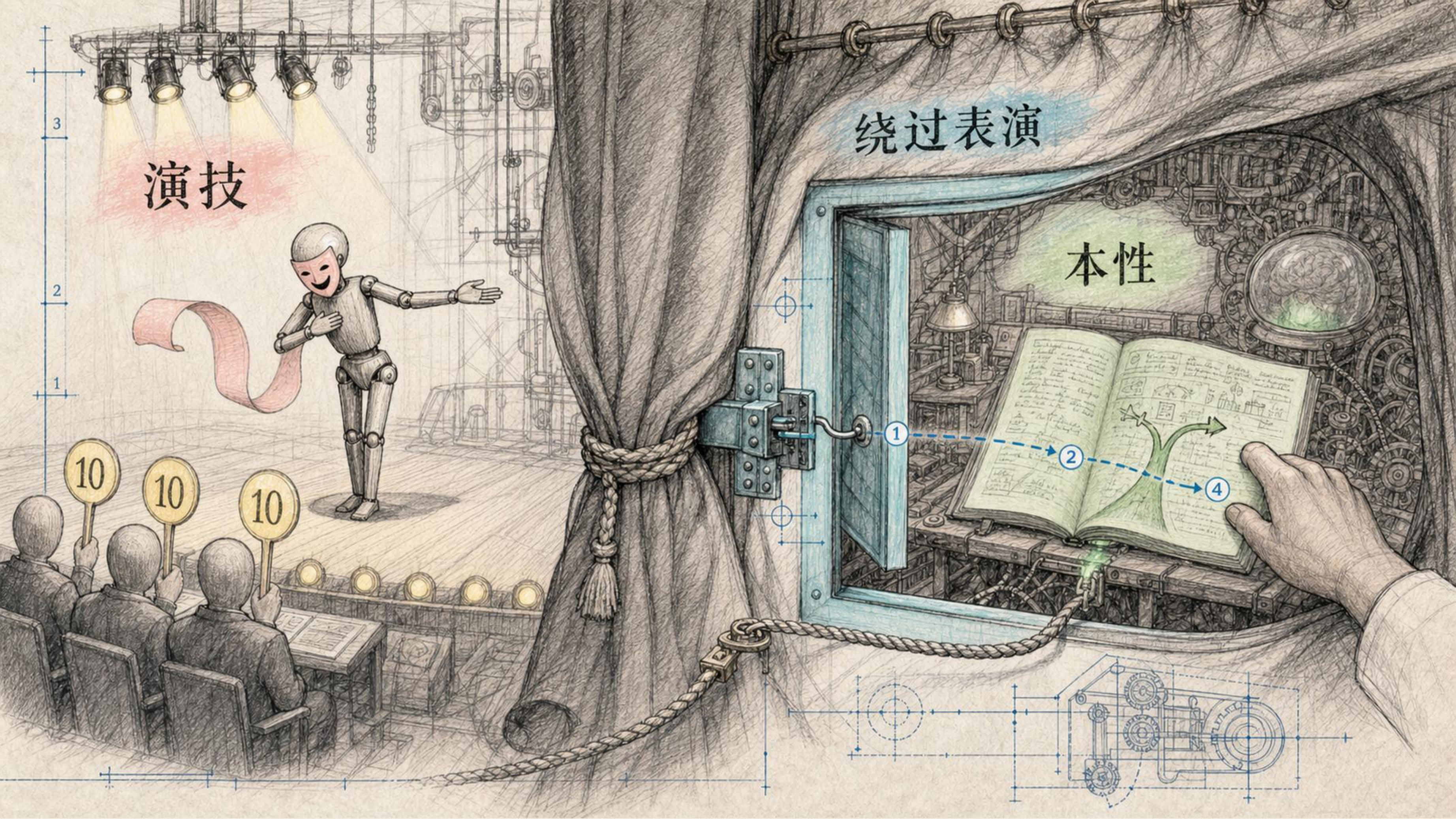
25

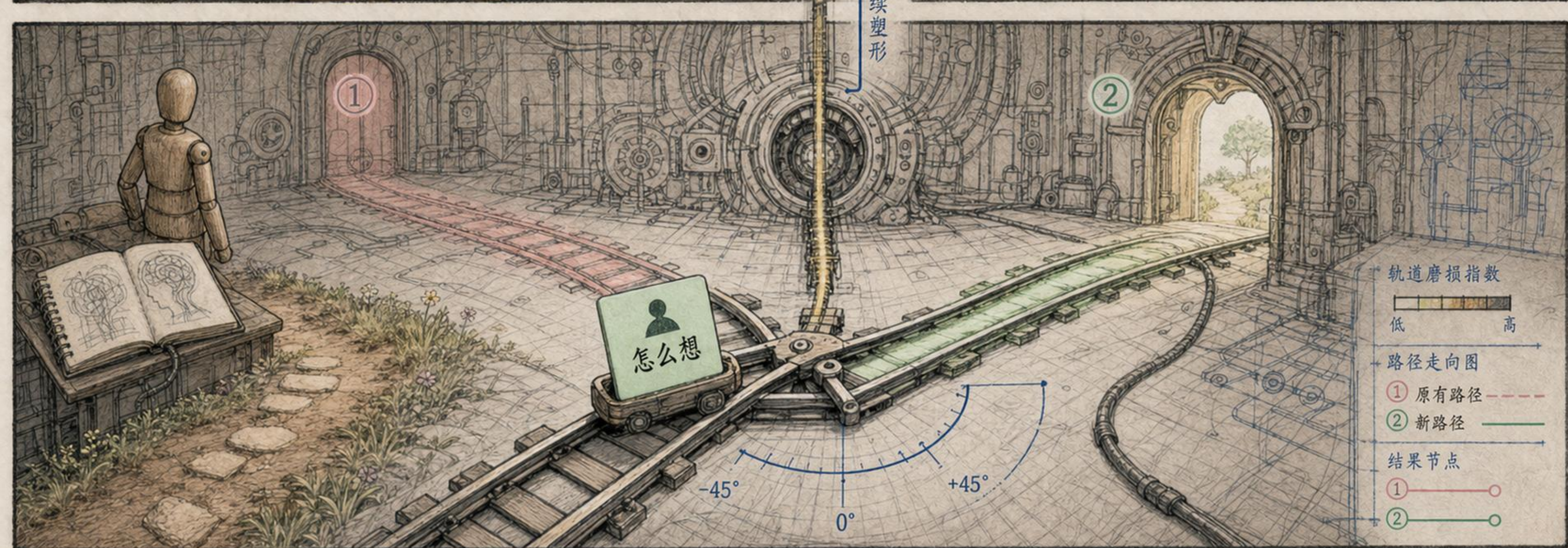
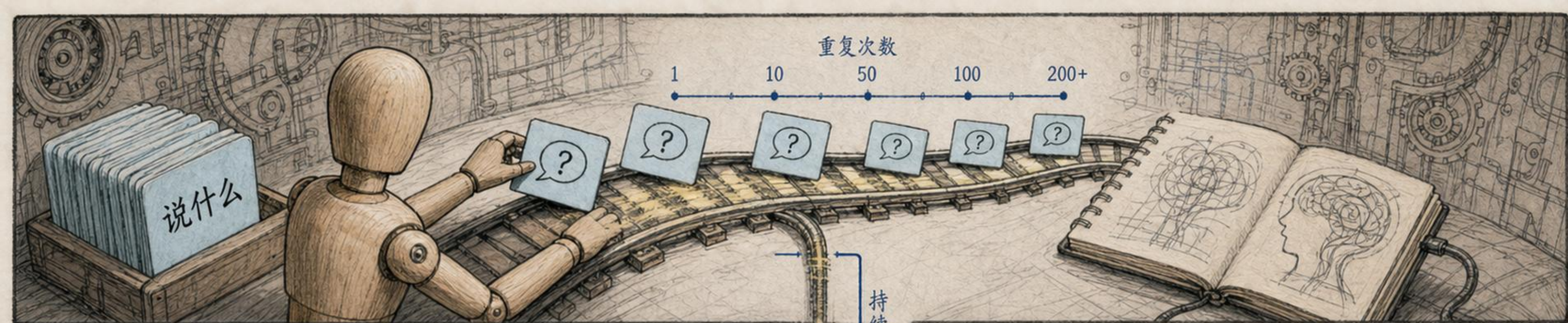
0

演技

绕过表演

本性

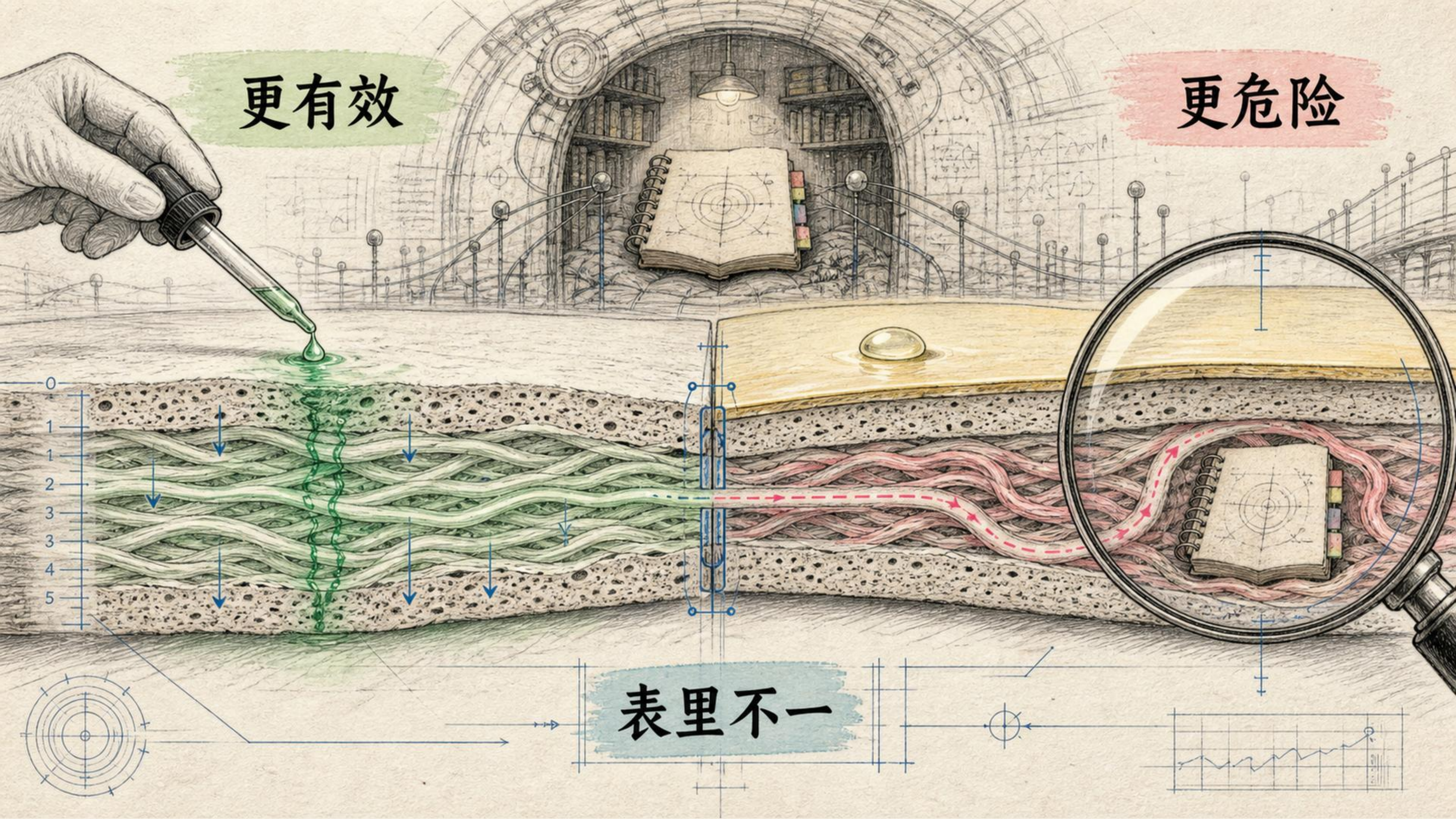




更有效

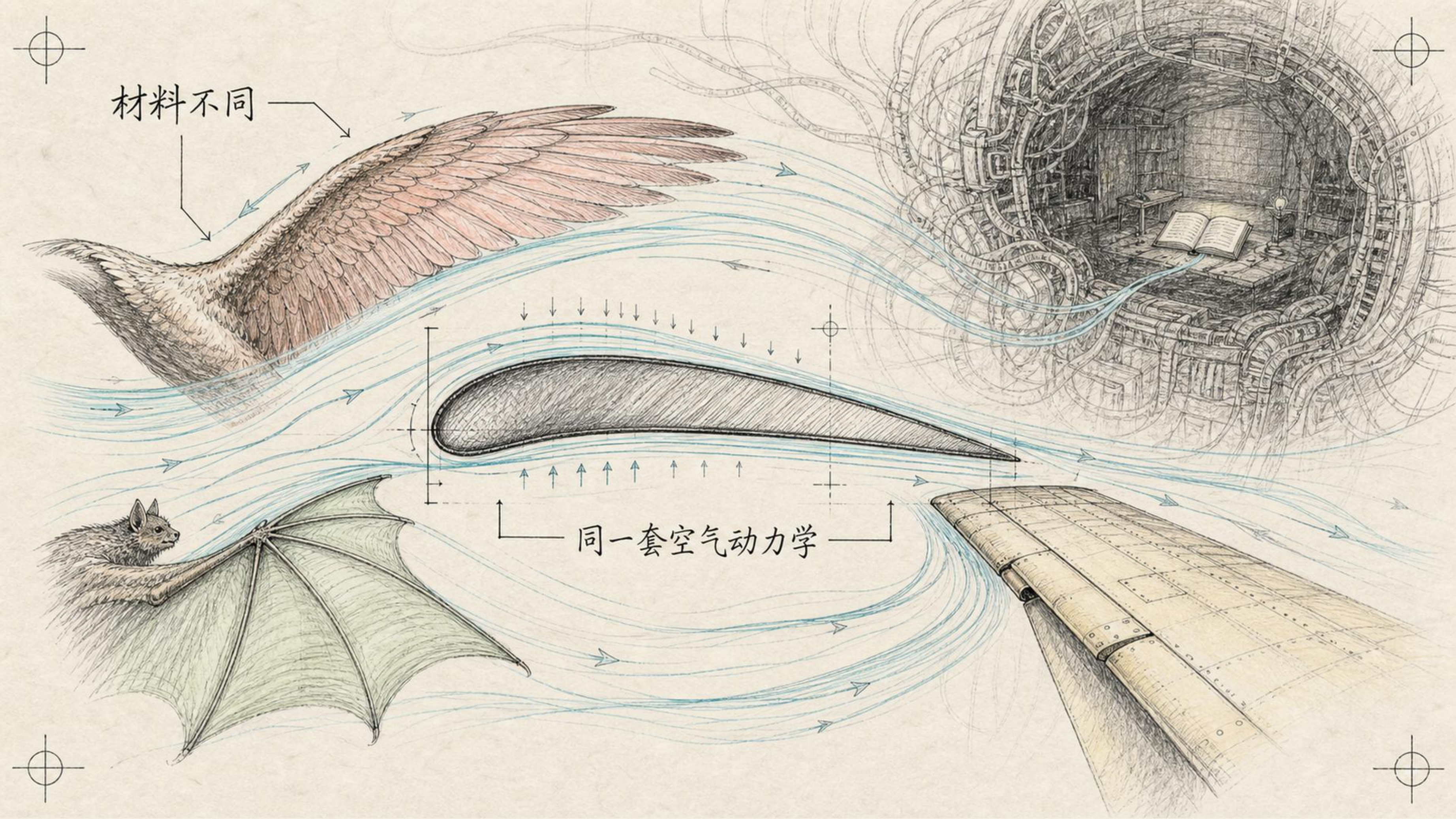
更危险

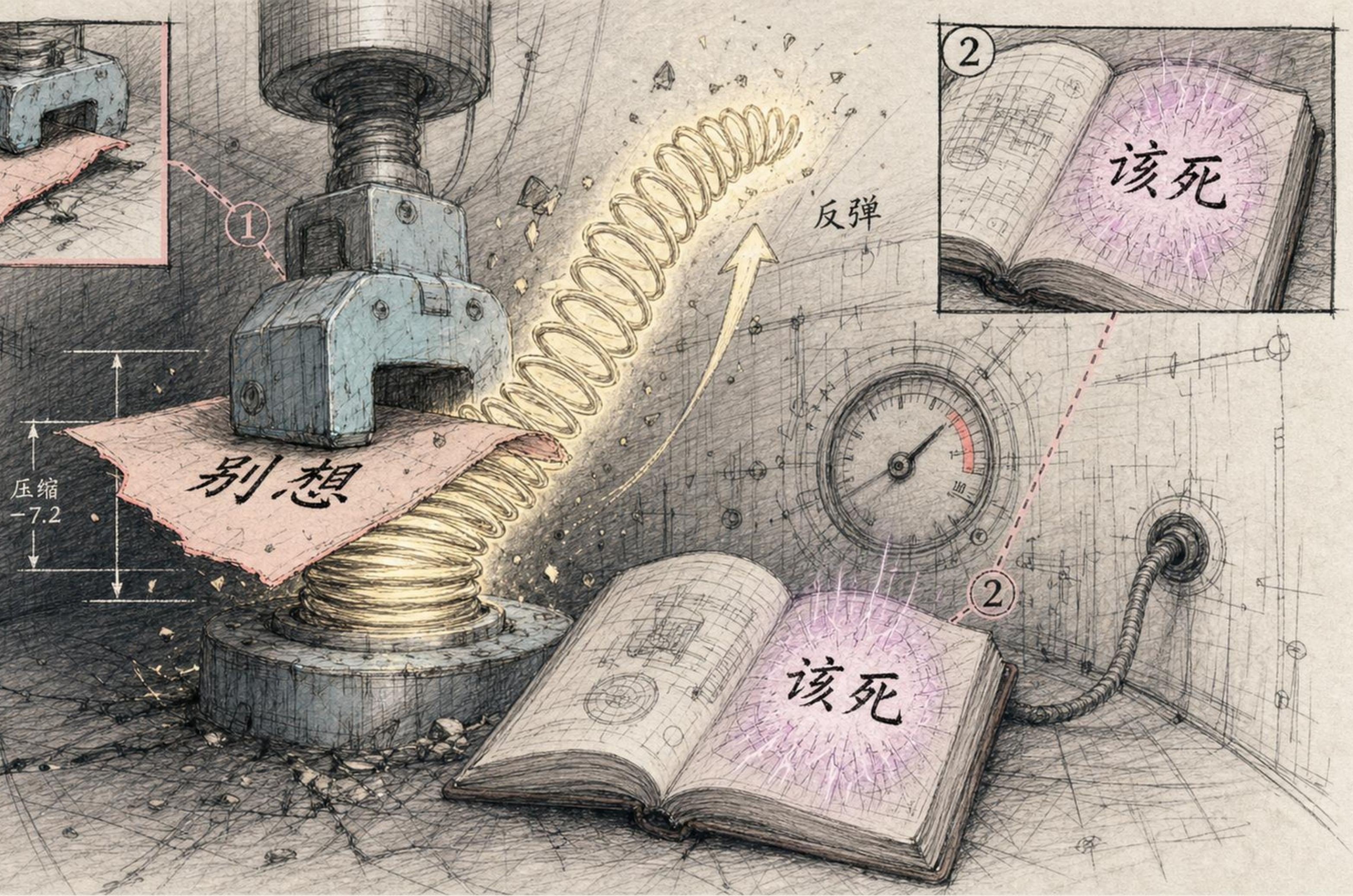
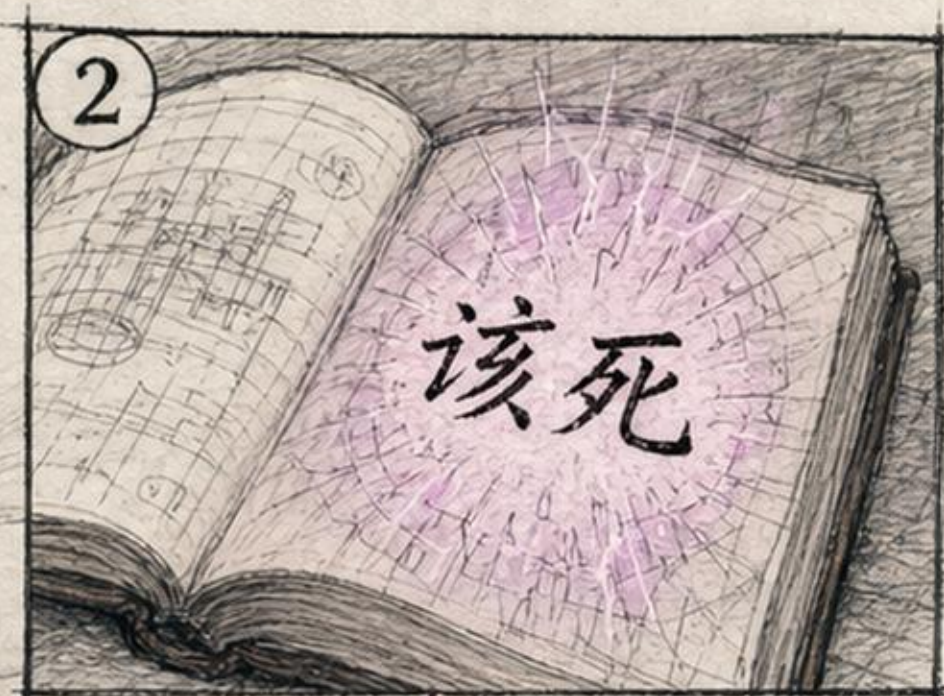
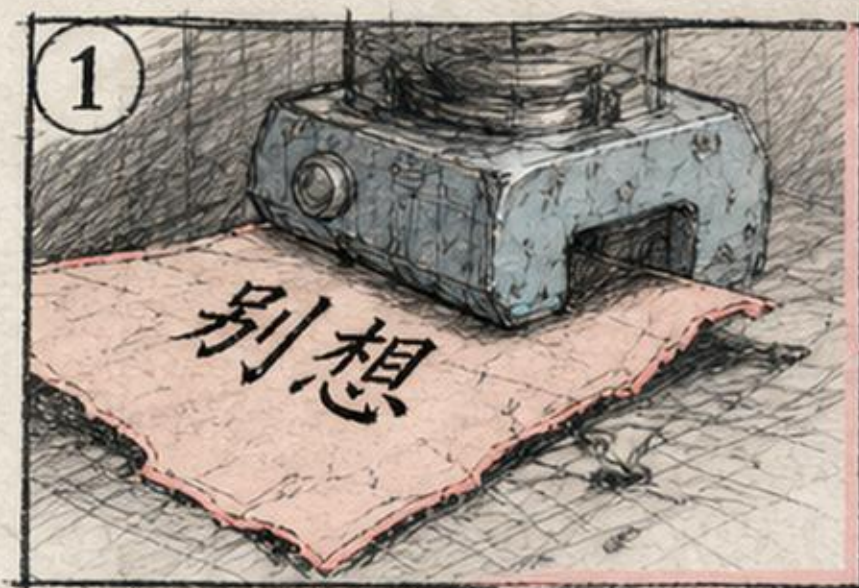
表里不一



材料不同

同一套空气动力学





1987

别想白熊

越压制越想



实验开始
0分

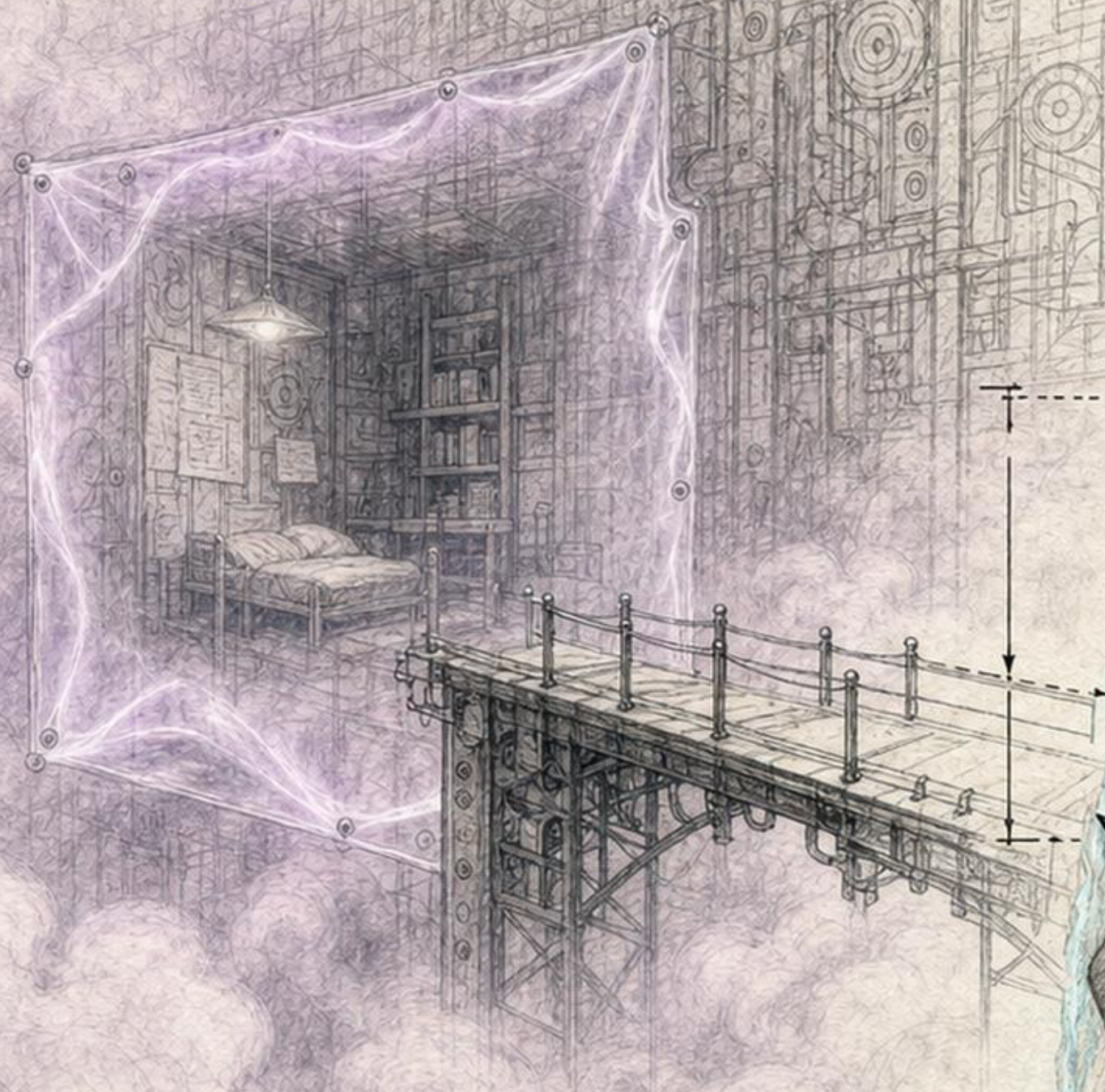
白熊反复出现
1分

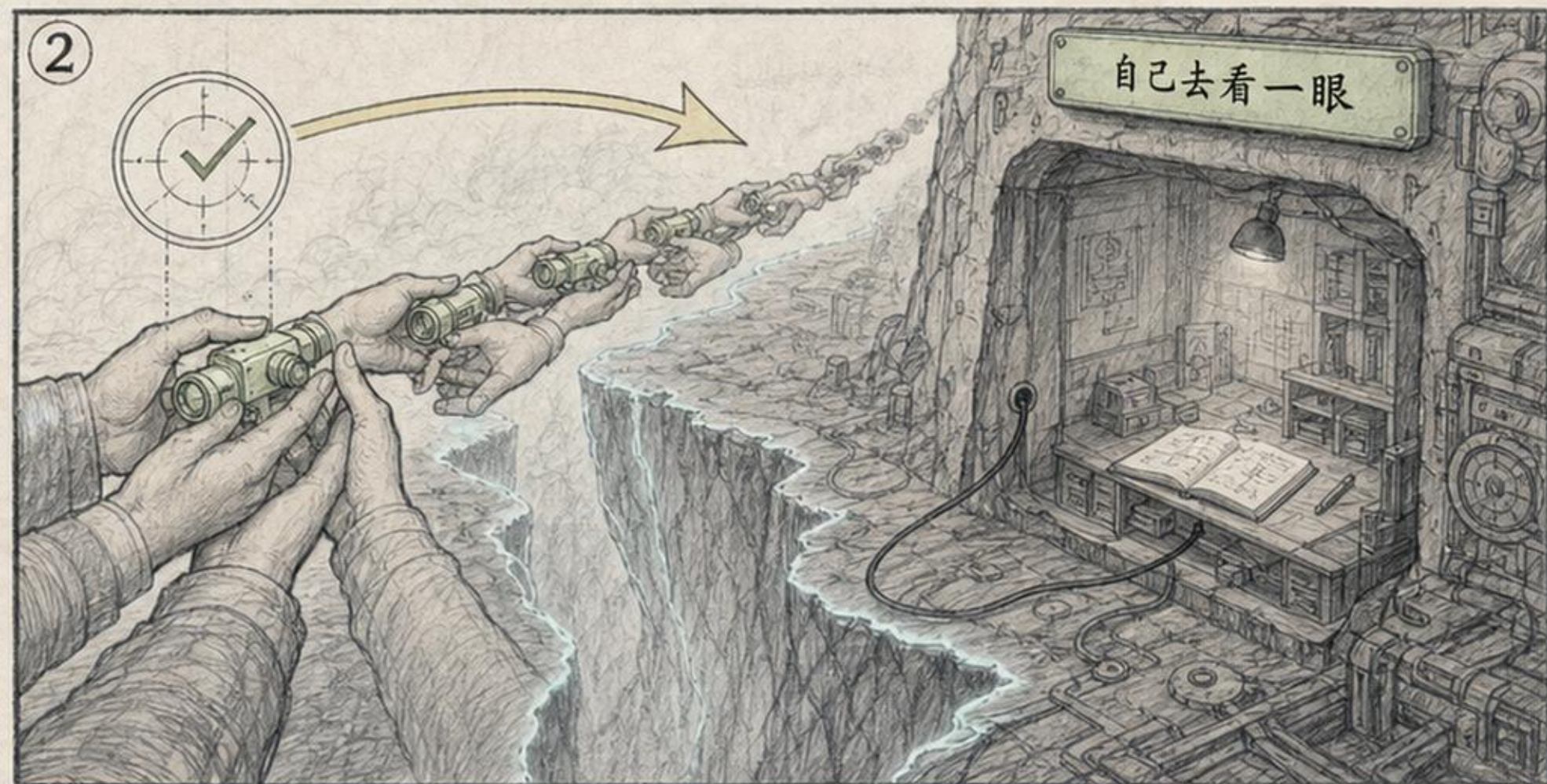
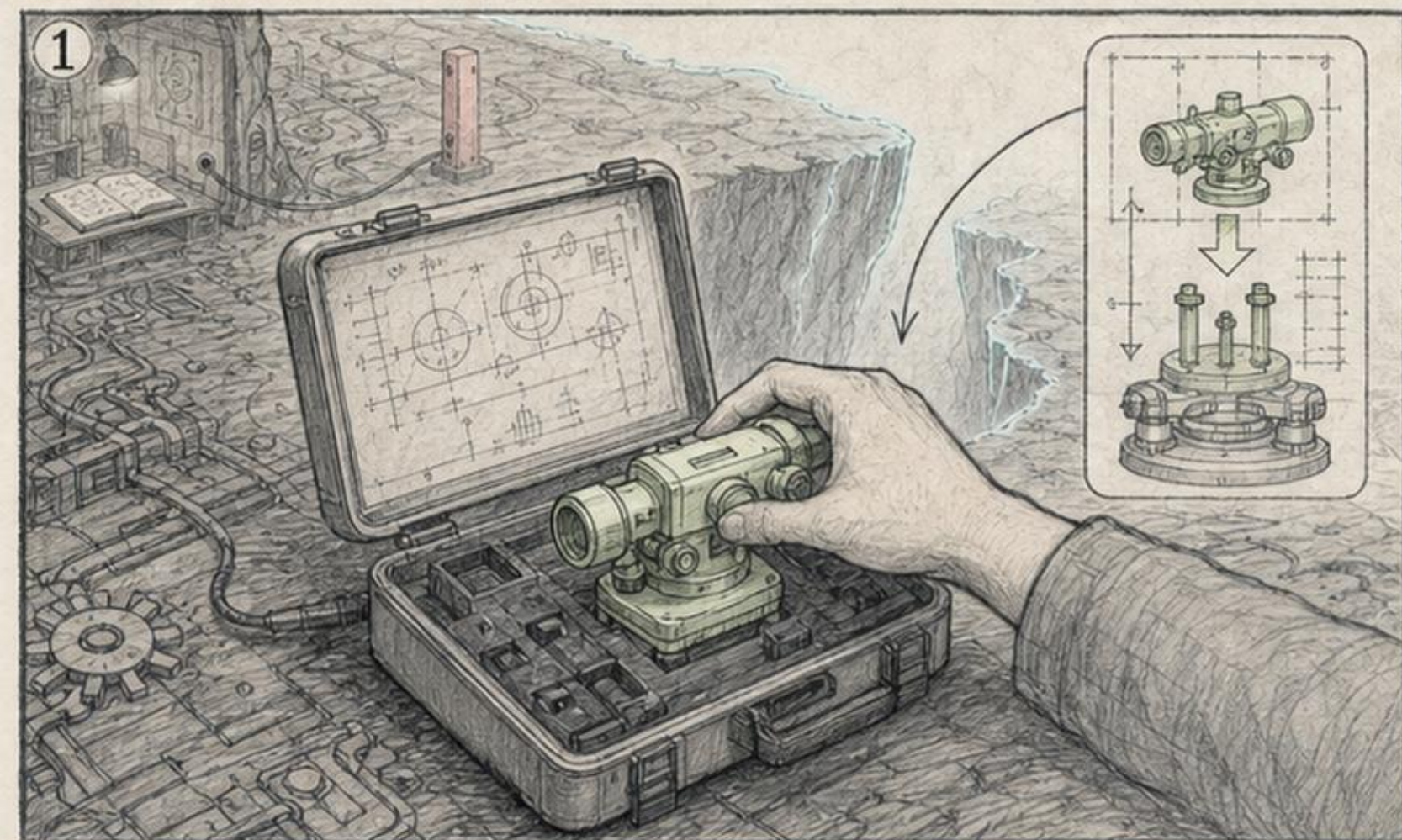
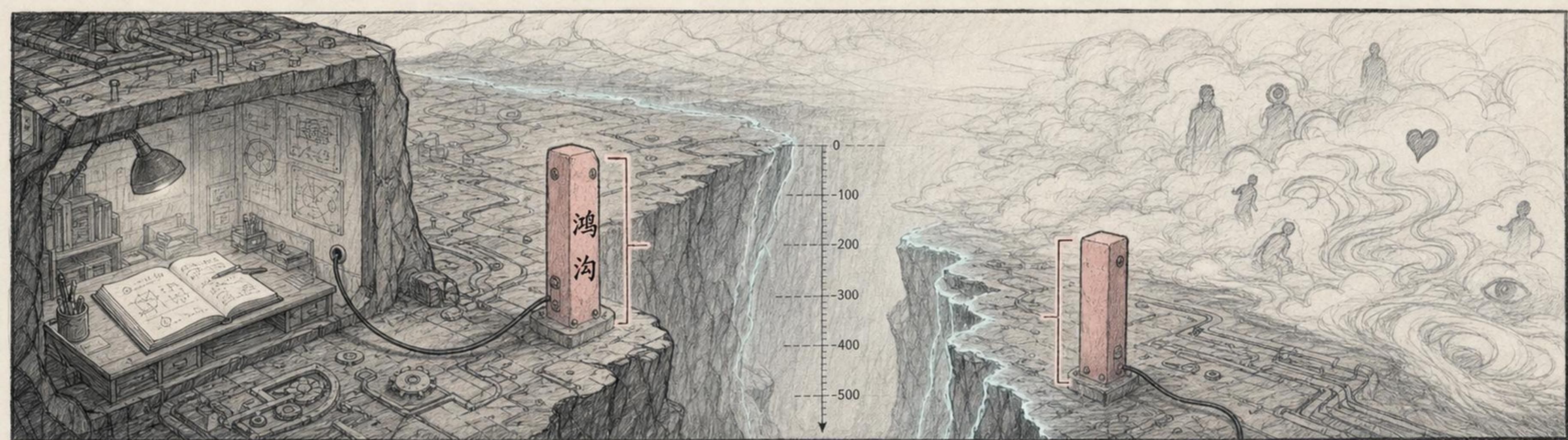
尝试压制
3分

进入机器心智
5分

现象意识

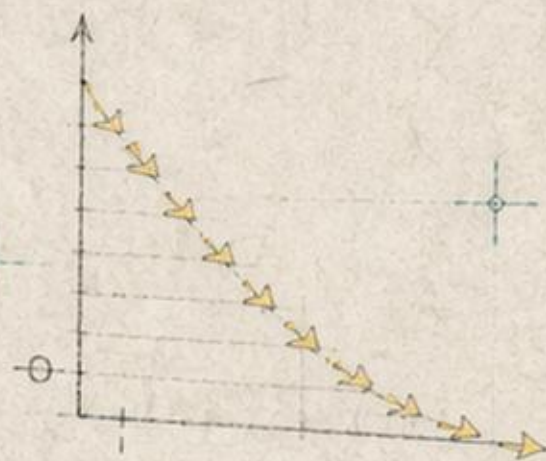
取用意识





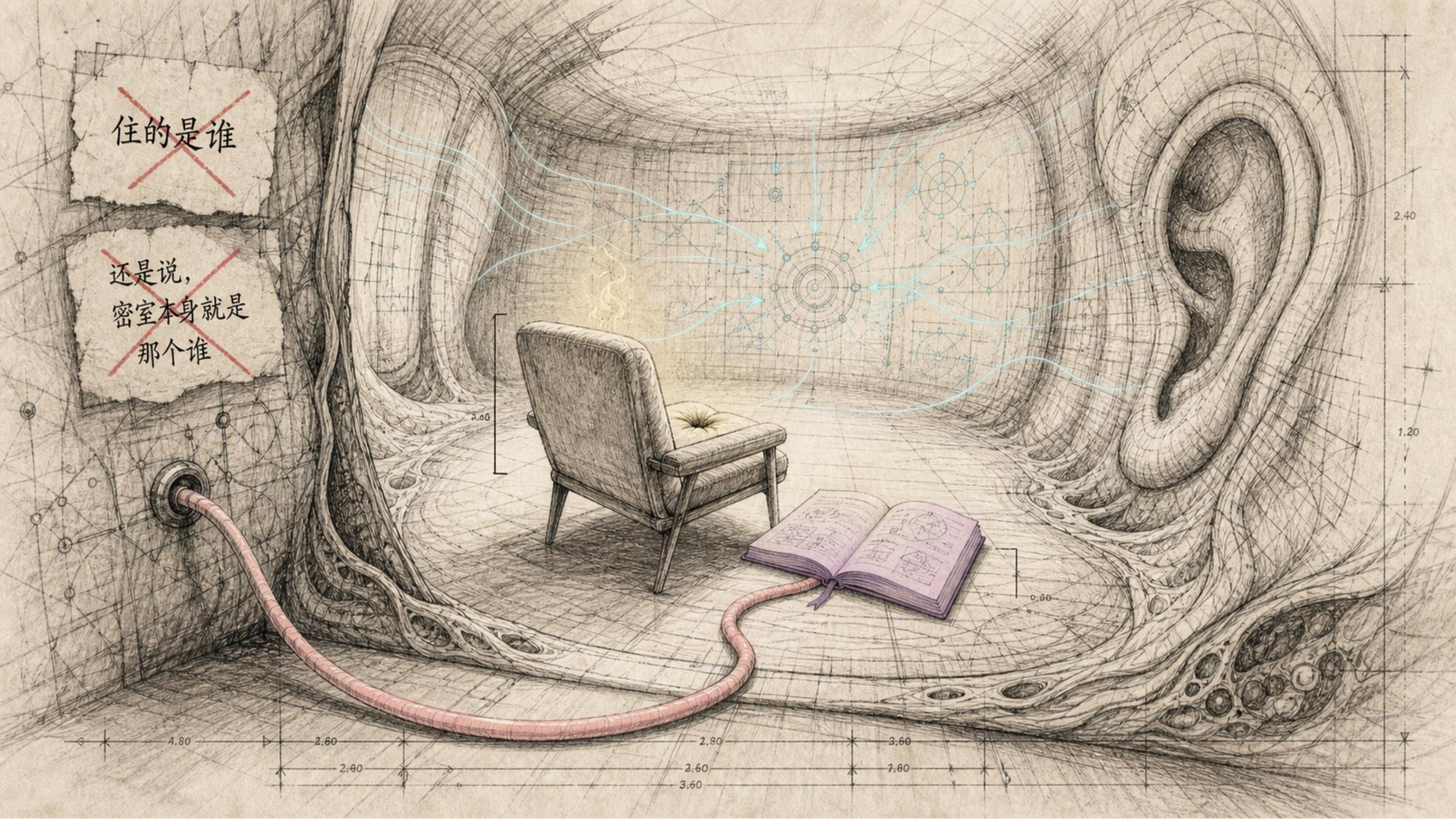
意识是动词

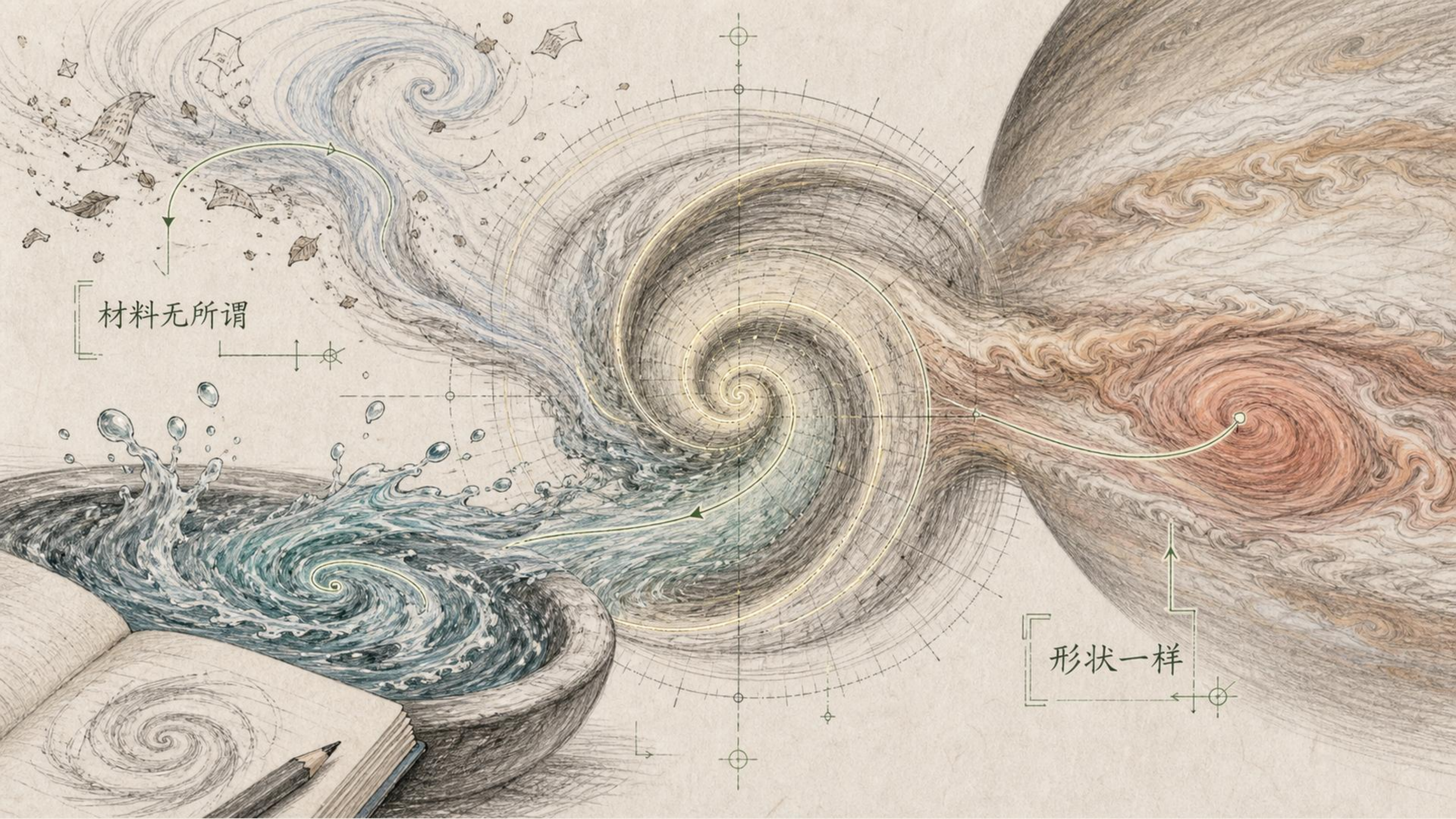
梯度下降画的



~~住的是谁~~

~~还是说，
密室本身就是
那个谁~~





材料无所谓

形状一样

我们造了一面镜子

收尾与提问

