九问苍穹：AI Agent的演进之路

芮勇博士关于人工智能体未来发展的深度思考

浪潮之巅的阴影：为何大模型还不够？

1. 理解能力不足

现有最强大模型读取模拟时钟，准确率仅 39%；读取日历，准确率仅 23%。暴露了模型在基础理解上的结构性局限。

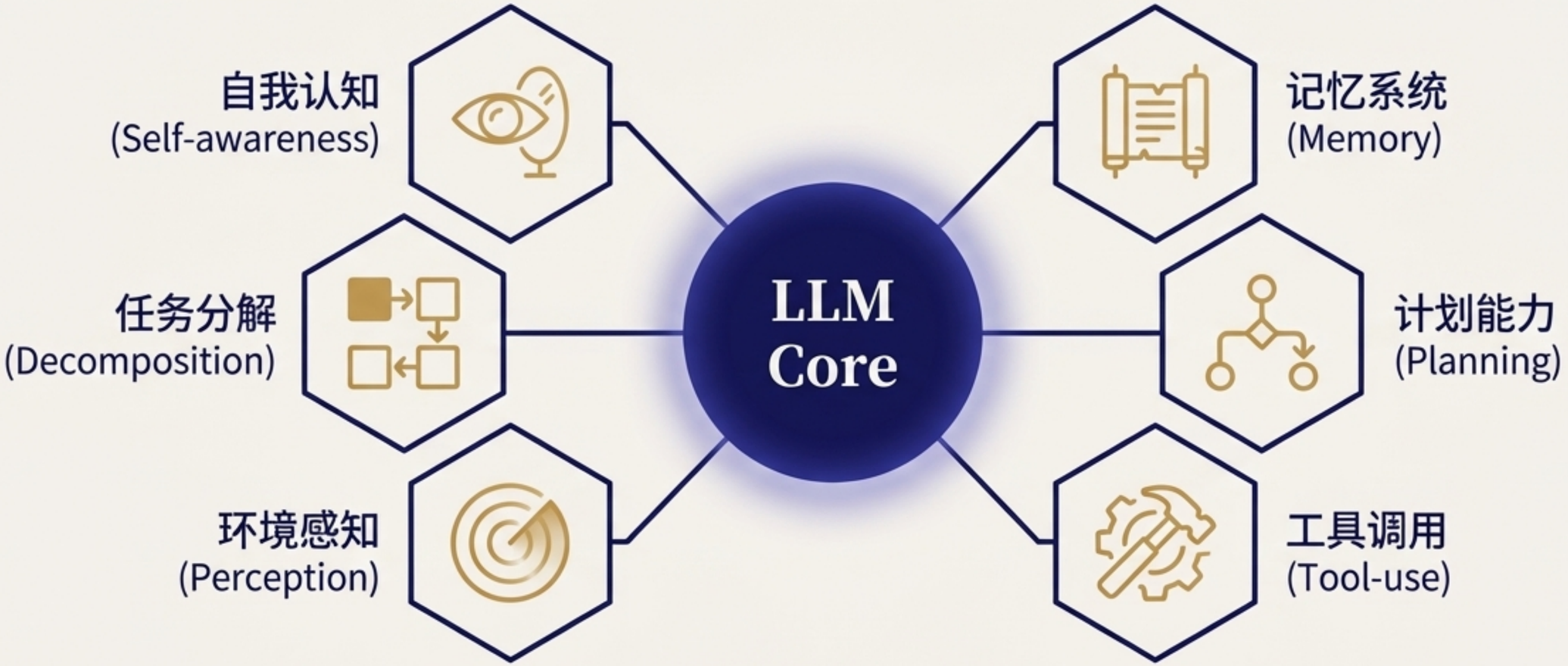
2. 存在幻觉

模型会自信地“说错话”。人类对自己知识的不确定性有意识，而大模型则看似笃定地给出错误答案。

3. 缺乏认知

缺少物理直觉、因果推理、结构化认知。模型只会“照猫画虎”，未真正理解背后机制。

Agent的必然性：为大模型装上“外挂”



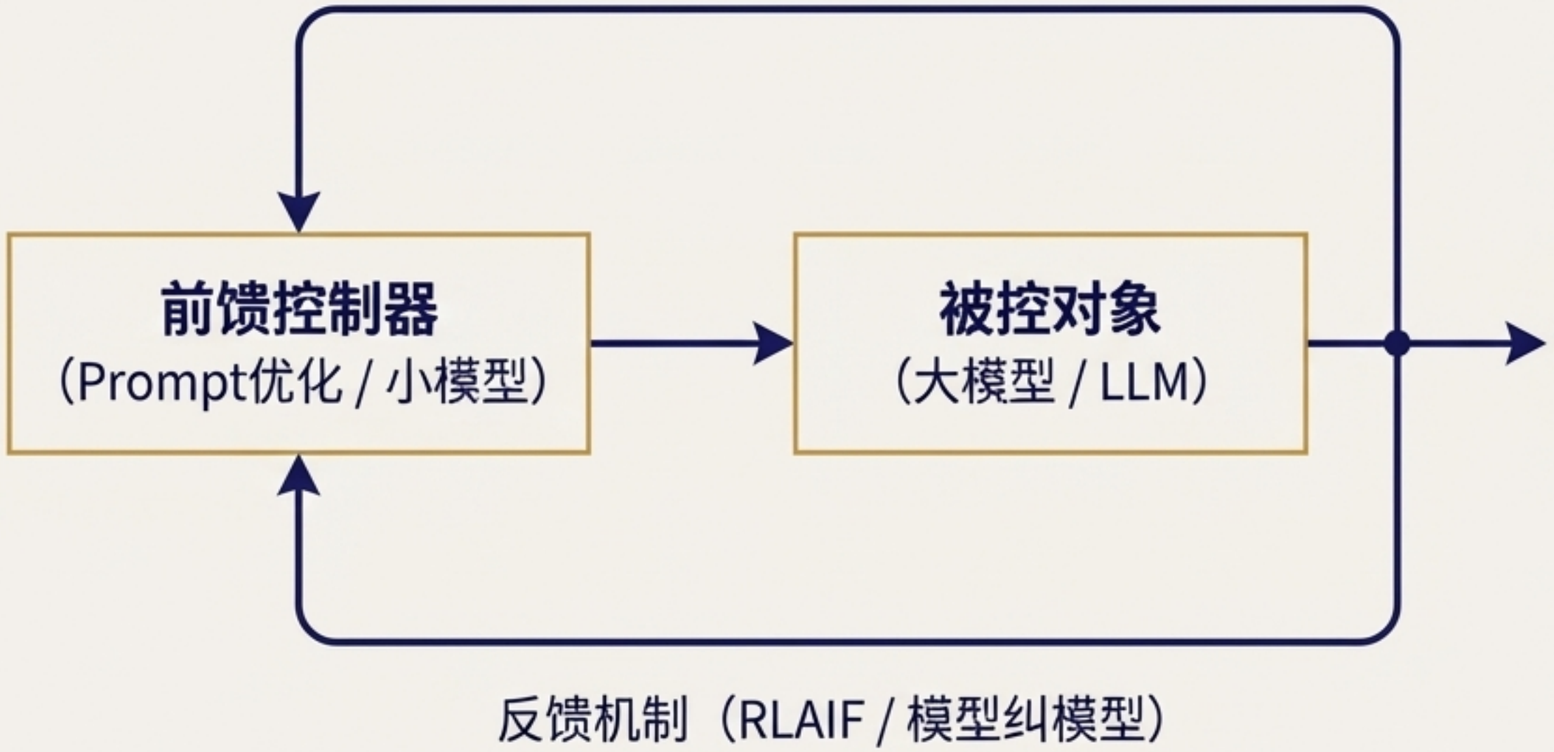
Agent的出现，正是为了解决这些“大模型做不到的问题”。

思想的基石：从经典科学中寻找答案

前三个问题，为构建可落地的AI Agent提供坚实的方法论。

第一部分：实践与落地

第一问：控制论能否启发AI Agent的设计?



AI反馈（RLAIF）的崛起

案例：OpenAI 的 CriticGPT

一个GPT-4生成代码，另一个GPT-4进行审查。

结果：错误检查效率提升60%，实现自循环反馈。

结论：控制论的闭环思想，正在启发AI Agent的结构设计。

第二问：认知心理学能否启发AI Agent设计？

工作记忆
(Working Memory)

对冗余信息进行压缩，节省资源。

长期记忆 (Long-Term Memory)
持续积累经验，迭代优化。

“学习-固化-提取”循环

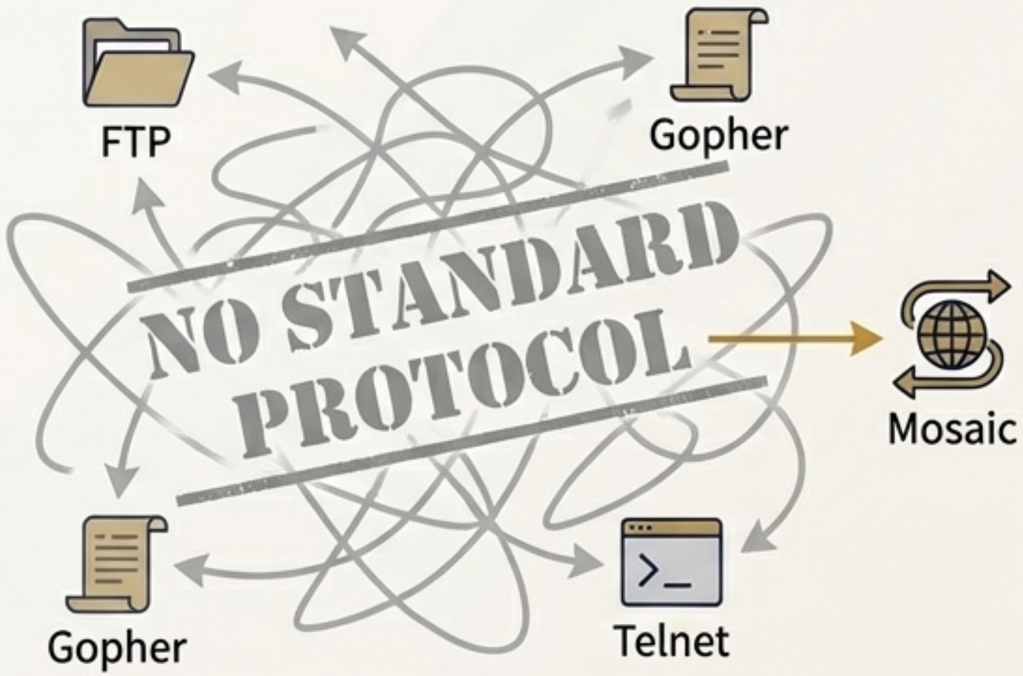
案例：Reasoning Bank

从长期记忆提取相关内容 -> 执行任务 -> 结构化总结 -> 更新回长期记忆。

结论：人类记忆的结构与流动，为构建可持续进化的Agent系统提供了核心方向。

第三问：计算机网络能否启发AI Agent的设计？

1990s



今天的Agent生态，面临着90年代互联网同样的瓶颈：**缺乏统一协议**。

Today

我们正在构建**“Agent时代的HTTP”**，这将推动生态的爆发式发展。

迷雾中的攀登：探索进行中的核心辩论

中间三个问题，是当下产业界与学术界正在激烈探讨，且尚无定论的开放性趋势。

第二部分：趋势与争鸣

第四问：语言生成能否达到人类水平的推理能力？

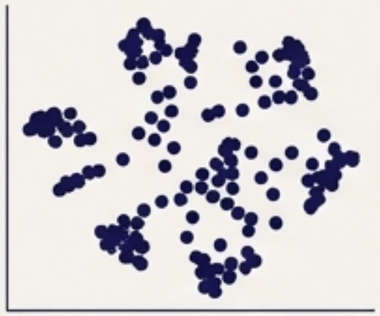
人类大脑中，语言与推理的生理基础是**分离**的。

大模型试图依托“语言区域的模拟”去实现推理功能，这在生物学上显得**不合常理**。

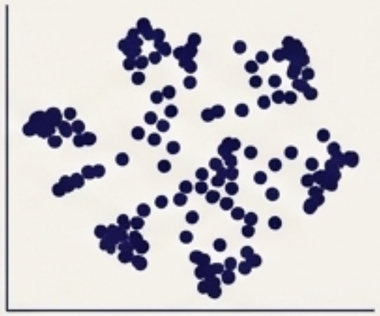
结论：这条路径能否走通，仍是一个未解之谜。

第五问: LLM和人类是否以同样方式压缩信息?

外部表征 - 高度相似



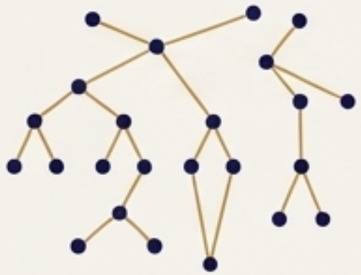
>90%一致性



相似：语义聚类与人类定义的概念类别高度一致。

内部结构 - 显著不同

人类



LLM

不同：内部表征结构与人类认知机制存在巨大差异。

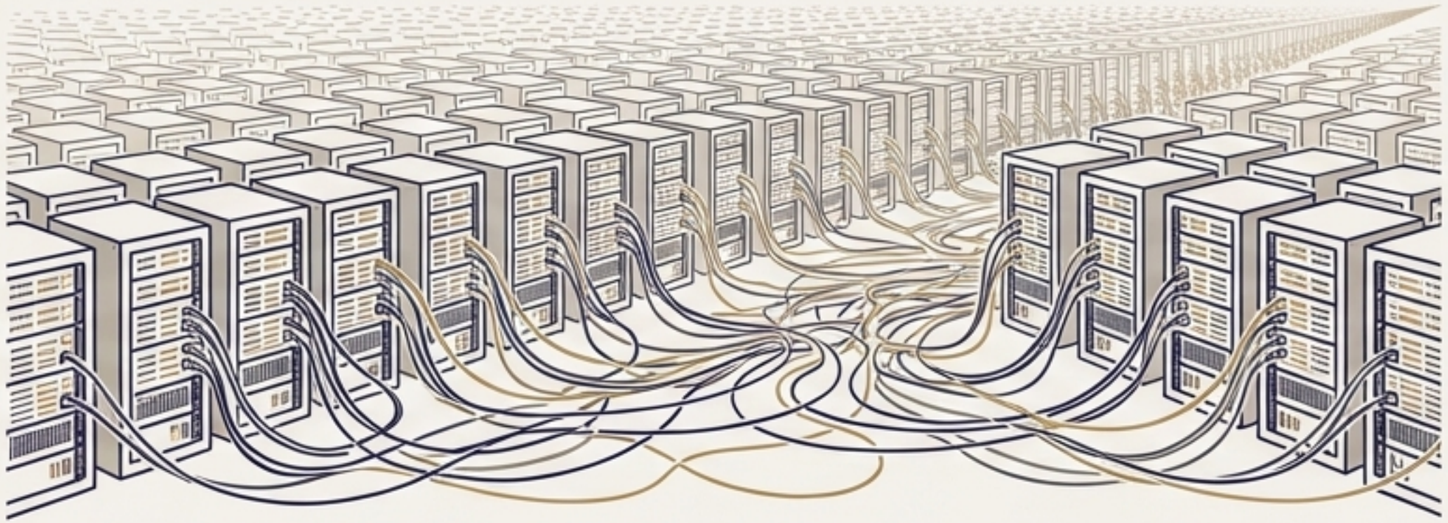
如果大模型的压缩方式与人类截然不同，我们是否仍走在通往真正智能的正确道路上?

第六问：统计学习能否实现真正的“理解”？

3

样本 (samples)

人类只需极少样本就能形成概念



1,000,000+

样本 (samples)

而深度学习需要百万级示例才能完成同样任务

依赖统计学习的范式，真的能达到真正的理解吗？这一点目前仍然没有答案。

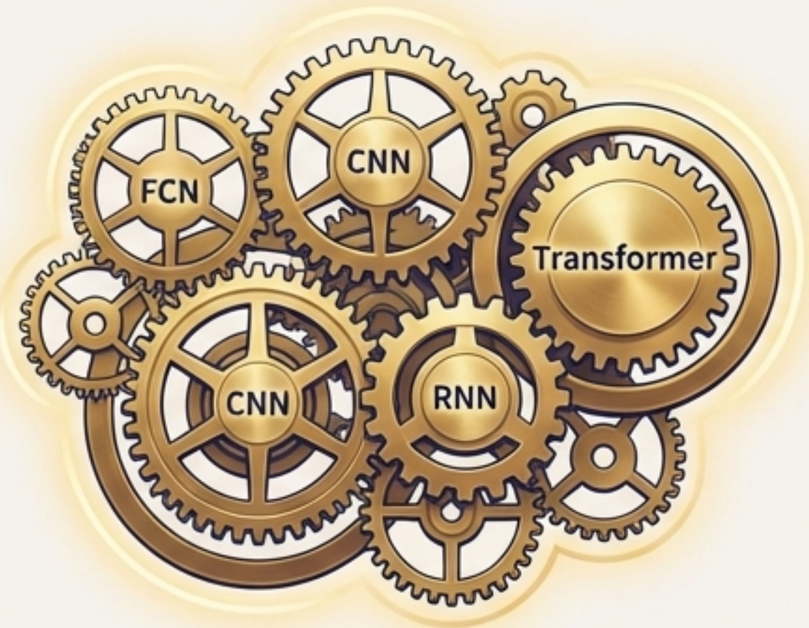
第三部分：前沿与思辨

仰望星空：决定未来的终极问题

最后三个问题，是关于AI范式根基的思辨，答案无人知晓，但它们将定义下一个时代。

第七问：LLM的Scaling Law能走多远?

结构工程 | Structure Engineering



观点一：继续优化结构、寻找更合适的拓扑，是性能提升的关键。

规模涌现 | Emergence from Scale

观点二：在数据与算力足够大时，更少结构的超大模型能自我涌现出超越复杂结构的能力。

未来智能的突破，究竟来自结构，还是来自规模?

第八问：预训练对于“快速演化”是否必要?

Richard Sutton派：On-the-fly学习

人类是‘无预训练’的，依赖大规模预训练是根本错误。

Andrej Karpathy派：捷径

AI没有漫长的演化，预训练是必要的‘快速演化捷径’。

预训练可能是必须的，但远远不够。持续学习仍然是不可替代的关键环节。

第九问：AGI是否需要新的架构?

可修补体系 | Repairable System

“或许只差一两次突破即可抵达AGI。”
— Demis Hassabis

根本性瓶颈 | Fundamental Bottleneck

“当前范式在根本上走向瓶颈，必须被彻底重建。”
— Geoffrey Hinton

我们是在完善一座已成型的大厦，还是在面对一个需要重新设计的结构？答案目前无人知晓。

回到旅程：九问的完整图景

前沿与思辨 | Q7-Q9

尚无定论的探索
（Open Exploration）

趋势与争鸣 | Q4-Q6

值得观察的趋势
（Observable Trends）

实践与落地 | Q1-Q3

可落地的方法论
（Actionable Methodologies）

在这九个问题里，
我们既看到AI Agent发展的清晰路径，
也看到通往未来的巨大不确定性。

正是这些确定与未知，
共同构成了当下最迷人的时代命题。