

使用大模型 AI 的 第一性原理

孙志岗

[AI师傅](#)创始人、CEO

原网易教育事业部，首席产品架构师

原哈工大计算机，副教授



你是否有下面的困惑？



AI 为啥懂这么多？它「知识」从哪儿来？



它是真的“思考”吗？跟人类的大脑有什么不同？



为什么同一句话有时候回答得很好，有时候答非所问？



跟 AI 聊天需要说“谢谢”吗？这会影响结果吗？



AI 会不会偷偷记住我的隐私信息？



我输入的问题会被公司用来训练别人的模型吗？



提示词怎么写才算“专业”？有万能模板吗？



为什么有时答案听上去非常自信，却可能是胡编的？



AI 会不会带有偏见或歧视？这些偏见是怎么产生的？



它会“撒谎”吗？如果回答出错，我怎么判断？

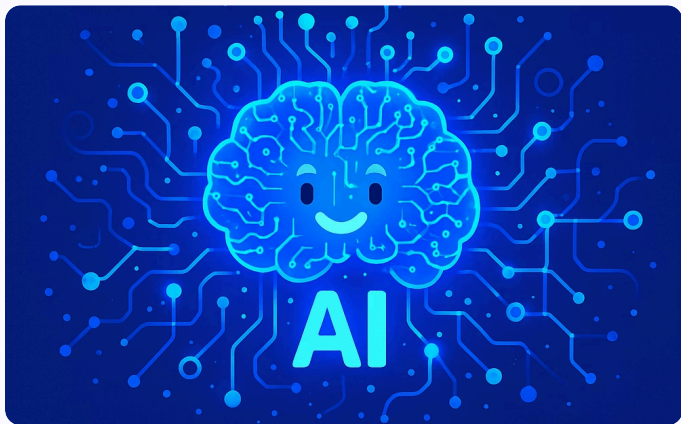


AI 真的会取代我的工作吗？哪些岗位最危险？



如果 AI 爆火之后“迭代太快”，我怎么跟得上学习？

今天探讨三个问题



大模型的原理



怎样随心控制大模型



AI 时代的个人发展

大模型的原理

没有公式，没有代码，文科生也能懂

为什么大模型
都是一个字一个字地
输出？

不是字，也不是词，而是 token

- ① **Token** 是大模型处理数据的最小单位。通常可以是一个字、词、单词、字符，甚至是 $\frac{1}{2}$ 、 $\frac{1}{3}$ 个字、单词，具体取决于模型如何将文本分割为更小的可处理片段。

GPT-4

GPT-4o & GPT-4o mini GPT-3.5 & GPT-4 GP

阳光、务实、自驱、坚持

Clear Show example

Tokens	Characters
14	11

阳00、务实、自00、00持

GPT-4o

GPT-4o & GPT-4o mini GPT-3.5 & GPT-4 GP

阳光、务实、自驱、坚持

Clear Show example

Tokens	Characters
9	11

阳光、务实、自驱、坚持

DeepSeek-V3

```
$ python3 deepseek_tokenizer.py
阳光
、
务 实
、
自 驱
、
坚 持
```

智能之源

只要能够非常好的预测下一个 **token**，就能帮助人类达到 **AGI**。

——前 OpenAI 首席科学家 Ilya Sutskever

智能就这么简单吗？

你说：锄禾日当午

甲接：汗滴禾下土

乙接：难过很喜欢

丙接：What's the meaning?

你认为谁智能谁不智能？



大模型是怎样预测下一个 token 的？

- 大模型里保存着所有 token 同时出现的**概率**
- 推理程序**（inference）基于输入，找**概率**大的下一个 token

输入	输出
我周末最喜欢的是	和
我周末最喜欢的是和	朋友
我周末最喜欢的是和朋友	们
我周末最喜欢的是和朋友们	爬山
我周末最喜欢的是和朋友们爬山	。

⚠ 没有检索、匹配、数据库、思考等等。只有**概率**！

「温度」参数可以控制不总是用概率最大的那个 token，产生多样性

大模型是怎样记住知识的？

训练数据

1. AI 正在改变我们的生活方式.....
2. **AI 技术在医疗领域有广泛应用**.....
3. AI 可以提高企业的生产效率.....
4. AI 算法能够预测市场趋势.....
5. AI 在自动驾驶汽车中扮演重要角色.....
6. AI 有助于个性化教育的发展.....
7. AI 机器人可以执行复杂的任务.....
8. **AI 技术正在推动智能家居的普及**
9. AI 在金融分析中发挥着关键作用
10. **AI 技术正逐步应用于艺术创作。**

- ✓ 大模型记住的是字词组合的**概率**，仅此而已
- 训练数据原文的半个字都没保存
- 这其实是一种信息压缩，和人一样

模型参数

「注意力」机制示意：

1. 「AI」后是「技术」的概率 30%
2. 「AI」后是「在」的概率 20%
3.

① 模型**参数**，也叫**权重**，是指在机器学习模型训练过程中，通过学习数据特征而调整的值，决定模型的输出

② 思考：为什么大模型有幻觉？

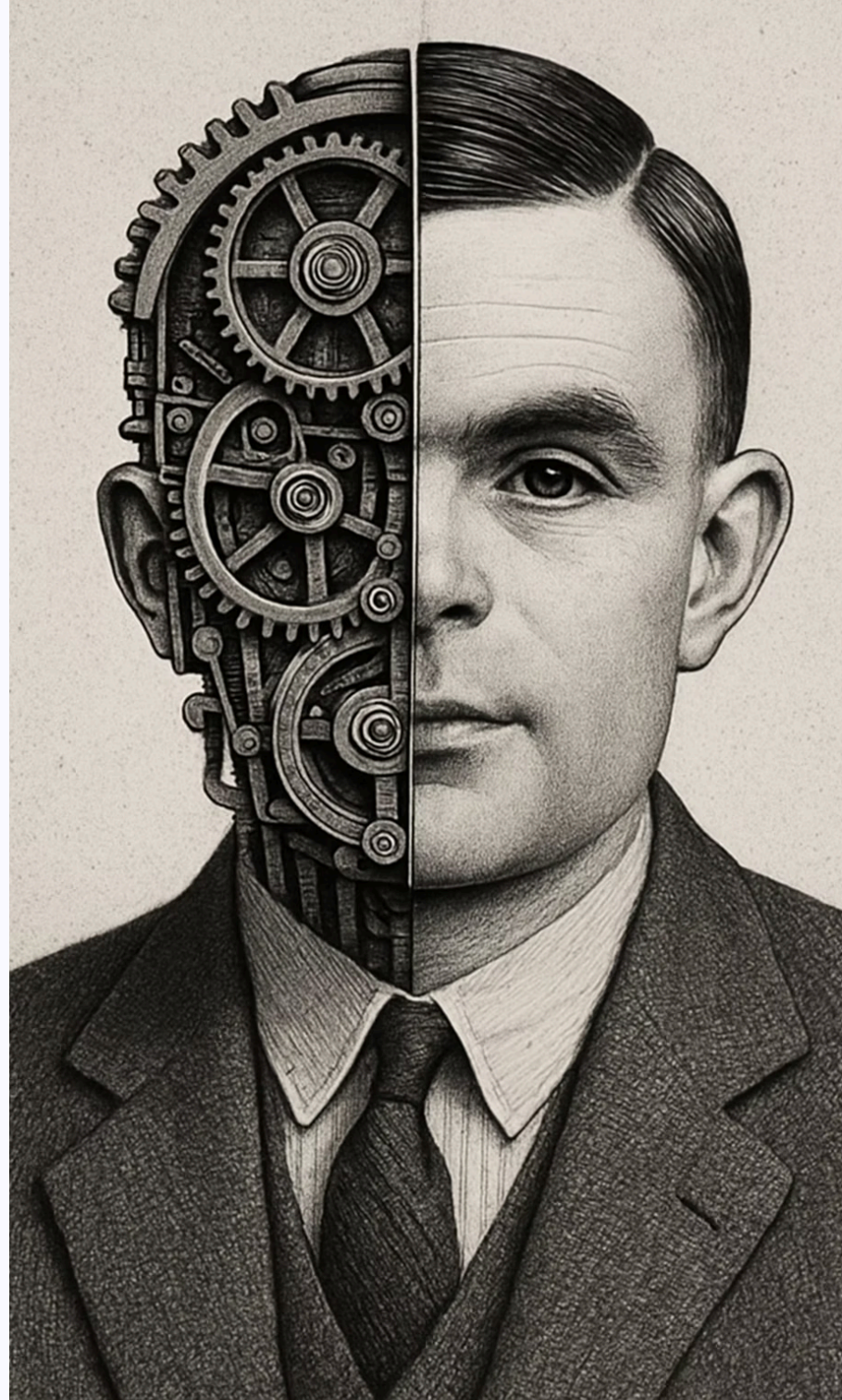
机械 vs. 智能

如果一台机器绝对不会犯错，它就不可能是智能的。

——阿兰·图灵 1947 年在[伦敦数学学会的演讲](#)

- ① • AI 的目标不是 100% 正确，而是比人类正确率高
- 应用 AI 的过程充满不确定性，一切努力都是为了提高正确率

- ☑ 在机械与智能之间寻找最佳结合点，是解决问题的关键



揭秘一下训练过程

所谓机器学习，和人类学习没多大区别

预训练

狂背书，掌握大量知识

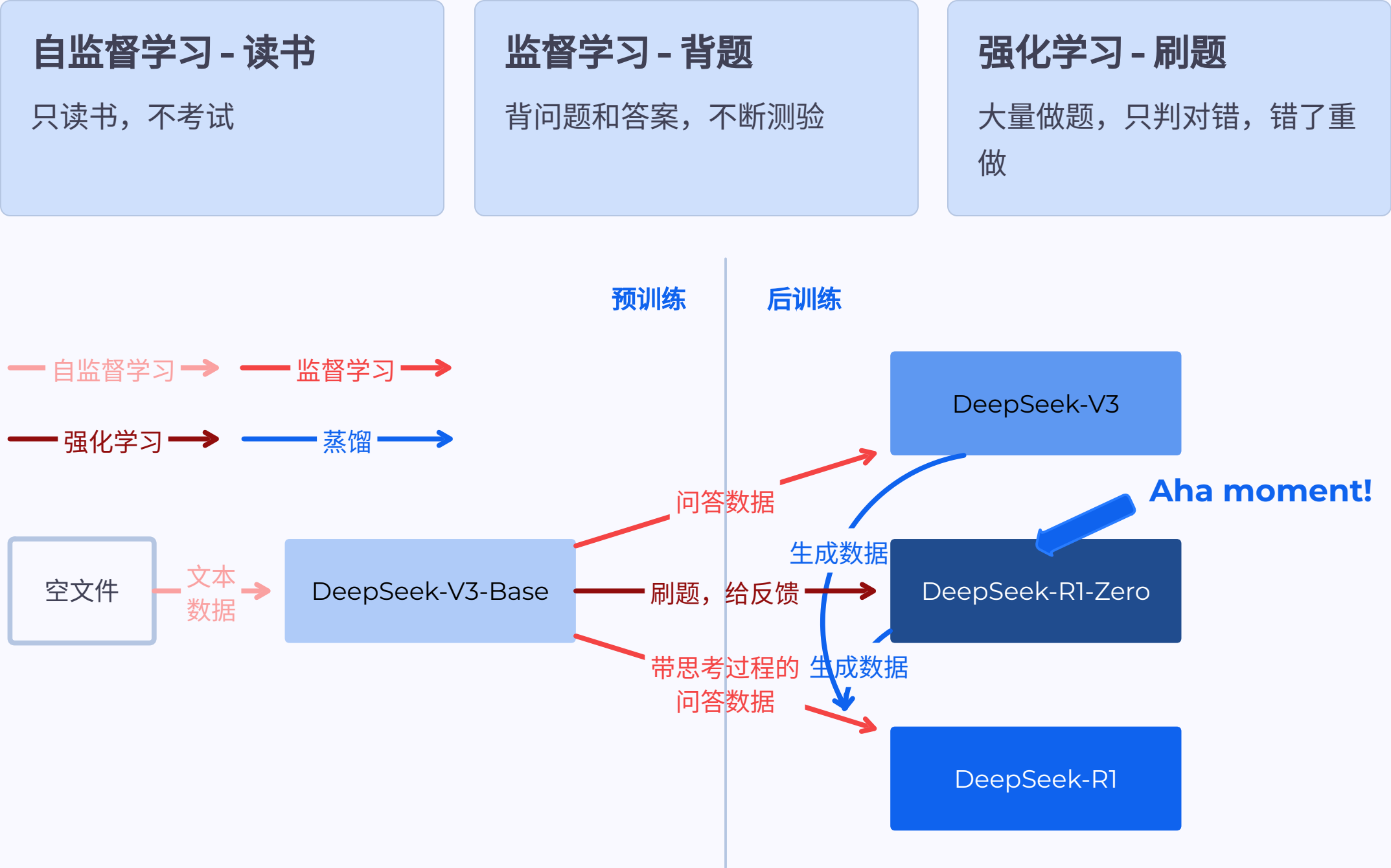
- 训练数据是几十 T 的文本、代码
- 很费钱很费钱，99% 成本花在这里
- 决定模型的基本素质

后训练

做实习，知道该怎么把知识用起来

- 训练数据主要是输入输出对
- 数据成本可能 > 训练成本
- 对模型行为、能力有决定性影响

DeepSeek 的训练过程



DeepSeek-R1-Zero 的 Aha Moment

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\dots$

$$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

$\dots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\dots$

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation: $\dots$

$\dots$

把 AI 当人看

相同

1. 基于神经网络，数学原理是相同的
2. 都是用「管理学」来合作

不同

1. 硅基和碳基的物理特性不同
2. 人类可实时训练，AI 不可以
3. 人类记忆不可被操纵，AI 可以

☑ 控制 AI 的记忆，就能控制 AI 的一切



怎样随心控制大模型

没有公式，没有代码，文科生也能懂

不是「用 AI 做事」，而是「教 AI 做事」

☑ 把 AI 当人看。如果不满意，就当孩子/学生/下属，教 ta，赋能 ta



提示工程

讲清楚



RAG

给资料

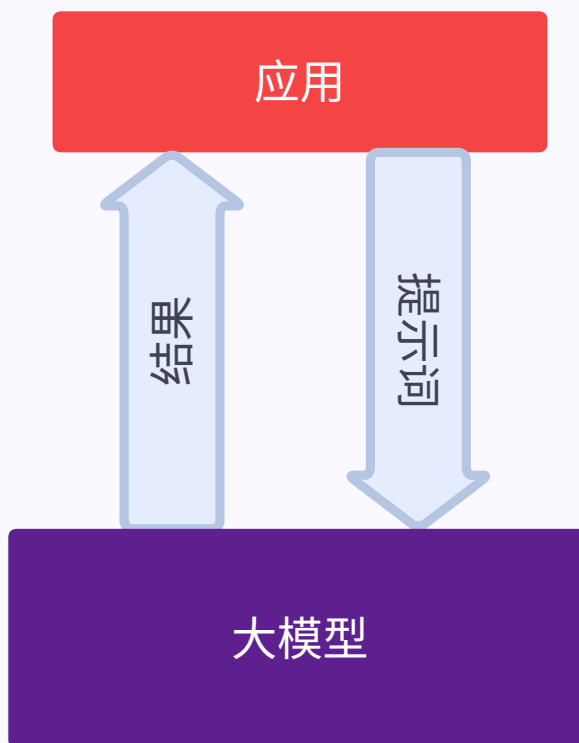


微调

刻意练习

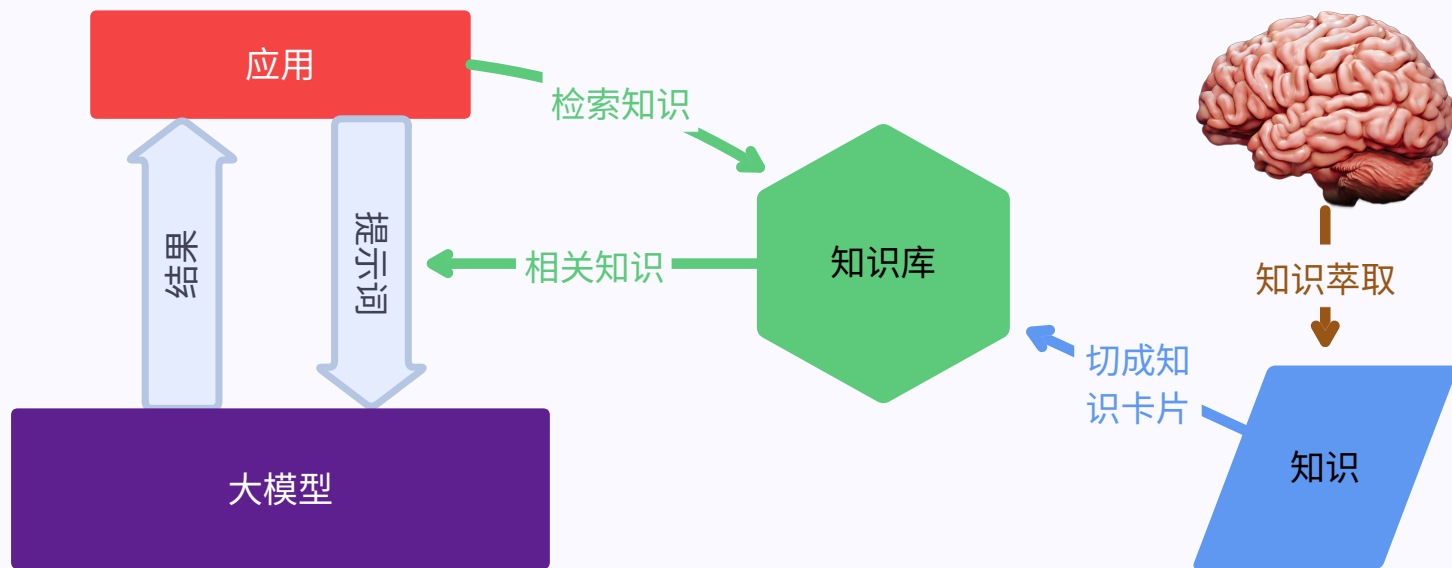
提示工程，就是手把手地教

- ① 给细节明确的指令
- 要多细，取决于需求和模型的基础能力
- 为什么思考型的大模型更厉害？因为自己思考出了更多细节



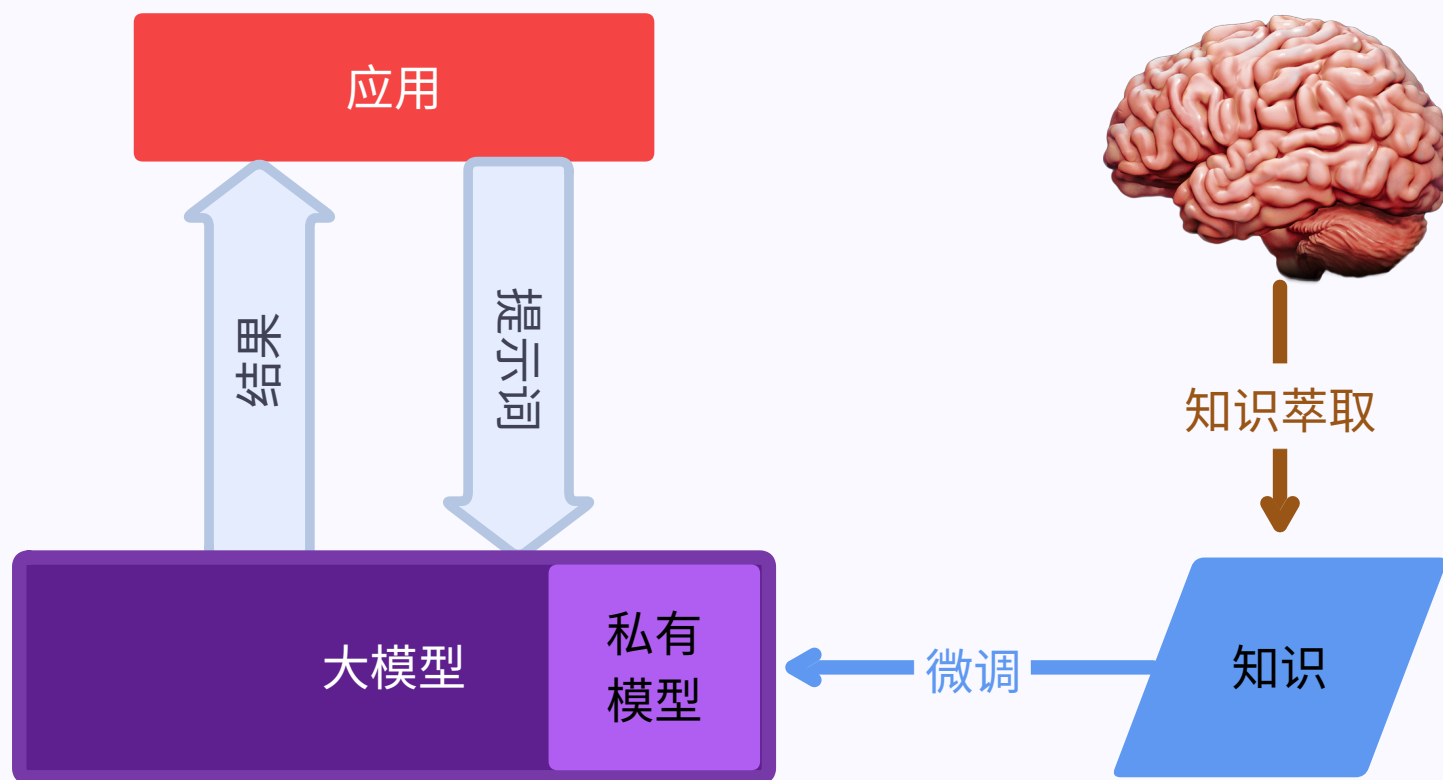
RAG，就是给资料

- ① 建立知识库，库里都是知识卡片
- 解决问题时先搜知识库，把搜到的放入提示词



微调，就是题海战术

- ① 准备大量问答数据
- 用微调程序训练模型



❗ 复杂任务，只一个人是不行的，一个 AI 也不行

workflow - 人决定 AI、工具和人的流程与分工

长文写作 workflow



智能体 - AI 决定 AI、工具和人的流程与分工

典型 Agent：夸克、钉钉AI助理、Manus、Coze 空间、Genspark、ChatGPT Operator

机械 vs. 智能

机械 - 代码和机器

精确可控，但处理不了巨复杂、模糊的事务

智能 - 人和 AI

不总对，不总听话，甚至很不听话，但能处理复杂和模糊

怎么选？

现在

机械手段占上风

接下来

智能手段会越来越多

长期

两者你中有我，我中有你

下面的困惑有答案了没？



AI 为啥懂这么多？它「知识」从哪儿来？



它是真的“思考”吗？跟人类的大脑有什么不同？



为什么同一句话有时候回答得很好，有时候答非所问？



跟 AI 聊天需要说“谢谢”吗？这会影响结果吗？



AI 会不会偷偷记住我的隐私信息？



我输入的问题会被公司用来训练别人的模型吗？



提示词怎么写才算“专业”？有万能模板吗？



为什么有时答案听上去非常自信，却可能是胡编的？



AI 会不会带有偏见或歧视？这些偏见是怎么产生的？



它会“撒谎”吗？如果回答出错，我怎么判断？



AI 真的会取代我的工作吗？哪些岗位最危险？



如果 AI 爆火之后“迭代太快”，我怎么跟得上学习？

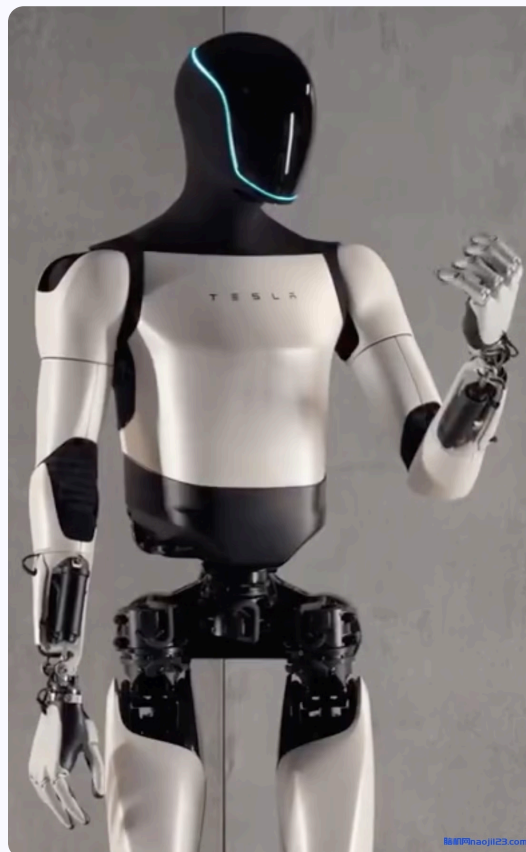
**AI 必将无处不在，
社会需要怎样的人才？**

AI 代替人的时间表

1 2030 年，AGI 到来，
代替大多数脑力劳动



2 2035 年，具身智能普及，
代替大多数体力劳动



3 未来.....



AGI 时代的人

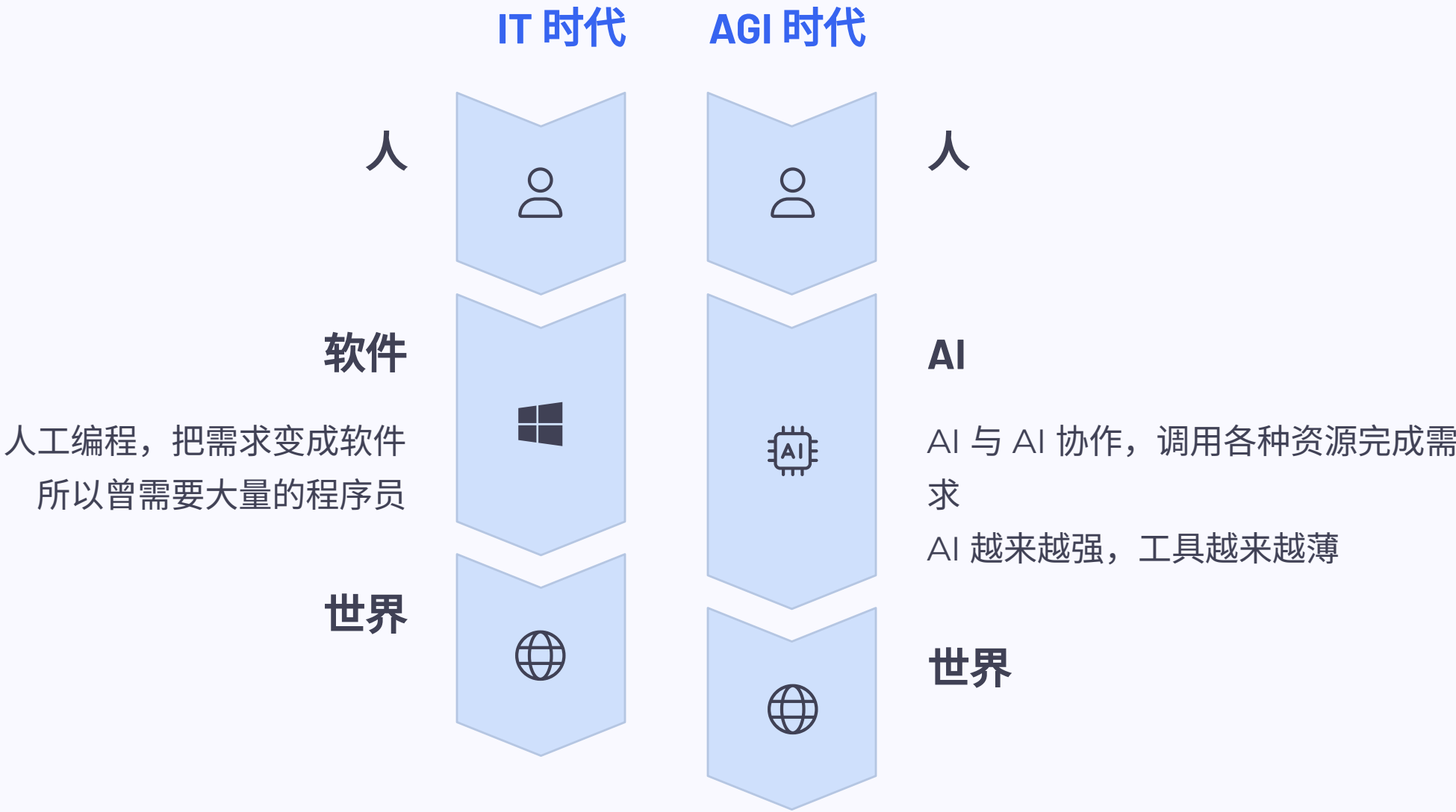
- 
- AGI + 具身智能，百倍提升生产力
 - 衣食住行成本百倍降低
 - 多数人类不需要用工作保障生活
 - 人人都是「富二代」



对 AI 的两大错误认知

✗ 使用 AI，就是使用 AI 工具

✗ 做 AI 产品，是技术高手的事



✓ 使用 AI，关键是教，是定制

✓ 做 AI 产品，技术已经不是最重要的

AI 创客的核心能力

懂业务 - 知道一个行业是怎么回事

- 技术
- 商业
- 人员
- 组织

懂 AI - 知 AI 善任

- 掌握 AI 的基本原理 - 相对稳定
- 明晰 AI 的能力边界 - 不断变化
- 会操纵 AI 达成既定目标 - 保持手感
- 设计整体解决方案 - AI 专家

✅ 通常，懂业务的人学 AI，比懂 AI 的人学业务，要容易

无处不在的 AI 智能体

扎克伯格：未来AI智能体
会达到数亿甚至数十亿
最终数量可能会比人类还多

扎克伯格

Meta CEO

我认为我们将生活在一个有数亿

人类的永恒价值？

管控 AI

教 AI 正确地做正确的事

承担责任

判断 AI 对不对

创造幸福

人类需要幸福，AI 不需要

- ✓ **AI 能力定律：**
AI 能力的上限，是使用者的判断力

最后倡议

- 1

跨过岗位边界

 - 横向扩展自己的视野、认知和判断力
 - 自己当老板，把 AI 当员工
- 2

天天用最好的 AI

 - ChatGPT o3
 - Gemini 2.5 Pro
 - Qwen 3
 - DeepSeek V3 & R1
- 3

到智能体/ workflow 平台试验想法

 - 创造 AI 应用原型
 - 无代码做更深的定制
- 4

到 AI 师傅学 AI 常识

 - ai-shifu.cn
 - 《跟 AI 学 AI 通识》
 - 免费兑换码：**eleme**
24 小时有效!
- 5

关注我的视频号

孙志岗AGI指南🌟
科技博主



扫一扫二维码，关注我的视频号
- 6

开始交流

谢谢!