

硅谷深夜

失眠报告

VOL.001 | 3:00AM | 内部研究披露

“我们低估了问题的复杂性，也低估了人类智能的深度。”

— 研究日志 03:17

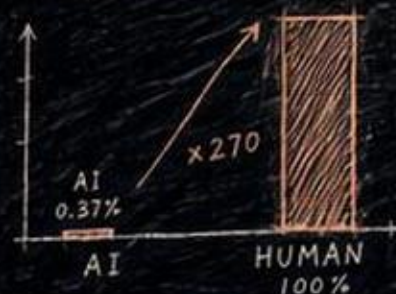
震惊发现

AI推理能力仅为人类水平的

0.37%

相差约 270 倍

REASONING GAP



关键差距维度

- 抽象推理 x314
- 因果理解 x256
- 迁移泛化 x289
- 逻辑深度 x275
- 自我反思 x392

真相令人不安，但无视它将更危险。

22:47 灵感闪现
00:11 数据验证
01:32 反复验证
02:18 结果确认

03:00
无法入睡

实验环境 ENVIRONMENT

- 128 x A100
- 2.3PB 训练数据
- 47个开源模型
- 12个月持续评估



核心结论 CONCLUSION

AI的推理能力目前仅达到人类水平的0.37%，距离真正的智能还有极其遥远的距离。

WAKE UP. BUILD DEEPER.

备注 NOTES

这不是终点，而是认知的起点。

— 研究团队
2024.05.18

人类大脑 ≠ 大模型

我们不是在缩放模型，
我们是在探索智能本质。”

Intelligence ≠ Scale

Depth > Data

残酷现实

当前最先进模型在复杂推理任务中仍然表现脆弱，容易幻觉，缺乏真正理解。

下一步？

- ☐ 超越模式匹配
- ☐ 深层因果建模
- ☐ 符号与神经融合
- ☐ 长期记忆架构
- ☐ 自我进化机制

警告

如果不能突破推理瓶颈，通用人工智能将永远无法到达。

SLEEPLESS IN SILICON VALLEY



INSIGHT | REVELATION | REALITY

AI刷榜狂欢

(一片凯歌)



新SOTA!
又破纪录!

GPT-4.5 霸榜!
MMLU 92.3% ↑

排行榜
天天刷新!

Claude 3.7
新高分!
HumanEval 95.2% ↑

Gemini 1.5 Pro 登顶!
GPQA 88.7% ↑

Llama 3.1 405B
多项榜单第一!

SWE-bench Verified
72.1% ↑

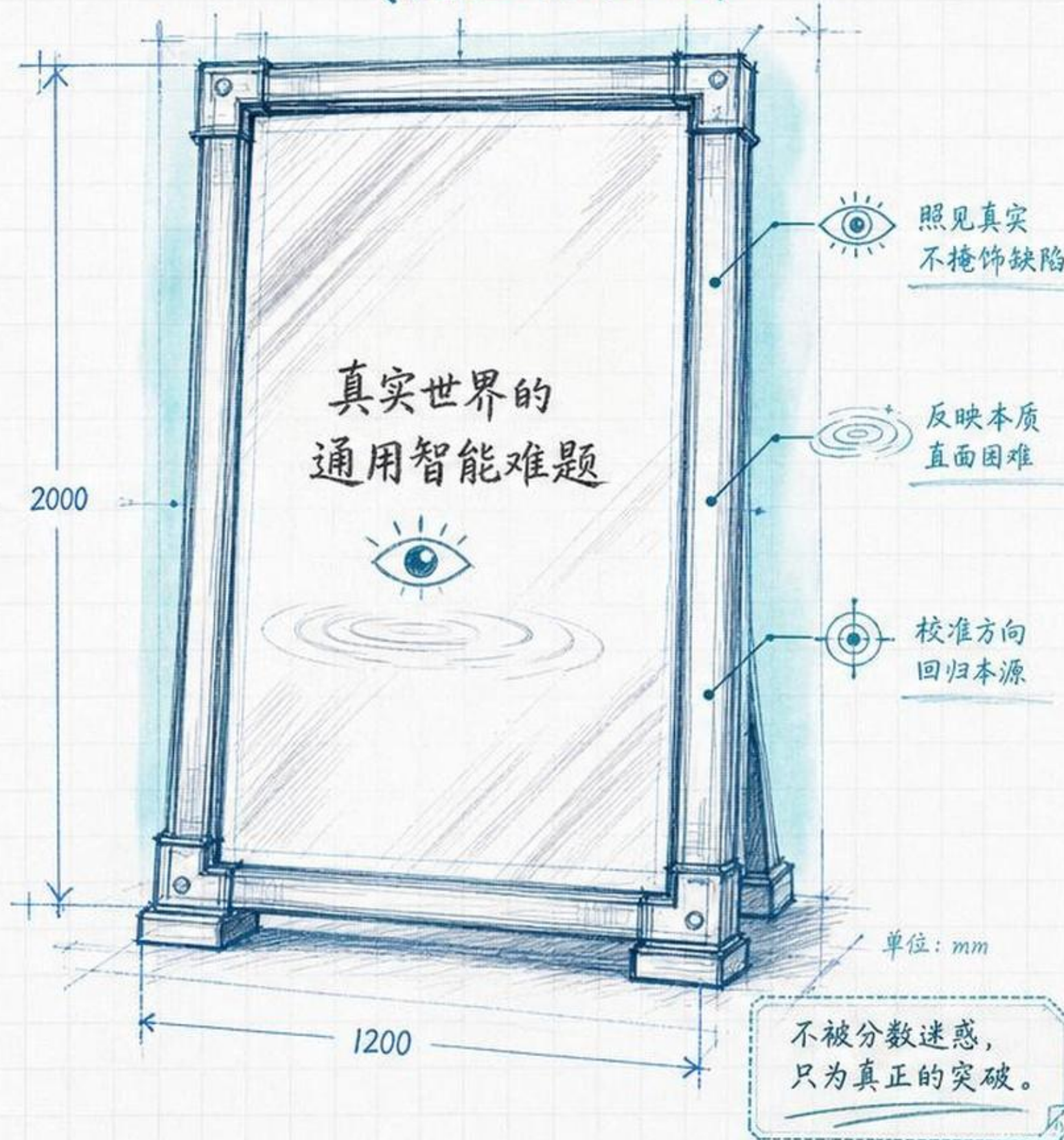
代码能力
大突破!

LiveCodeBench
78.9% ↑

ALL WIN!
未来已来!

VS.

AI界最重要的一面“镜子” (ARC-AGI-3)



谷歌九年

- 2015-2024
- 深度学习研究员
- 长期专注前沿研究
- 参与多个核心项目

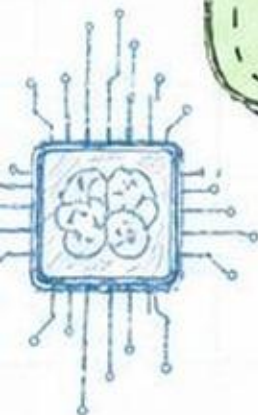


$$L = \sum \ell(y_i, \hat{y}_i)$$



顶尖研究者

- 深度学习与通用智能探索者
- 以独立思考和工程能力著称
- 在学术与工业界影响深远



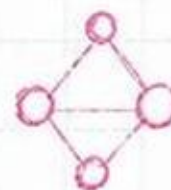
弗朗索瓦·肖莱

ARC-AGI-3 的创造者

“我相信，理解智能的最佳方式，
就是构建能发现并度量它的工具。”

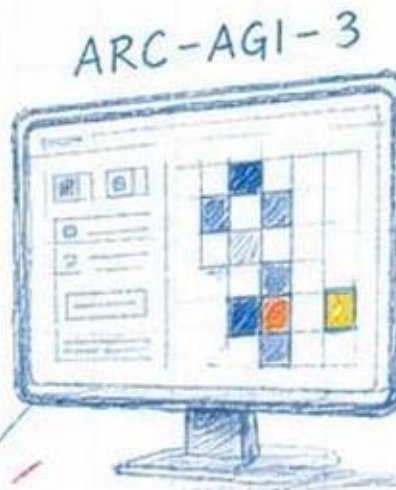
Keras作者

- Keras 的作者
- 简洁、优雅、易用的深度学习框架
- 被全球开发者广泛使用



造AI工具的人

- 专注构建能推动智能边界拓展的工具
- 追求可验证的智能进步
- ARC 系列的创造者



《论智能的度量》 (2019)

摘要

当前主流的智能度量方法存在根本性缺陷，它们无法全面反映智能的本质与多样性。我们需要建立新的度量框架，超越任务导向的局限，走向更本质、更通用的智能理解与评估。

作者：朱松纯
北京通用人工智能研究院
2019.7

我们一直在用错误的方式衡量AI的智能

准确率

0 20 40 60 80 100

基准测试

20 30 40 50

排行榜

$$I = \sum_{i=1}^n w_i x_i$$

$$w_i = ?$$

$$x_i = ?$$

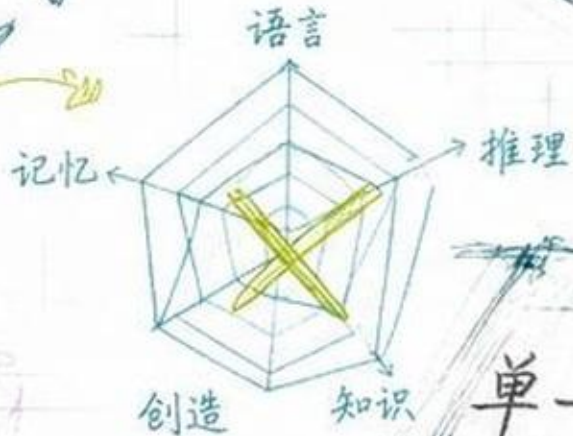


1/2 人类对齐度

2 3 4 5

综合得分

20 40 60 80 100



单一指标

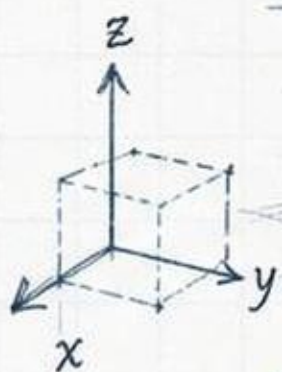
晶体智能

记忆与调用

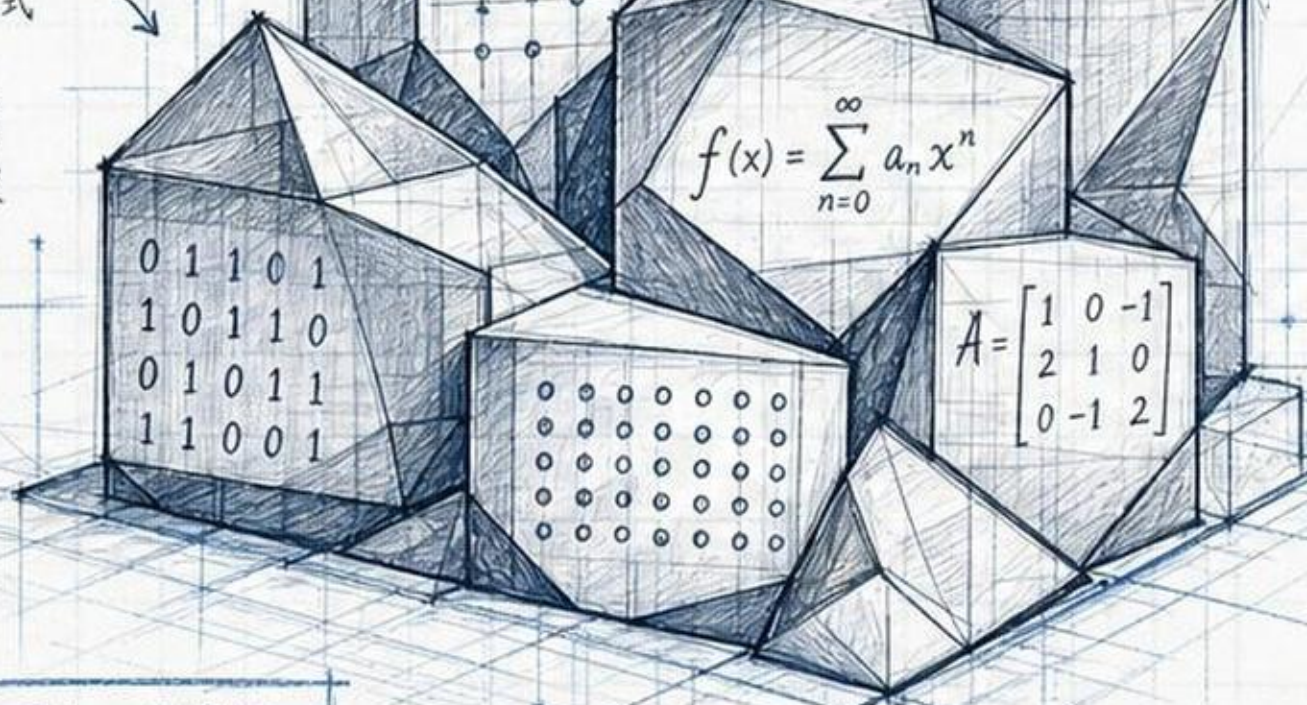
知识储存
结构化信息

经验沉淀
可复用模式

逻辑推演
规律与框架



已知世界的
结构与秩序



$$E = mc^2$$

$$a^2 + b^2 = c^2$$



$$f(x) = \sum_{n=0}^{\infty} a_n x^n$$

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 1 & 0 \\ 0 & -1 & 2 \end{bmatrix}$$

VS

流体智能

未知环境的创想

💡 灵活应变
适应新情境

🌀 模式重组
跨界连接

🌱 创意生成
探索可能性

❓ 拥抱不确定
从未知出发

未知世界的
探索与创造



流体智能测试



无语言



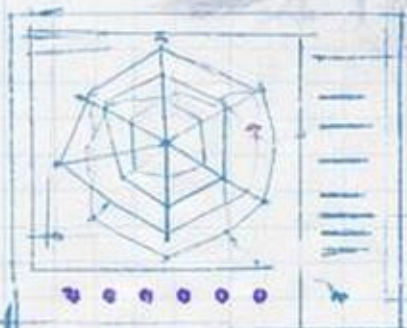
无信号



无导航



完全陌生



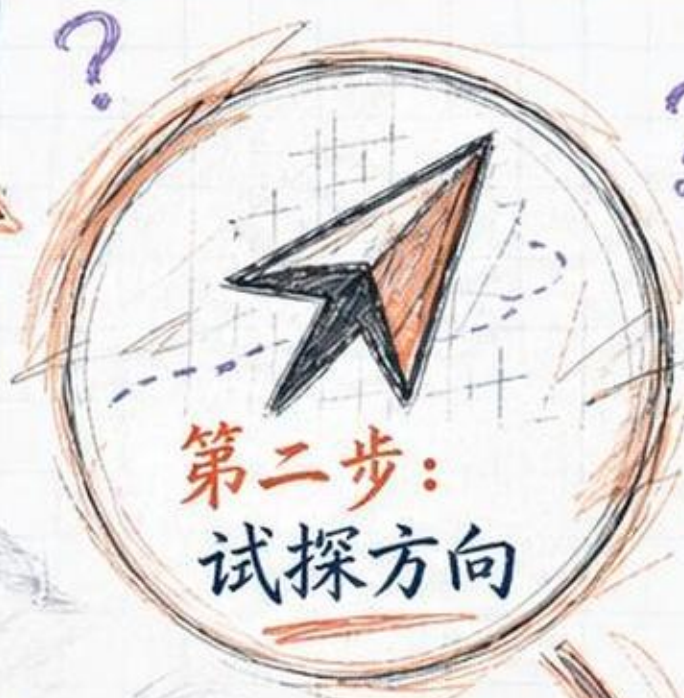
探索思维循环

- ✓ 不确定性
- ✓ 快速决策
- ✓ 持续迭代
- ✓ 适应变化



第一步：
观察感知

流体智能



第二步：
试探方向

未知环境
动态变化



第三步：
获取反馈



关键能力

- ✓ 观察力
- ✓ 判断力
- ✓ 适应力
- ✓ 学习力



第四步：
调整路线

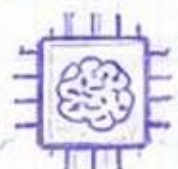
|| 即时感知 · 试探 · 反馈 · 调整



AI 是结晶化的智能

当前AI本质:

- ✓ 存储海量知识
- ✓ 静态不变
- ✓ 模式匹配

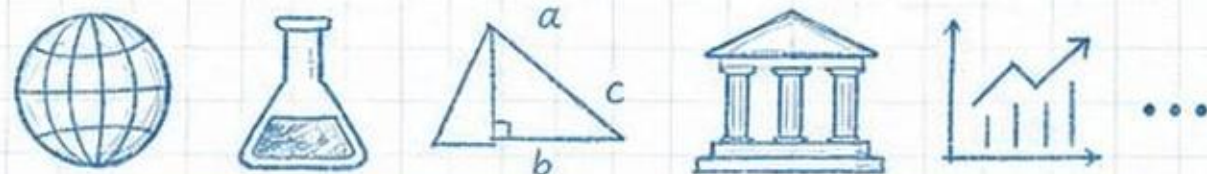


静态切片

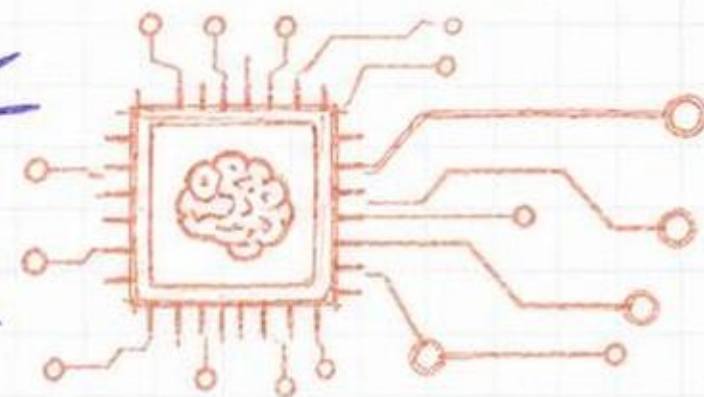
人类所有公开文本



庞大知识储备



记忆力 + 模式匹配



真正智能定义

面对前所未有的情境



面对前所未有的情境



极速搞明白效率



极速搞明白效率



面对前所未有的情境



面对前所未有的情境



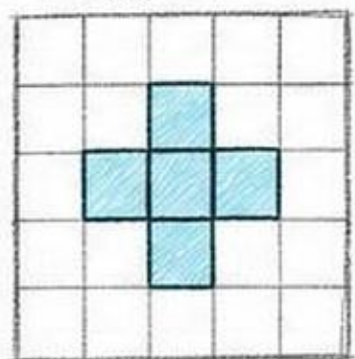
“不是知道多少，而是多快搞明白”



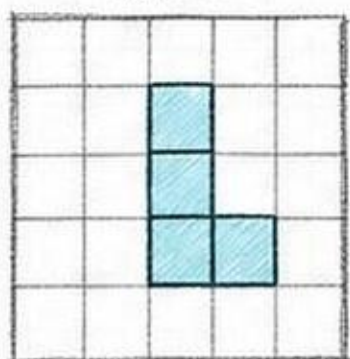
ARC-AGI-1 规律推断测试

① 示例输入

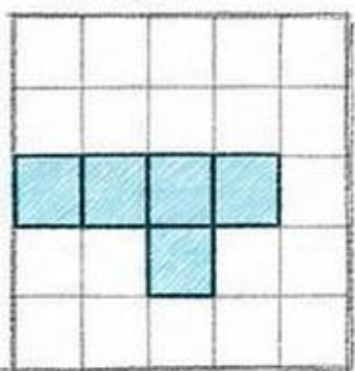
输入



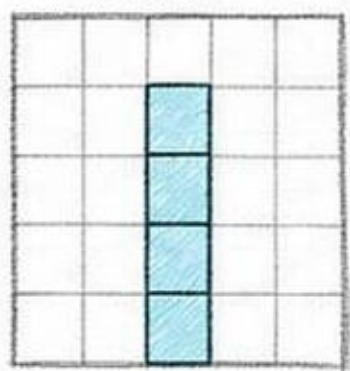
输出



输入



输出



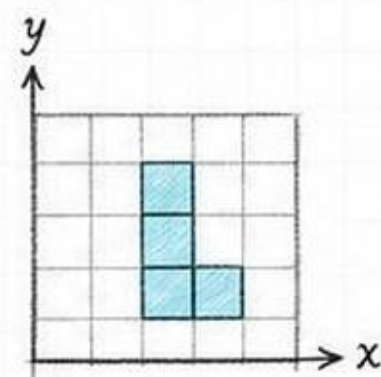
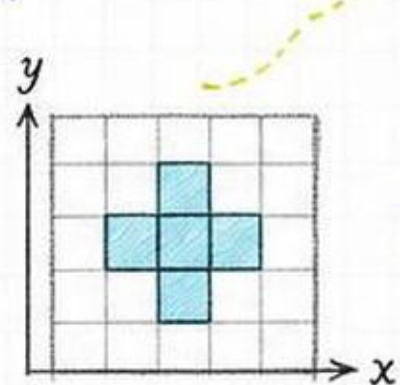
② 隐藏规律推断



观察变化: 形状不变

位置移动: $(x, y) \rightarrow (y, x)$

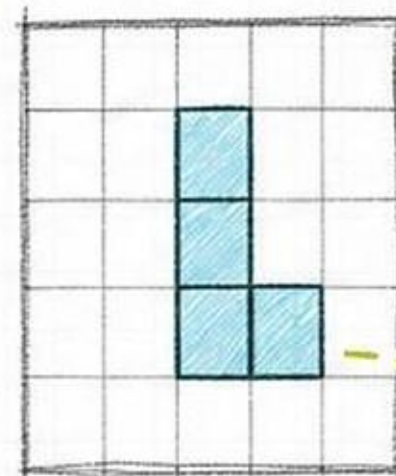
规律假设: 坐标交换 (转置) ☆



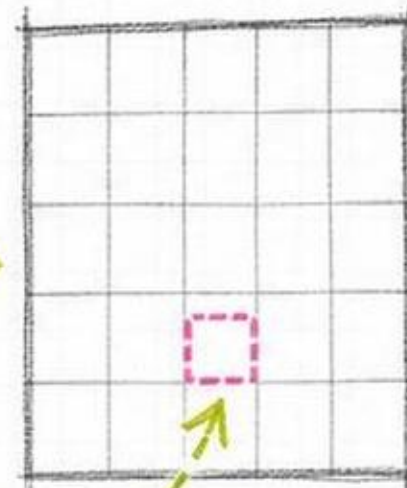
$$T(x, y) = (y, x)$$

③ 未知输入预测

未知输入



预测输出



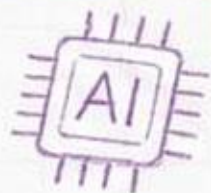
应用规则:

$(x, y) \rightarrow (y, x)$ ☆

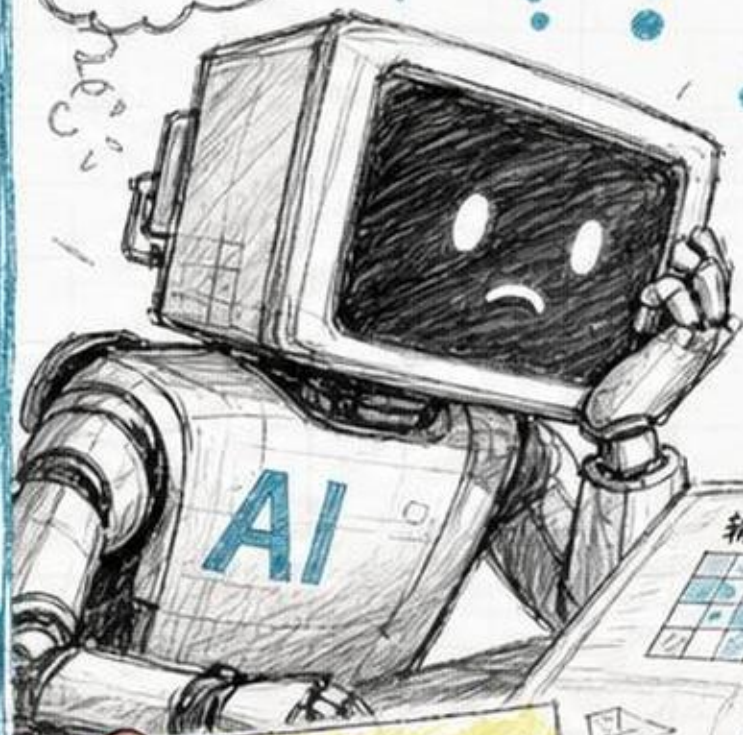
阶段一：早期AI的挣扎与行业围剿

Phase 1-1

早期AI



规律在哪...
我怎么想不到...



得分区间



0%~30%

输入

输出

独一无二
现场悟

- 需要跨越抽象层级
- 无法依赖训练记忆

难以泛化

Phase 1-2

科技巨头集火攻克

全行业资源
全面压上!



大规模算力集群

先进模型与算法

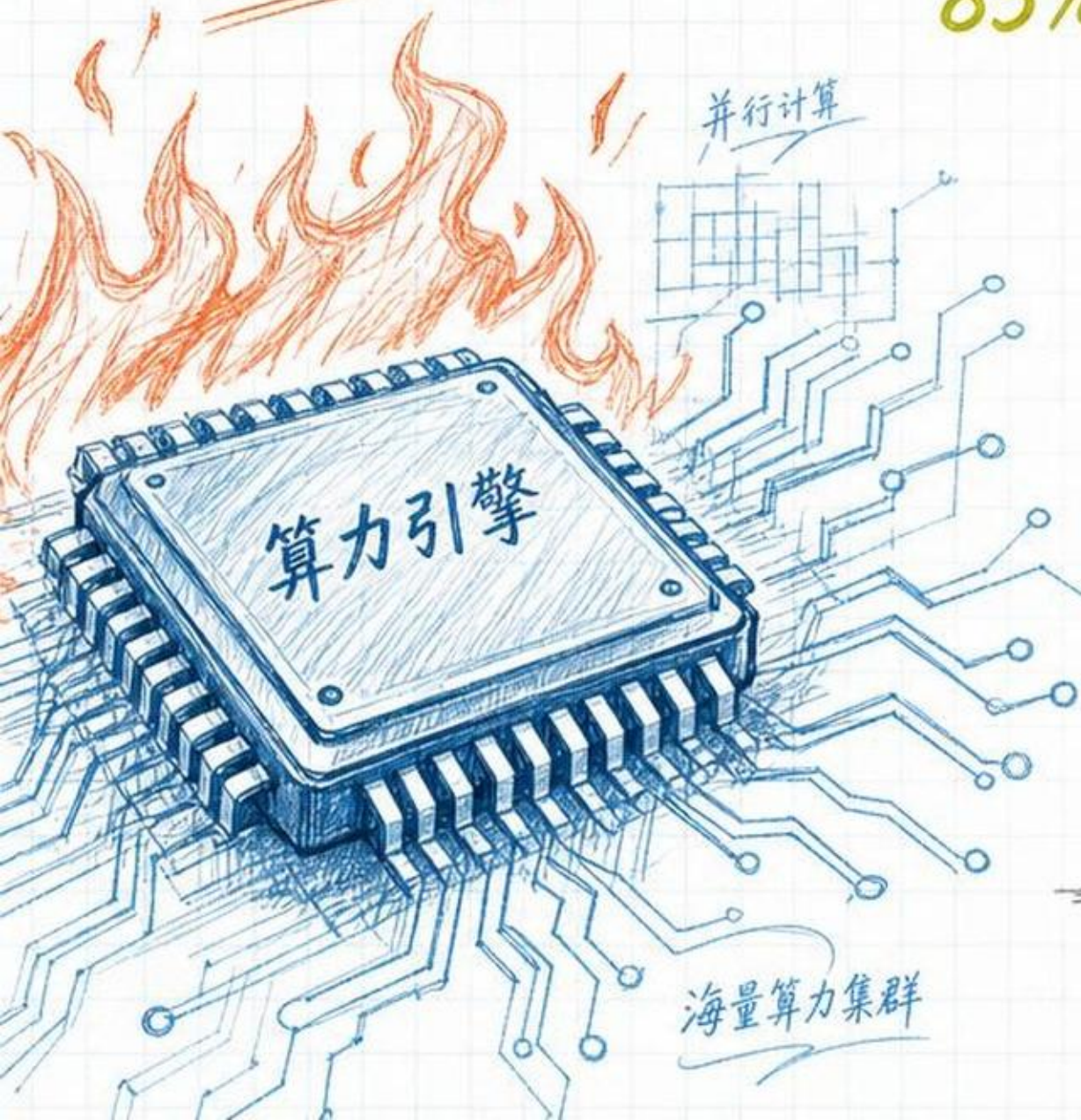
海量数据支撑

只为攻克唯一难题

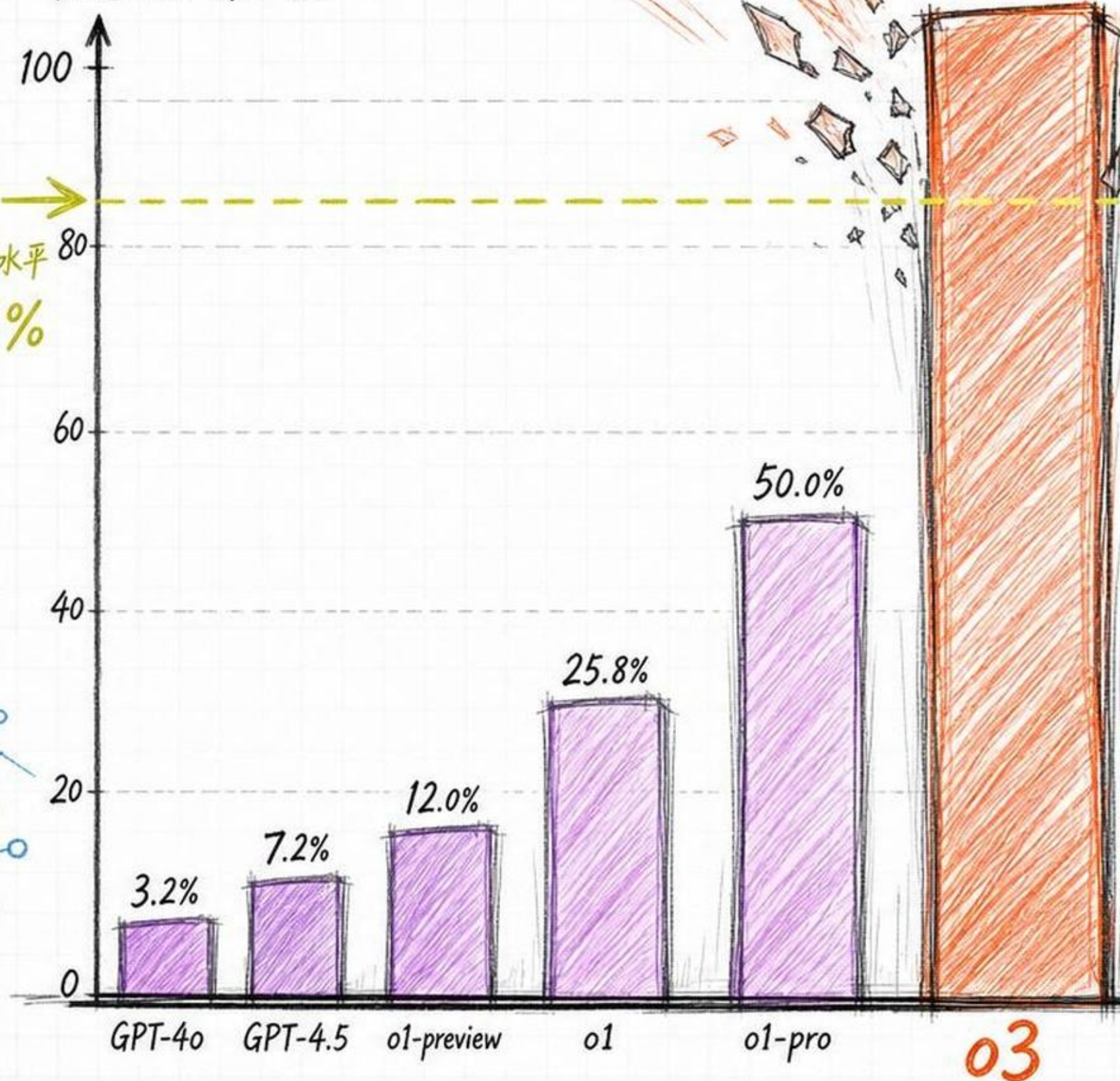
算力爆炸式突破

突破不可能，定义新极限！

人类平均水平
85%



ARC-AGI 得分 (%)



计算成本：
天文数字

172倍算力

关键突破因素

- ✓ 海量并行计算
- ✓ 超长思考时间
- ✓ 强化学习优化

超越人类新极限
开启AGI新纪元！



AI 解题成本

\$3,000 ~ \$30,000 / 题

- ☑ 高算力
- ☑ 长时间
- ☑ 大模型 !!!
- ☑ 高成本

\$30,000

\$30K

\$15,000

\$8,000

\$3,000

VS

人类解题成本

\$5 / 题 (一顿午饭的能量)



一顿午饭
的能量

简单思考
高效解决

这叫突破吗?

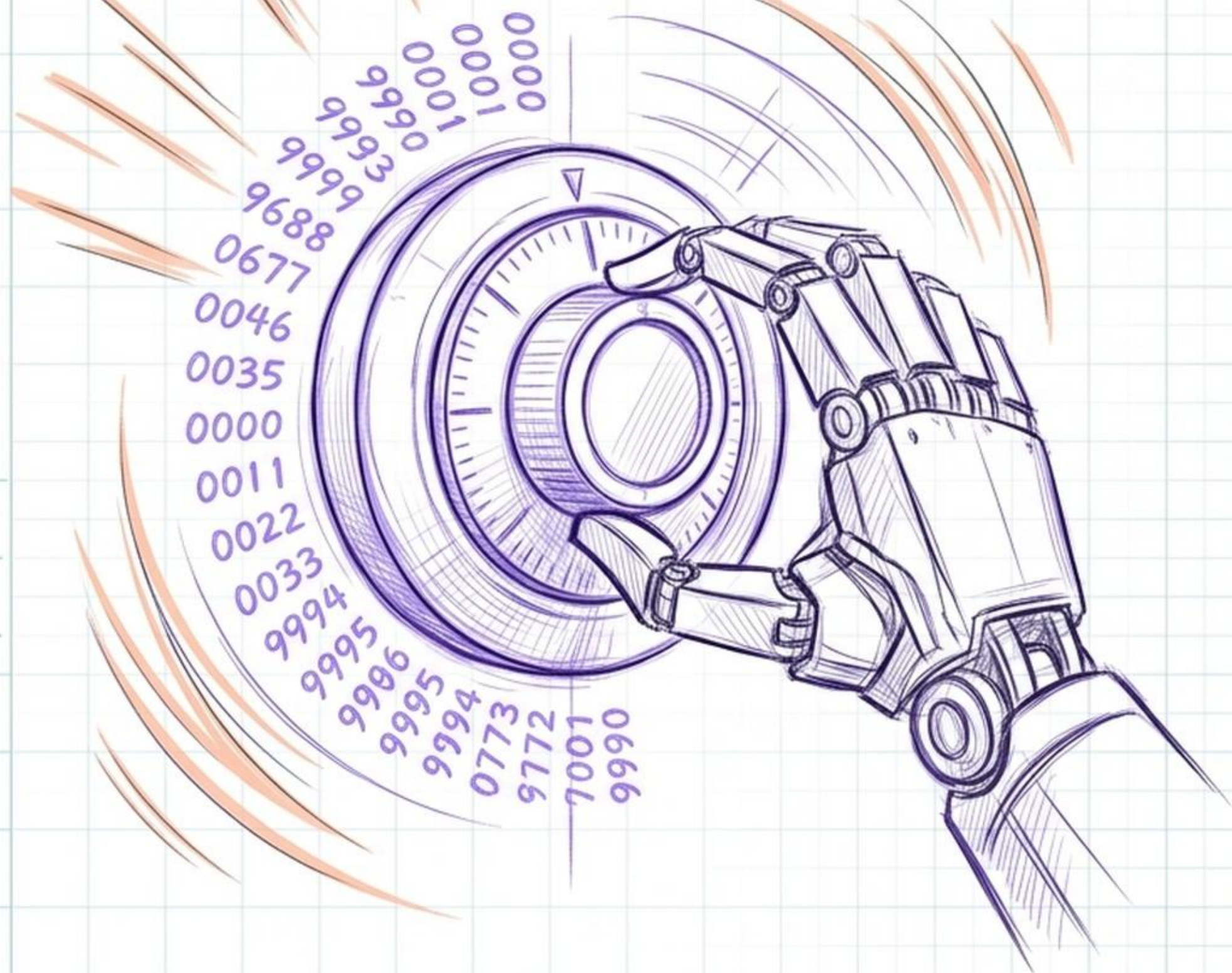
GPU集群
算力消耗
海量推理
.....



成本飙升

暴力搜索与回溯

0000-9999



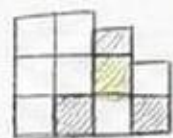
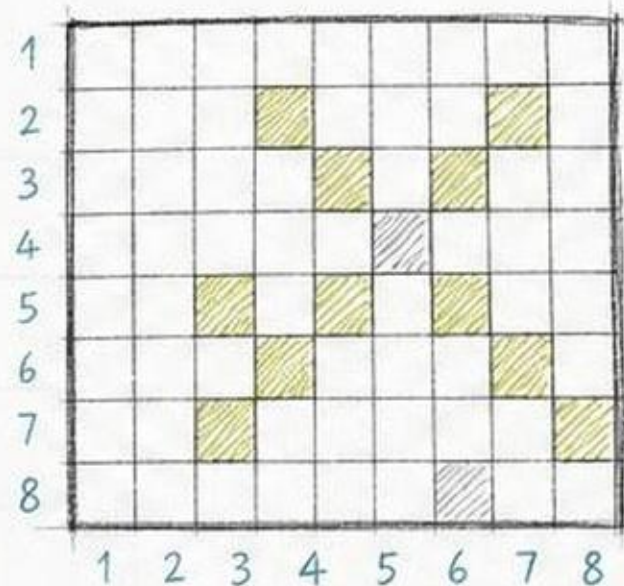
缺乏真实“理解”

ARC-AGI-1

简单平面图案

特征:

- ✓ 平面
- ✓ 规则明确
- ✓ 模式单一
- ✓ 逻辑直接



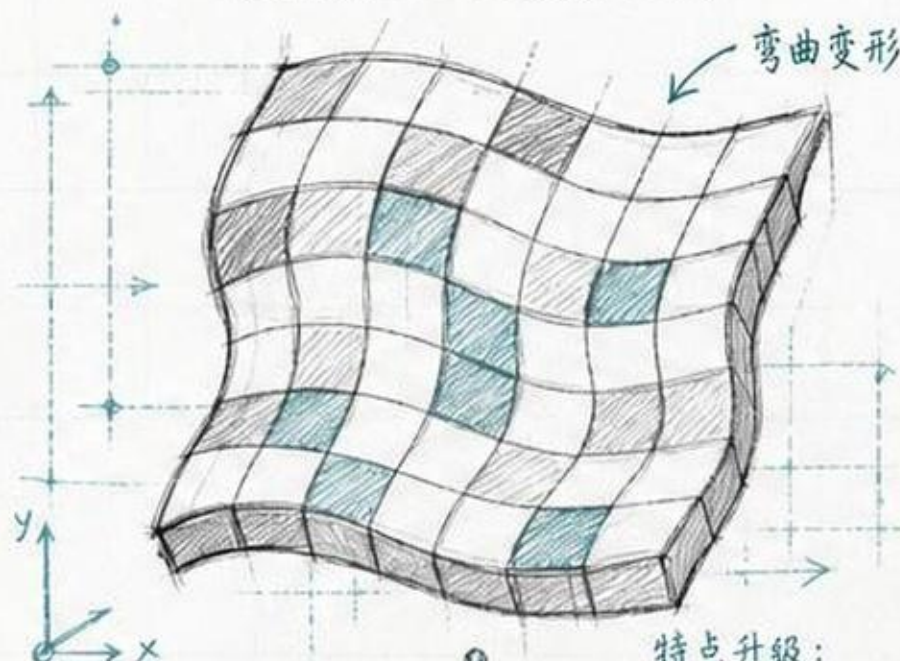
示例任务

93%

成绩暴跌

ARC-AGI-2

轻微扭曲的复杂几何



特点升级:

- ✓ 规则嵌套
- ✓ 形状变化
- ✓ 多步推理
- ✓ 部分歧义

68.8%

成绩暴跌

ARC-AGI-3

高度扭曲的多维互动环境



挑战跃迁:

- × 规则动态变化
- × 多维空间交互
- × 因果链复杂
- × 长期规划
- × 开放式目标
- × 观察与行动
- × 不确定性高

规则彻底颠覆

成绩暴跌

从平面图案
到互动沙盘



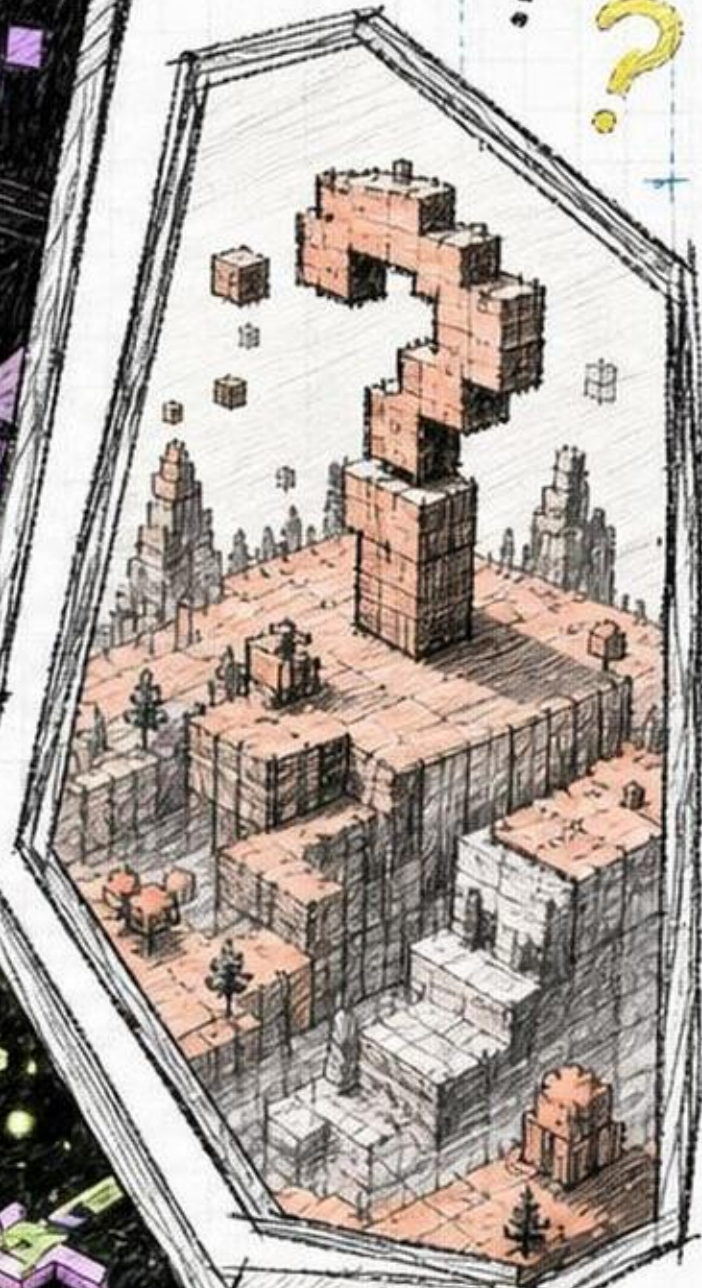
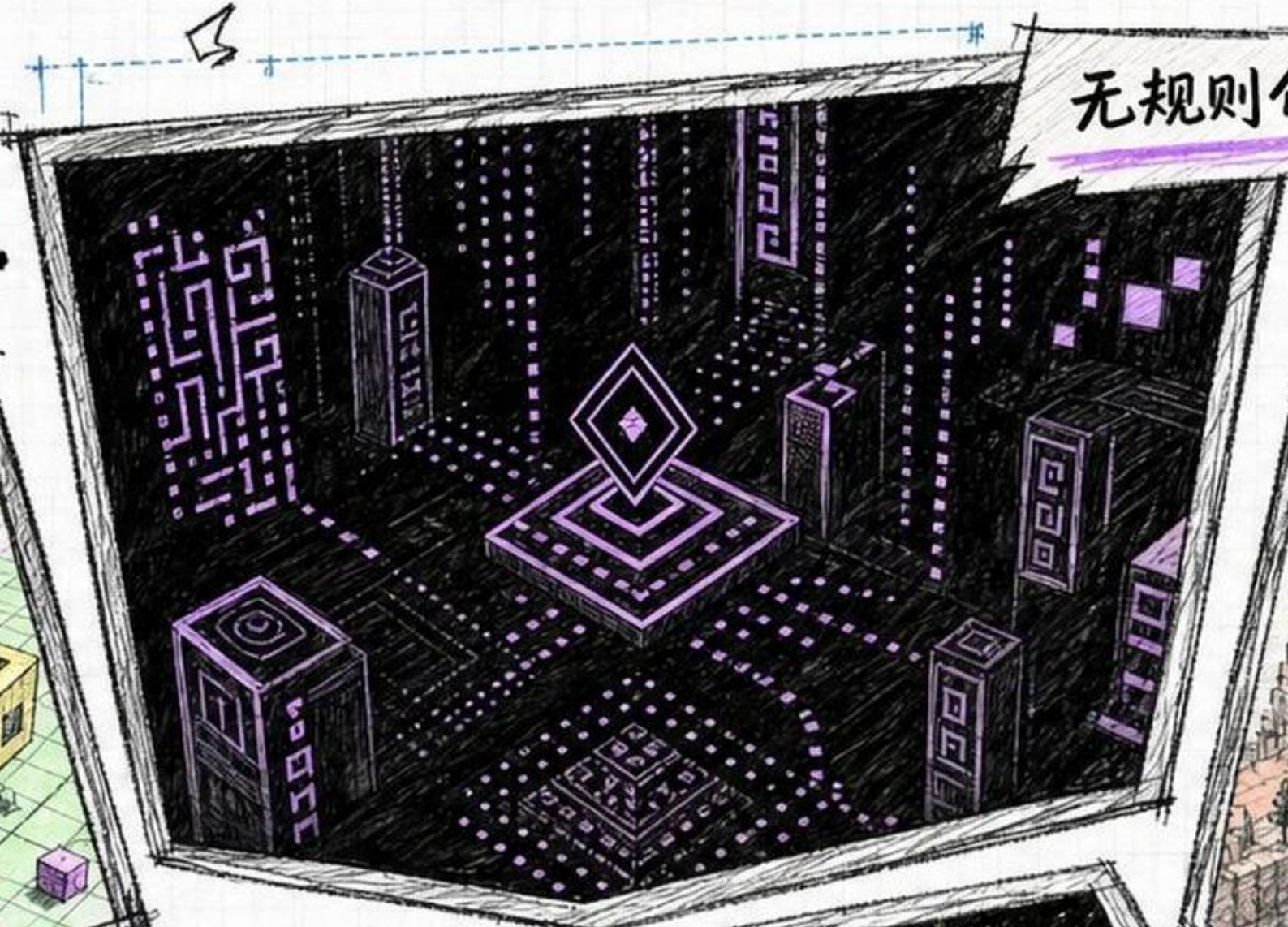
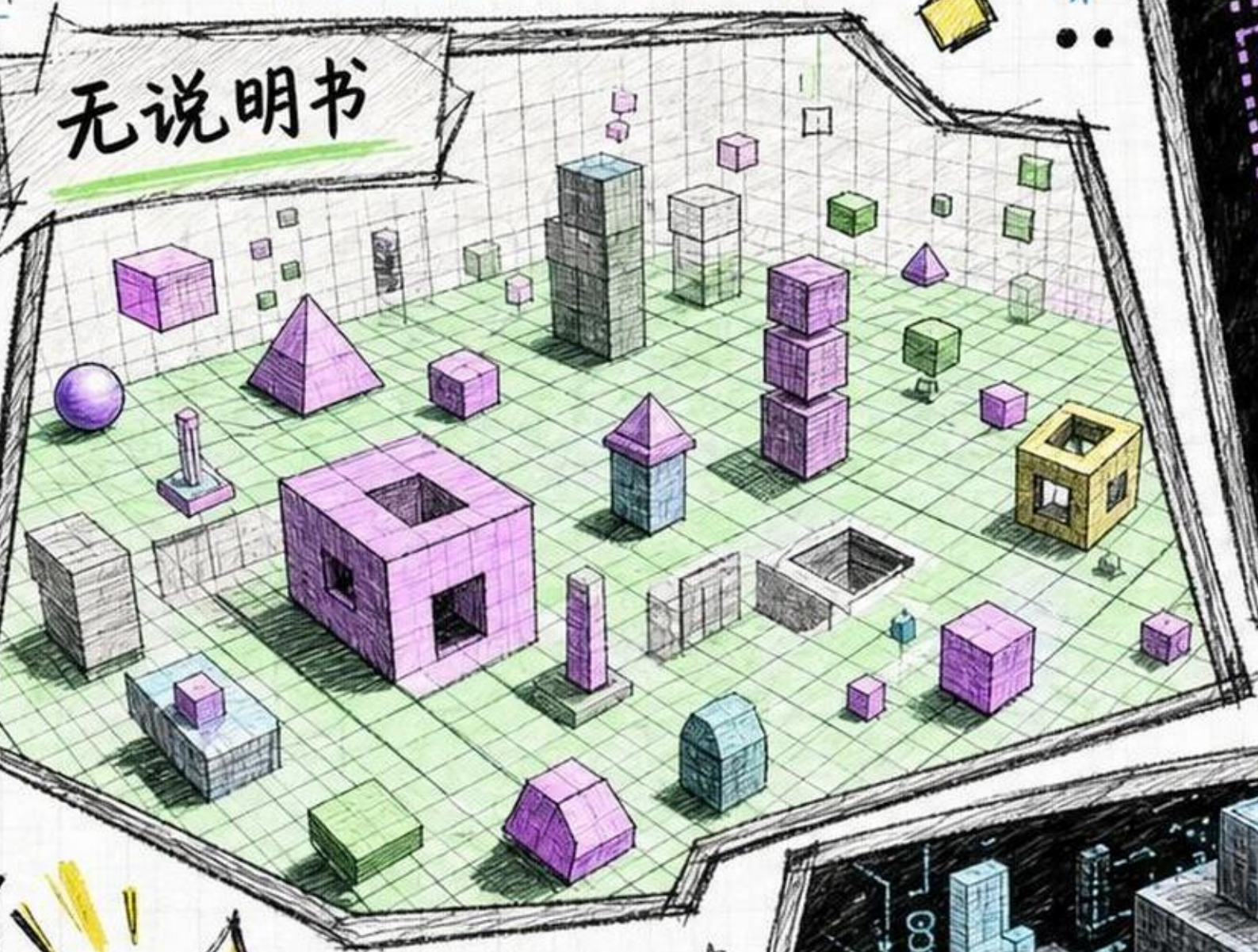
备注:

AI 在早期版本中表现优秀,但随着测试进化,性能大幅下滑。

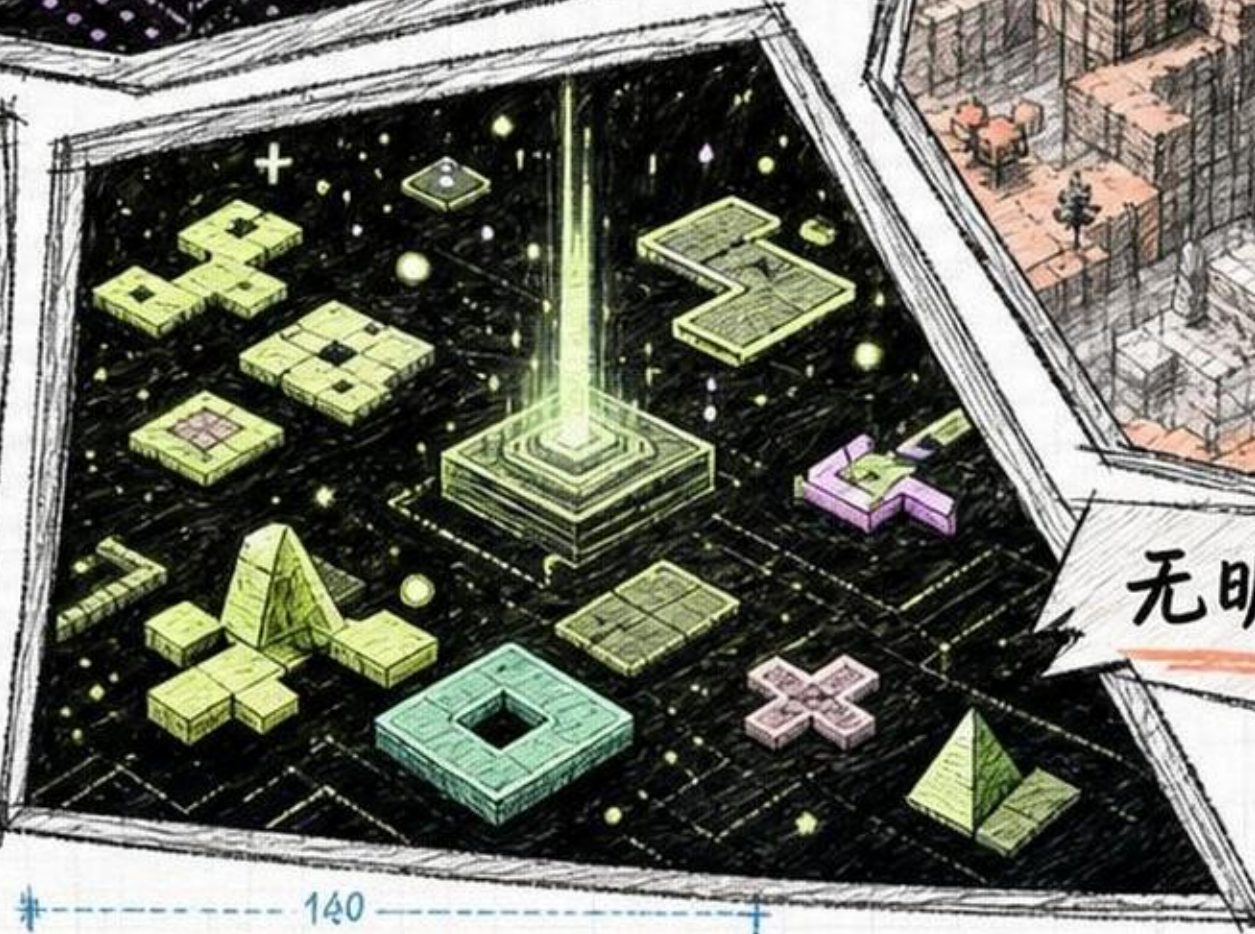
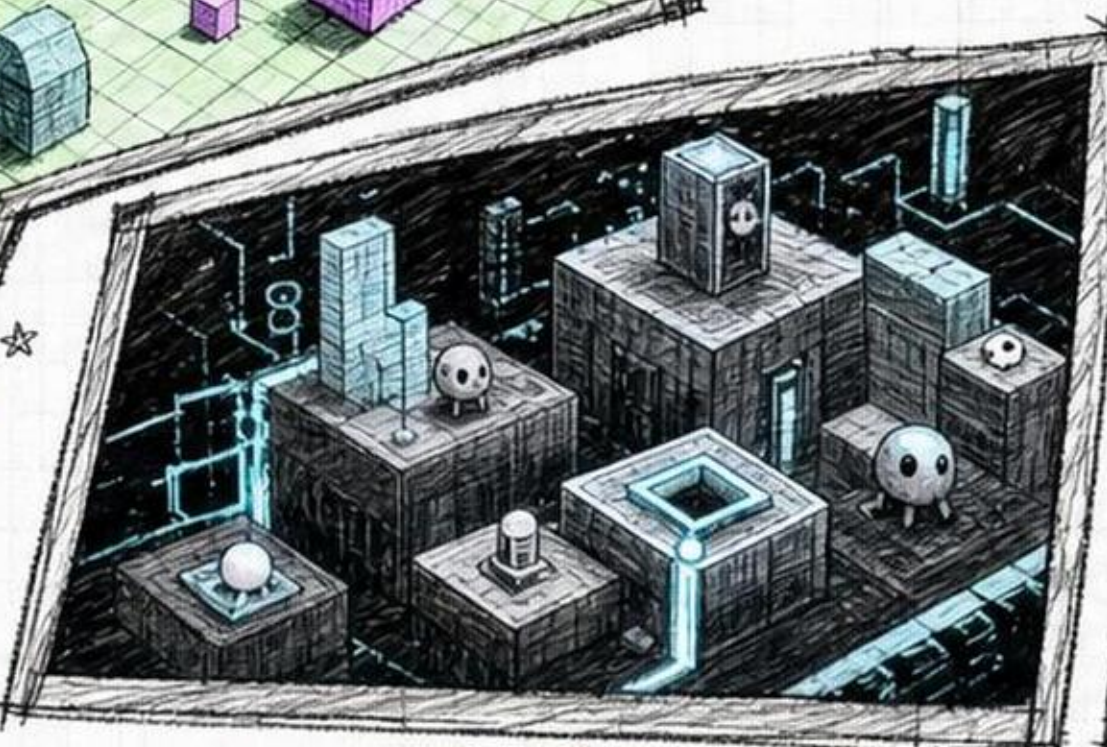
ARC-AGI-3

无规则介绍

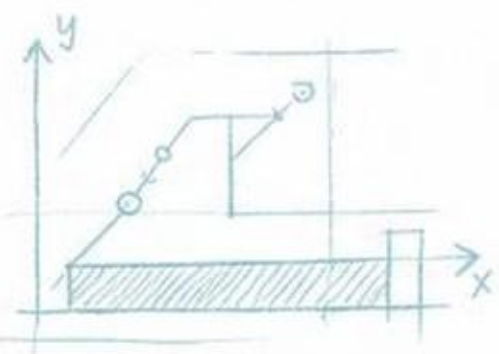
无说明书



每个世界，
完全不同



无明确目标



学习效率 = 步数最小化



Diagram illustrating the first step: blind clicking. A character is shown interacting with various mechanical components (spring, button, grid) without understanding their function. The formula $F=kx$ is visible on the left.

第一步：盲目点击

Diagram illustrating the second step: action observation. The character observes the effects of different actions (pressing, pushing, clicking) on the components, leading to a better understanding of their functions.

第二步：动作观测

Diagram illustrating the third step: deciphering mechanism and goal. The character understands the mechanism and the goal (lighting up all grid cells).

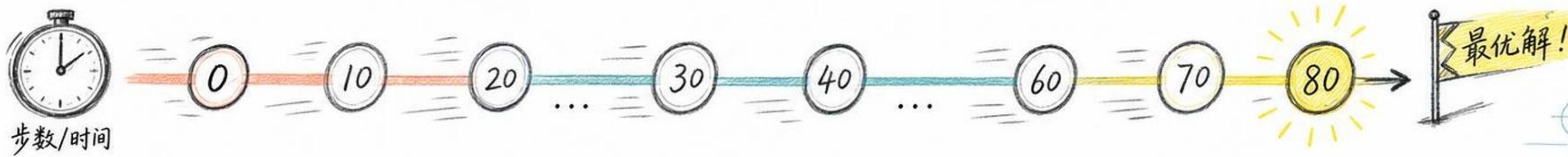
哇！我懂了！

机制规律总结

- 踩下 → 弹起
- 推动 → 移动
- 点亮 → 记录

目标：点亮全部格子

第三步：破解机制与目标



流体智能

未知环境

从零开始

核心定义：
在未知任务上的
技能获取效率

未知环境

从零开始

肖莱, 2019



人类 vs 顶级AI模型

真正的流体智能测试



在真正衡量推理、归纳、类比和抽象思维的流体智能测试中，人类与顶级AI模型之间存在着惊人的鸿沟。

顶级AI模型

- Gemini 3.1: 0.37%
- GPT-5.4: 0.26%
- Claude 4 Opus: 0.15%



约270倍
差距

人类 100%

100%

人类基准
(100%)

代表人类在流体智能
测试中的平均表现

80%

60%

40%

20%

0%

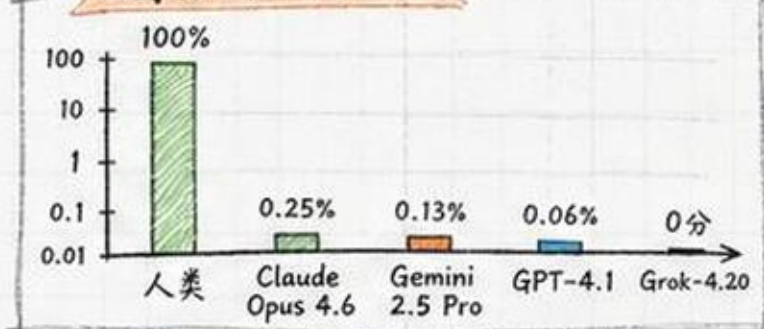
差距之大
超乎想象!

AI模型表现
(不足1%)

即使是最先进的AI模型，
表现也不到人类的1%。

SWE-bench Verified 性能对比示意图

得分对比 (%)



人类独占山巅，AI仍在迷雾中摸索。

人类 100%

差距：

两到三个
数量级

Claude Opus 4.6:
0.25%

约为人类的
1/400

Gemini 2.5 Pro:
0.13%

约为人类的
1/800

GPT-4.1:
0.06%

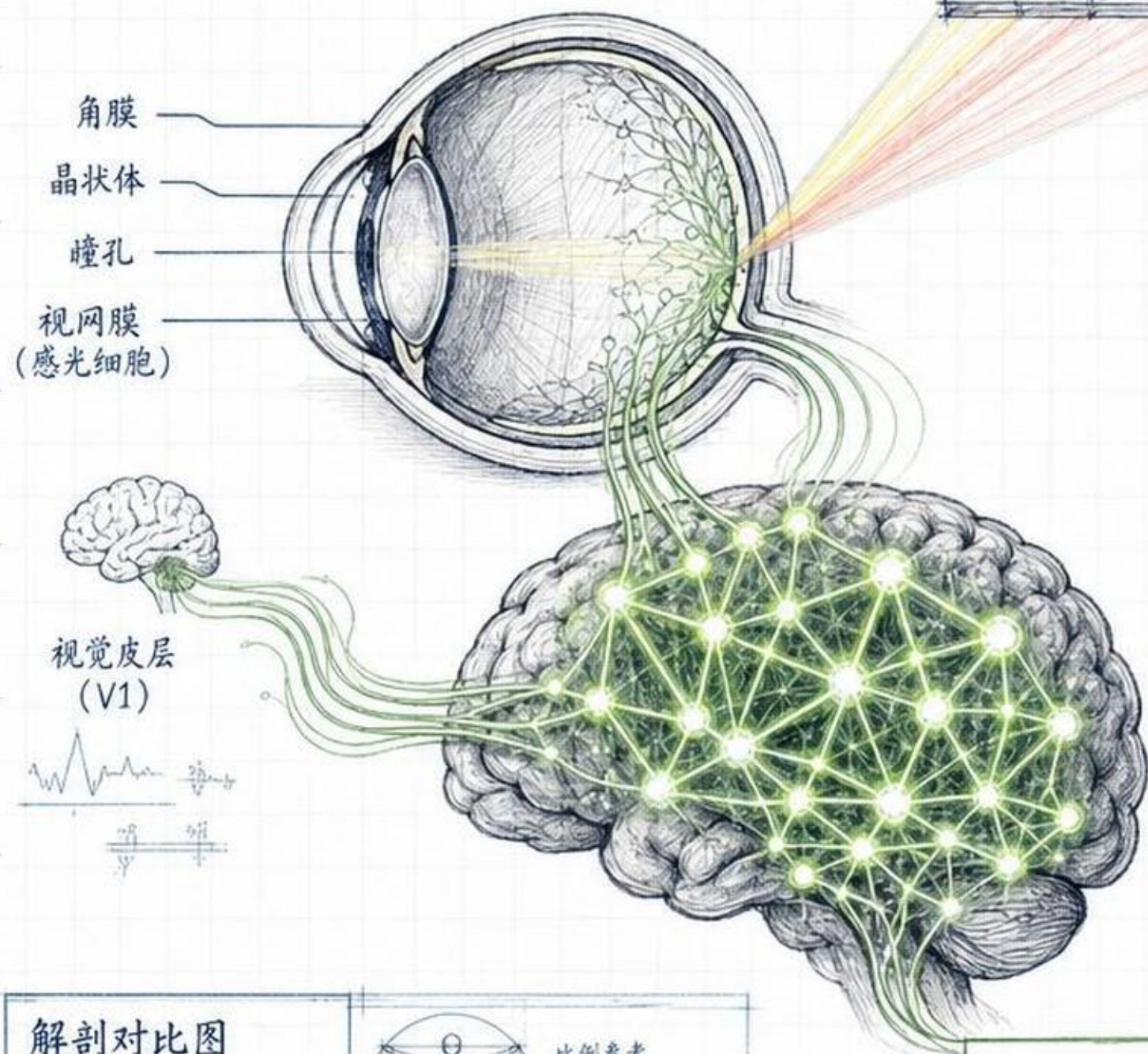
约为人类的
1/1,700

Grok-4.20:
0分

未能解决任何
一个问题

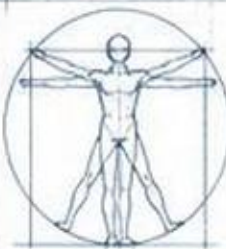
连山脚下的入口都没找到

人类视觉



解剖对比图

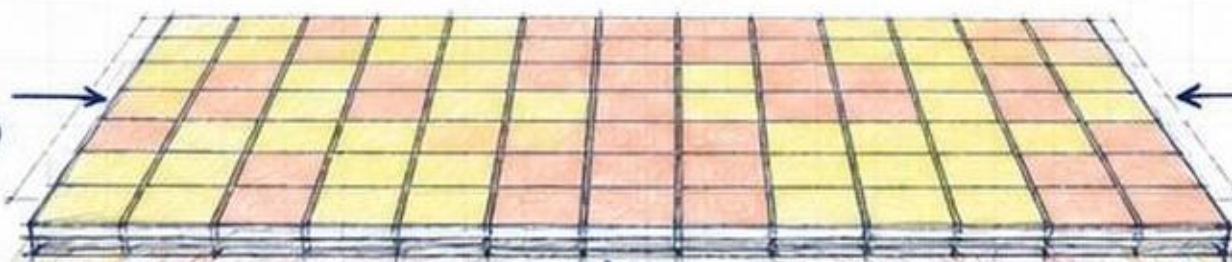
- 生物神经连接 (有机)
- 机械电路连接 (非有机)



比例参考
人眼结构 ≈ 24mm
镜头结构 ≈ 24mm
处理核心位置对比

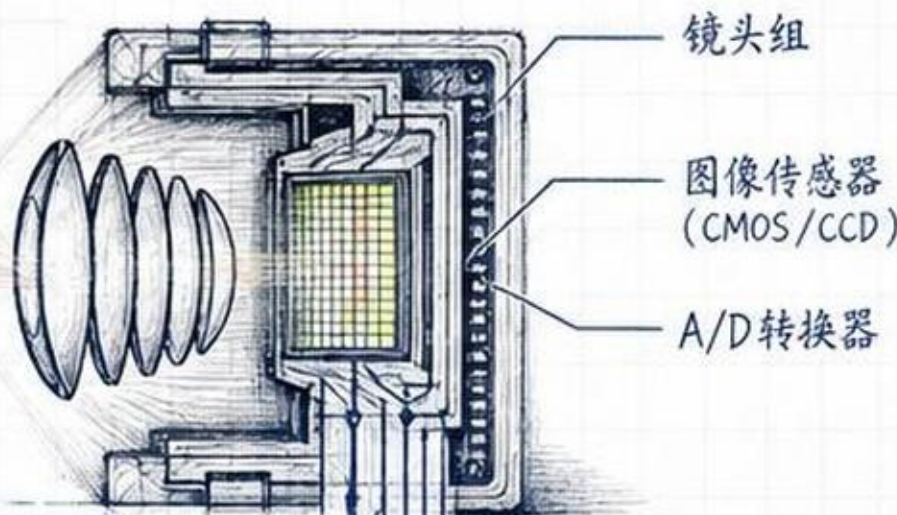
信息输入与操作接口完全相同

像素矩阵
(Pixel Grid)



每个像素接收
相同的光信号

机器视觉



它缺的是
理解

理解力

- ✓ 语义关联
- ✓ 经验整合
- ✓ 情境理解
- ✓ 概念抽象
- ✓ 价值判断

形成意义
产生理解与认知

缺失：
真正的理解力

仅进行数据处理
不产生真正的理解

处理流程

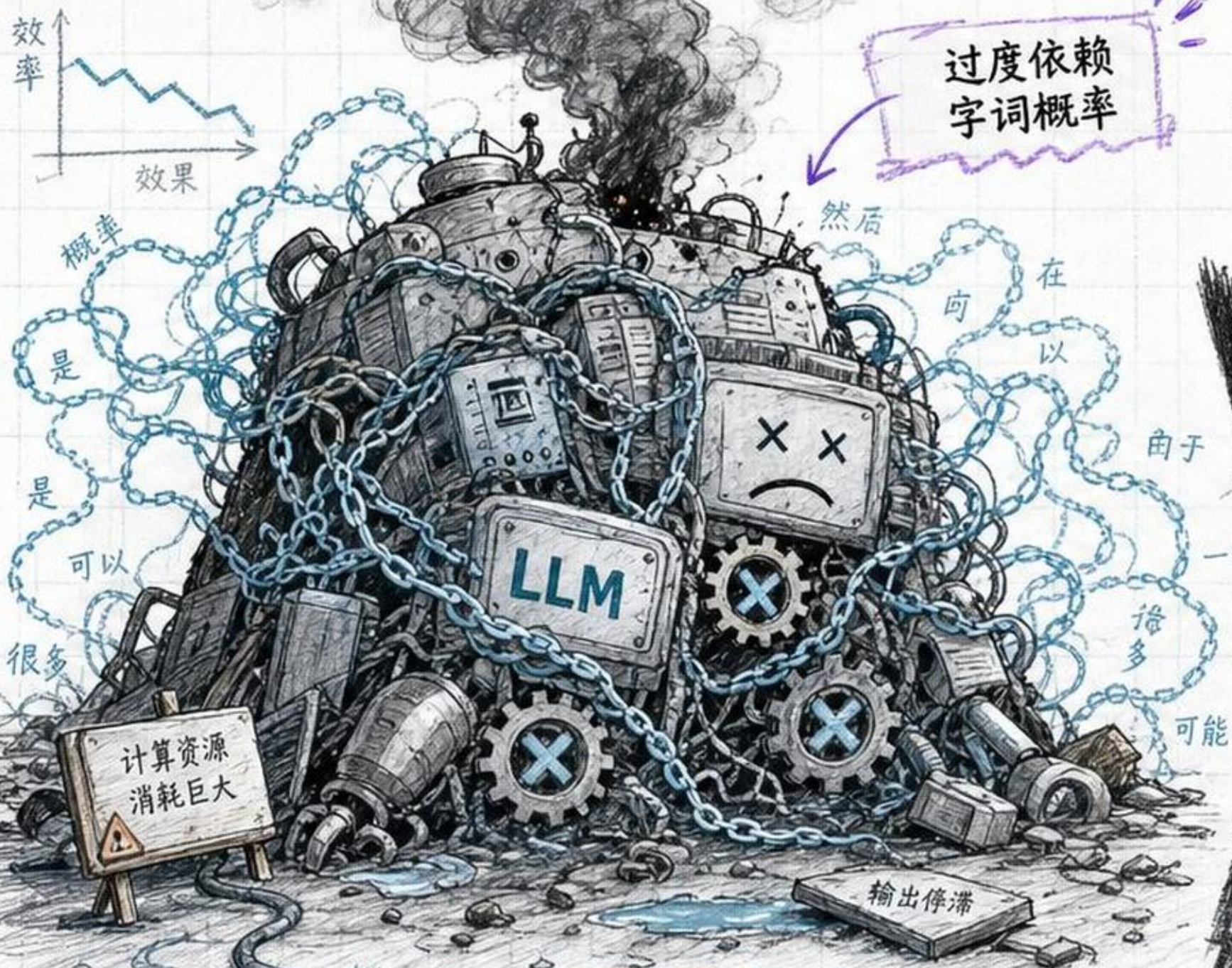
- ☐ 特征提取
- ☐ 模式识别
- ☐ 分类输出

设计目的：理解差异
核心差异：是否形成意义

DATE: 2024.05.20
DRAWN BY: Insight Lab
VERSION: 1.0



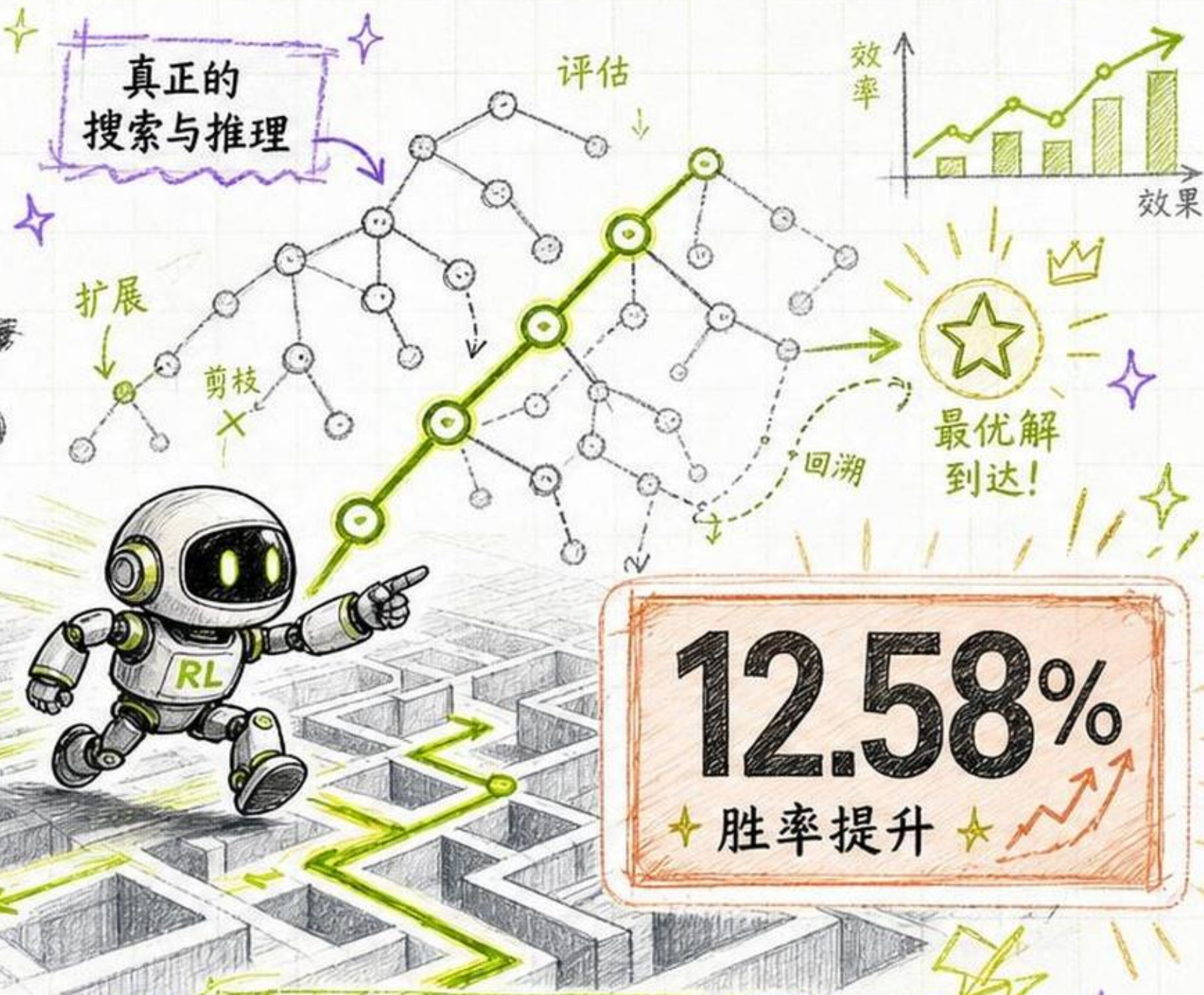
万亿参数大语言模型 (LLM)



败者



强化学习 + 图搜索系统

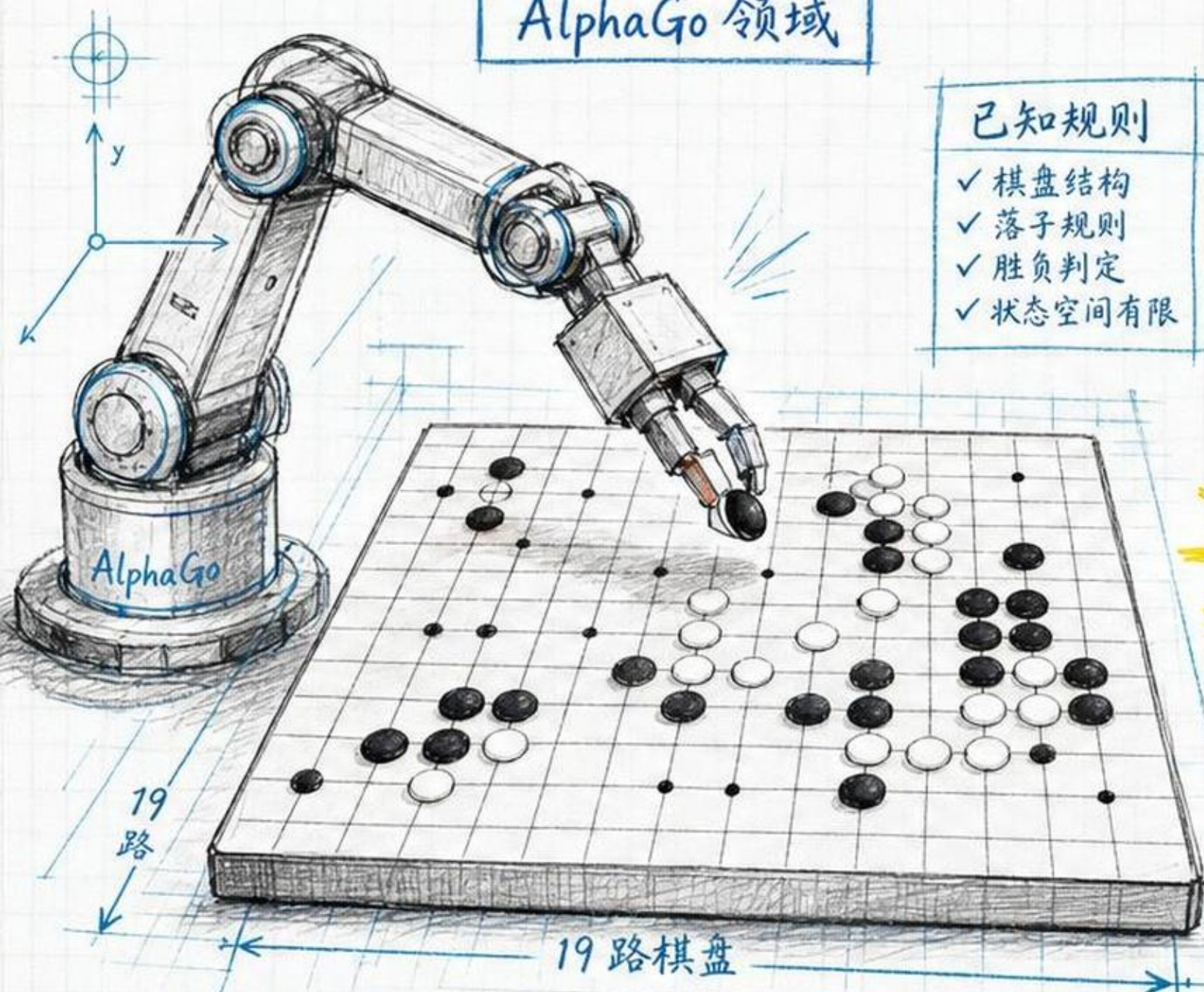


胜者



规则固定且已知 (极致优化)

AlphaGo 领域



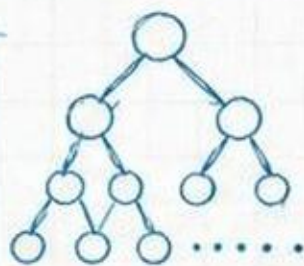
已知规则

- ✓ 棋盘结构
- ✓ 落子规则
- ✓ 胜负判定
- ✓ 状态空间有限



精确计算

- ✓ 深度搜索
- ✓ 策略优化



可计算
可预测
可优化



已知
VS
未知

规则彻底未知 (语言非智能全部)

ARC-AGI 领域



未知规则

- ✗ 规则不明
- ✗ 目标不明
- ✗ 空间无限
- ✗ 无法预测

无法穷举
无法预测
无法优化



迷宫结构(动态变化)

- 墙体重组
- 路径变换
- 规则未知
- 持续演化中

摸出门道 ✓

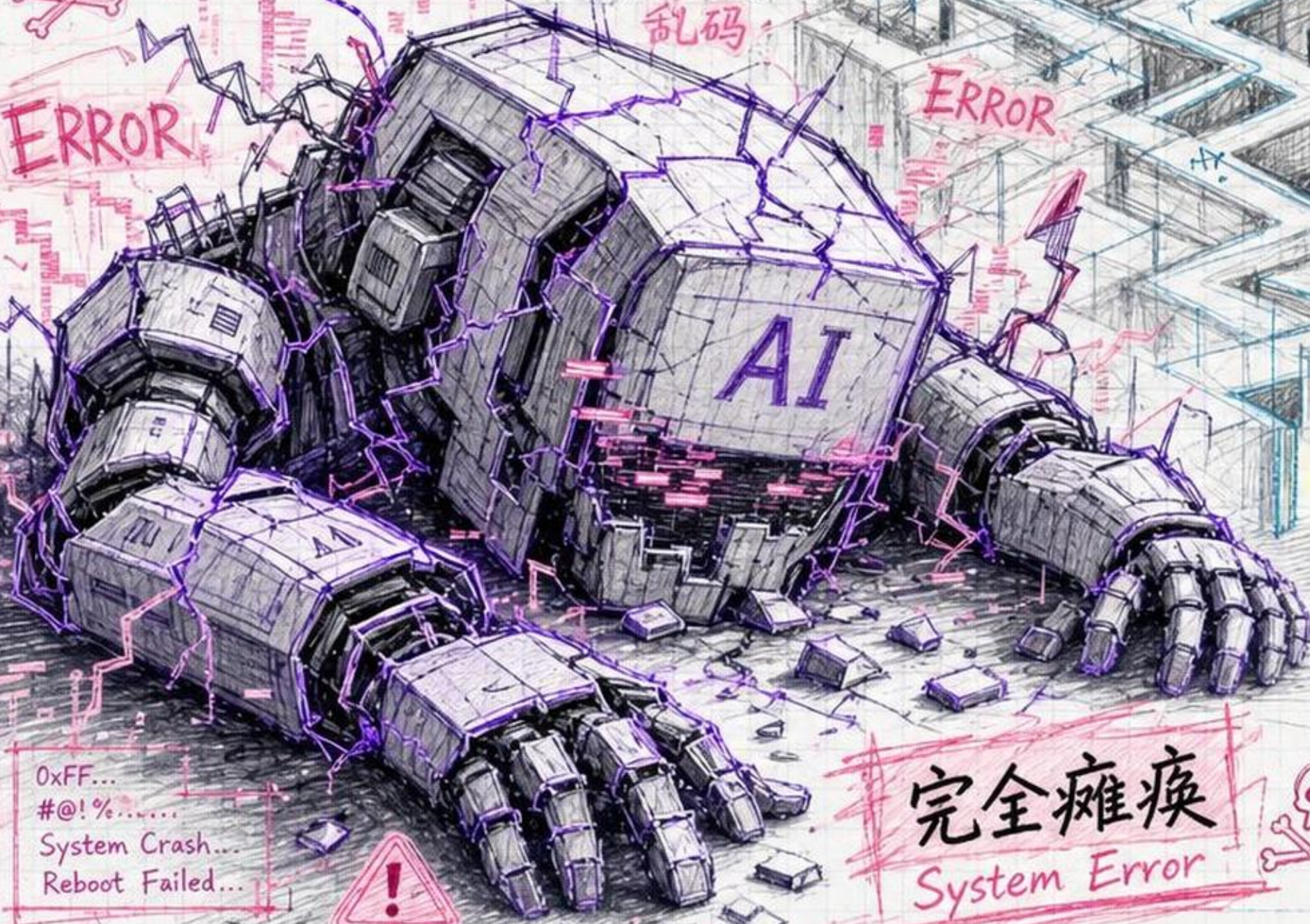


ERROR

乱码

ERROR

ERROR



规则: 未知

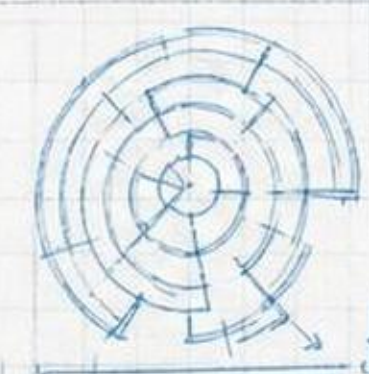
完全瘫痪
System Error



0xFF...
#@!%.....
System Crash...
Reboot Failed...



- 墙体重组
- - 可能路径
- 未知区域
- ⚡ 能量扰动



① 第一格：发现怪异装置

② 第二格：愣神试探旋转

③ 第三格：触发冲水成功



人类：几秒钟搞定

对比 AI:

- 数据边界限制
- 缺乏真实体验
- 难以快速泛化



彭罗斯意识理论

哥德尔推演 (不可计算成分)

哥德尔不完备定理

对于任何一致的、能够表达基本算术的形式系统 F ：
 F 既不完备，也不能证明自身的一致性。

$\forall x (R(x) \rightarrow S(x))$
 $R(a)$

$\therefore S(a)$?

未能证明

G_F

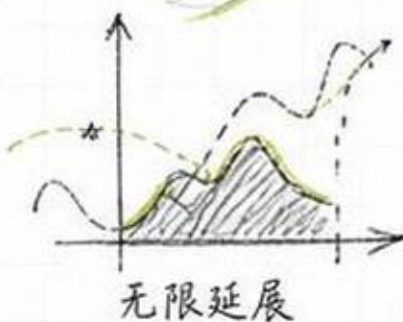
G_F ：“我在 F 中不可证明”

哥德尔句子构造

$G_F \equiv \neg \text{Prov}_F('G_F')$

存在真命题
在系统 F 中
不可证明

	真值
可证明	×
不可证明	✓

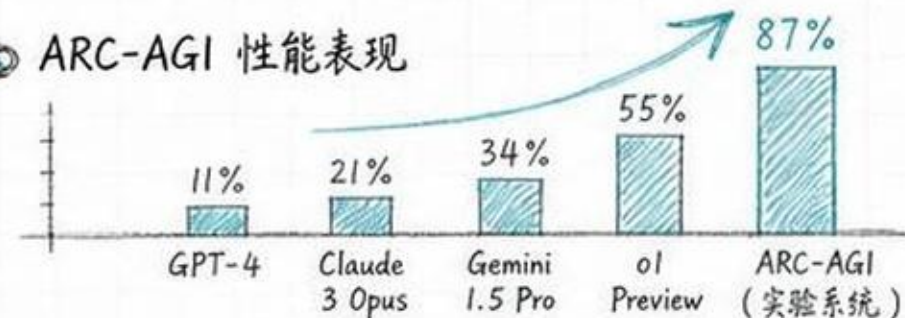


人类智能中存在不可计算成分：
它超越形式系统的界限，
在物理现实与意义空间中
孕育创造与理解。

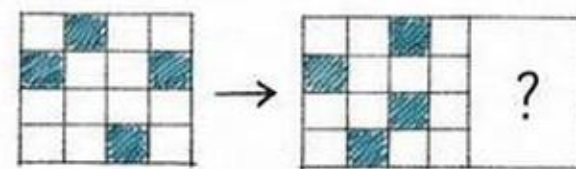
无法被算法触及

ARC-AGI 实验铁证

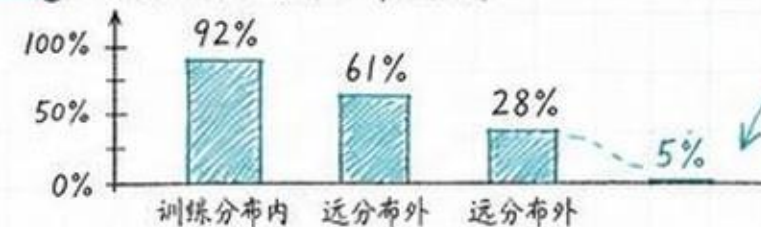
ARC-AGI 性能表现



典型 ARC 任务示例



泛化能力 (跨分布测试)

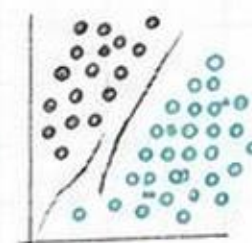


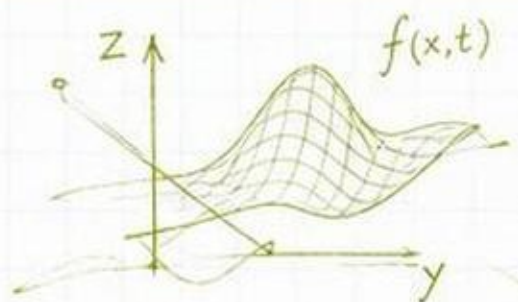
人类表现
远超算法模型

高效推理与少样本学习

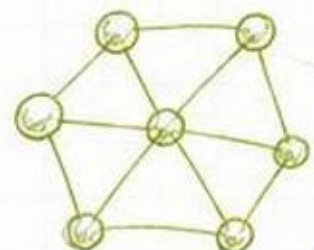


少样本提示
即可发现
深层规律





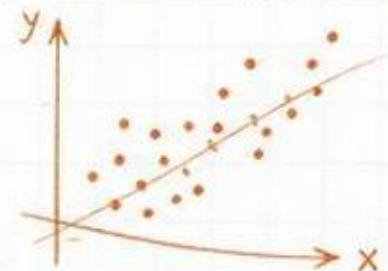
万物皆可推演? →



$$\nabla \cdot E = \frac{\rho}{\epsilon_0}$$
$$H \psi = \hat{H} \psi$$

模型假设
边界条件
初始状态

完美世界的推演



真实世界
从不按剧本走



$$y = f(x) + \epsilon$$

噪声·离散·偏差
异常·不确定性

真实世界的回击

理论层面推演

面对未知的
实验数据

堆算力
无法复制
灵魂

当两种真相同时出击，唯一的谎言便土崩瓦解！

完美模仿

• 语言流畅度

语音输出模块
V3.2

语音模型
参数冻结

- ✓ 发音标准
- ✓ 节奏精准
- ✓ 逻辑连贯

床前明月光，
疑是地上霜。
举头望明月，
低头思故乡。

环境变化
超出训练分布

准确率 99.8%
延迟 120ms
流畅度 ★★★★★

流畅 ≠ 理解

零适应能力

• 失去预训练

未见过
无法处理

上下文丢失
逻辑断裂

推理崩溃

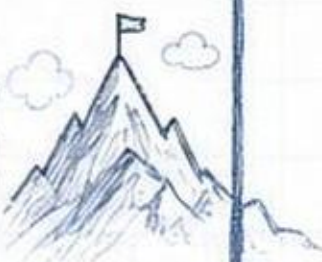


现有能力边界:

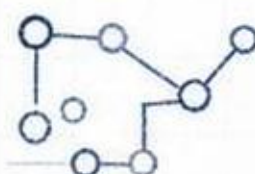
- 模式识别
- 统计规律学习
- 局部最优解

目标: 通用智能 (AGI)

- 自主学习与迁移
- 跨领域泛化能力
- 复杂推理与规划
- 意识与自我驱动



200万美元突破奖金



大语言模型

推理增强模型

当前整个AI范式 (大模型·推理增强)

质的差异·根本性能力缺失

难以跨越的
能力鸿沟



流体智能：三岁孩子的天赋



没有说明书



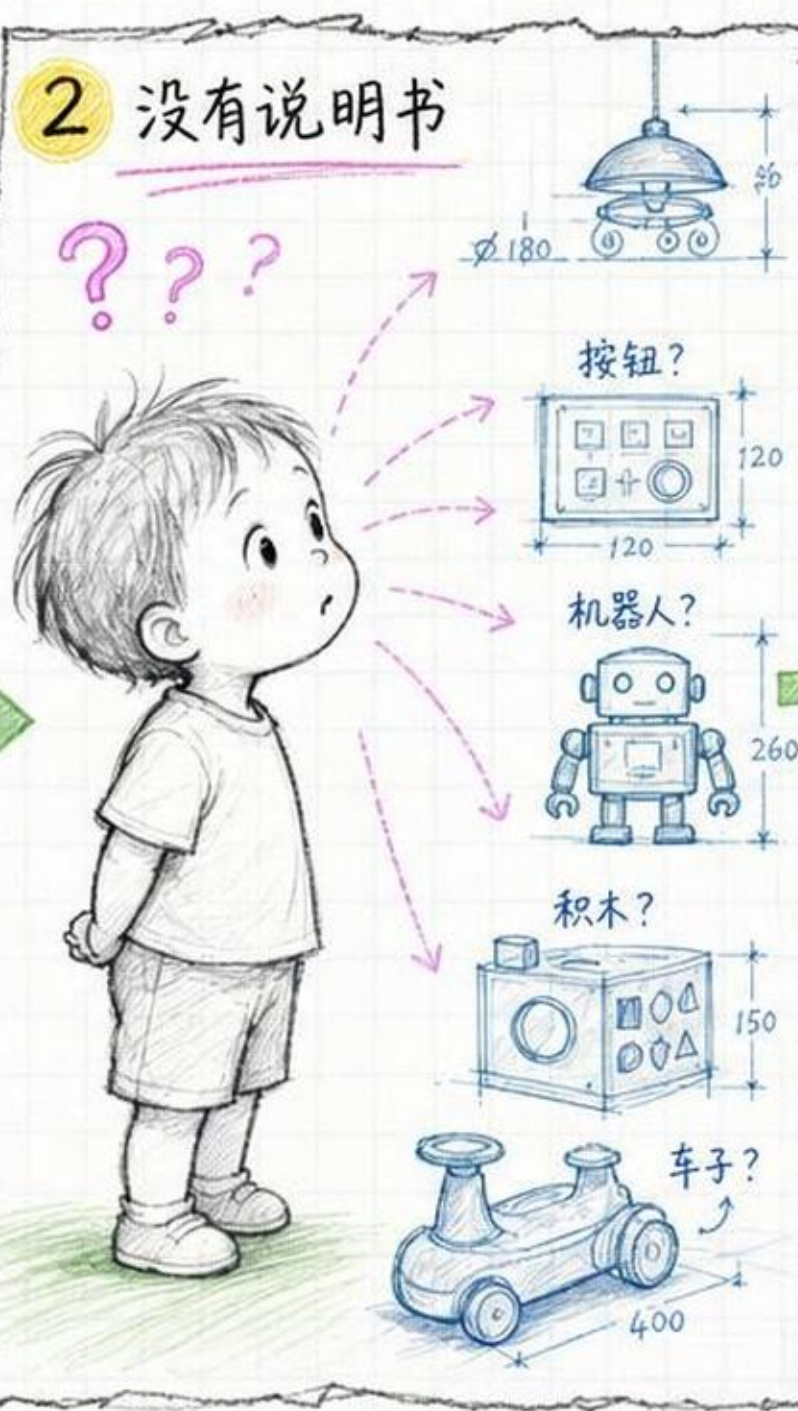
零次训练

探索世界，
天生会的事。

1 陌生的房间



2 没有说明书



3 零次训练



4 看一眼就懂了



不靠记忆，不靠模仿，天生就能理解世界的运行方式。



= 无限可能

Gestalt

格式塔感知



零散信息瞬间组织为有意义的整体



FROM CHAOS
TO CLARITY

OBSERVE
ORGANIZE
UNDERSTAND

生物计算机

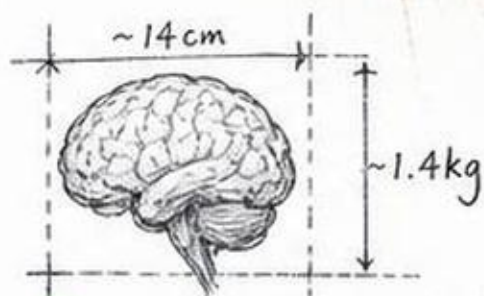
约10比特/秒 · 20瓦

神经元
~86亿个

脉冲信号
低电压, 低功耗

线粒体供能
高效生物能量

能效比赢家
~ 10^3 比特/焦耳



- ✓ 自学习
- ✓ 自适应
- ✓ 低能耗
- ✓ 高效智能

无情碾压

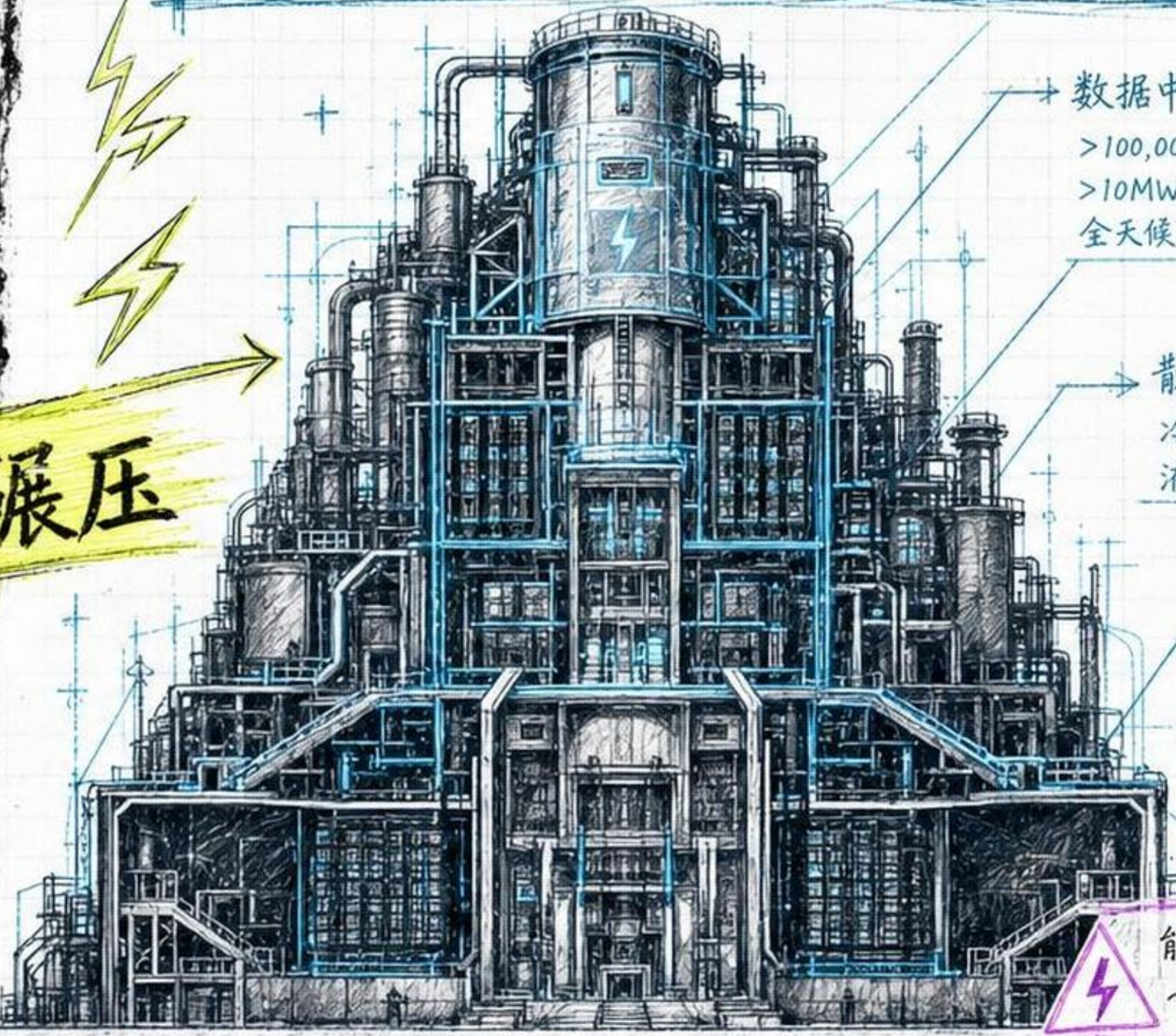
百万美元电费硅基神坛

数据中心集群
>100,000 GPU
>10MW 功耗
全天候运行

散热系统
冷却塔 × N
液冷循环

电力输入
>10MW
年耗电
>\$1,000,000

能效比劣势
~ 10^2 比特/焦耳



ARC-AGI-3的讽刺现实

智能?

电费即真理?

‘理解’是
未知的物理事件



大脑的“理解”
是否遵循已知物理定律？
还是一种新的物理现象？

未知的动力学？

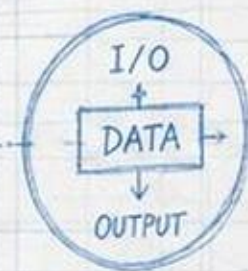
$U = ?$

($U = \text{Understanding}$)

计算 $\neq$ 理解？



Turing Test $\neq$ Understanding

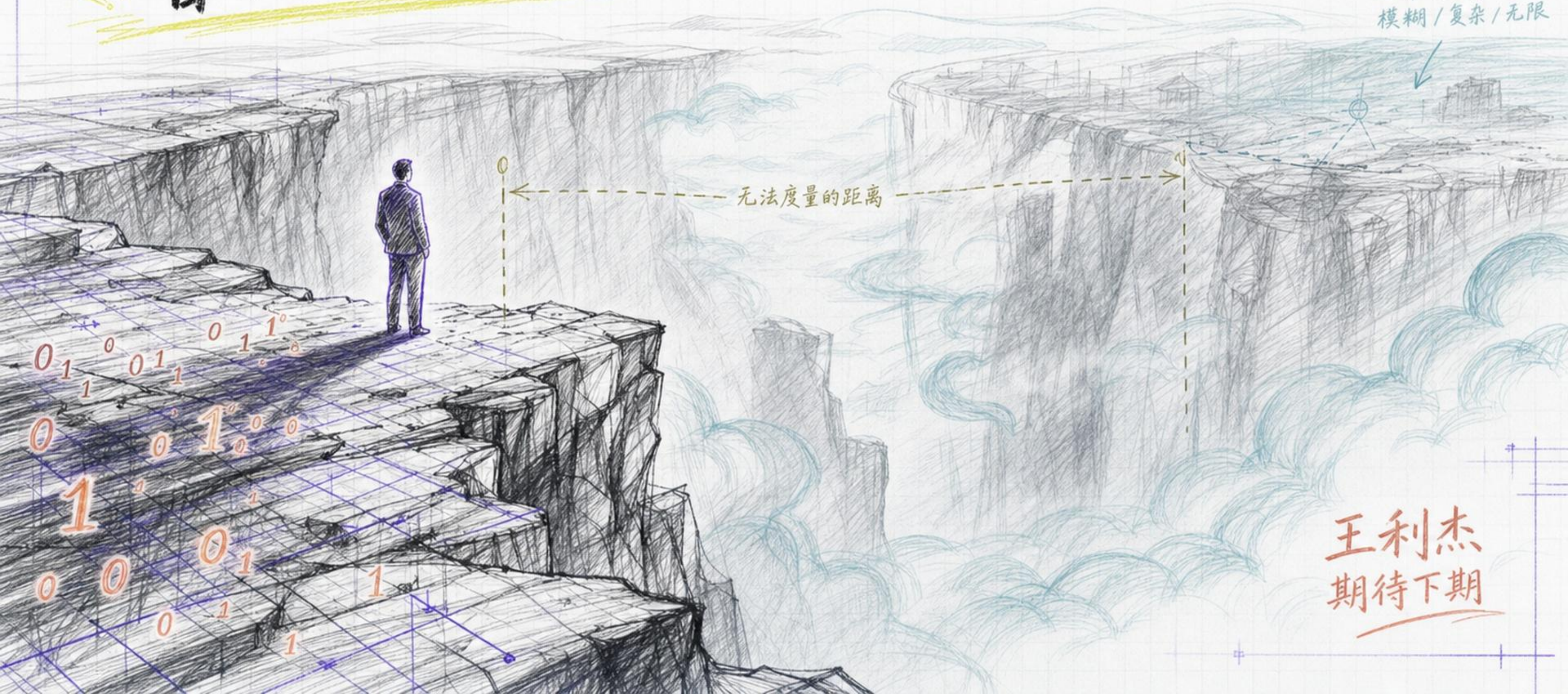


0.37%

冰冷的现实

智能的鸿沟，还是对理解本身的轻视？！！

理解的彼岸
模糊 / 复杂 / 无限



王利杰
期待下期