# 专家级 Agent 技术演进及应用

## —— 从通用到专业，跨越产业化门槛

演讲人：梁家恩 博士

云知声智能科技股份有限公司 董事长/CTO

AiCon

全球人工智能开发与应用大会

AiCon
全球人工智能开发与应用大会

# 目录

# 极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

## 4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

## 6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

## 8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

## 10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

## 12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

# 01 学以致用

—— 大模型为学，智能体致用

# AGI 进入“学以致用”阶段

大模型提供“语义与推理”的可计算框架

### 1、Scaling Law 深化

- ✓ 预训练(Pre–train)
- ✓ 后训练(Post training)
- ✓ 推理(Inference)


### 3、推理与规划 加强

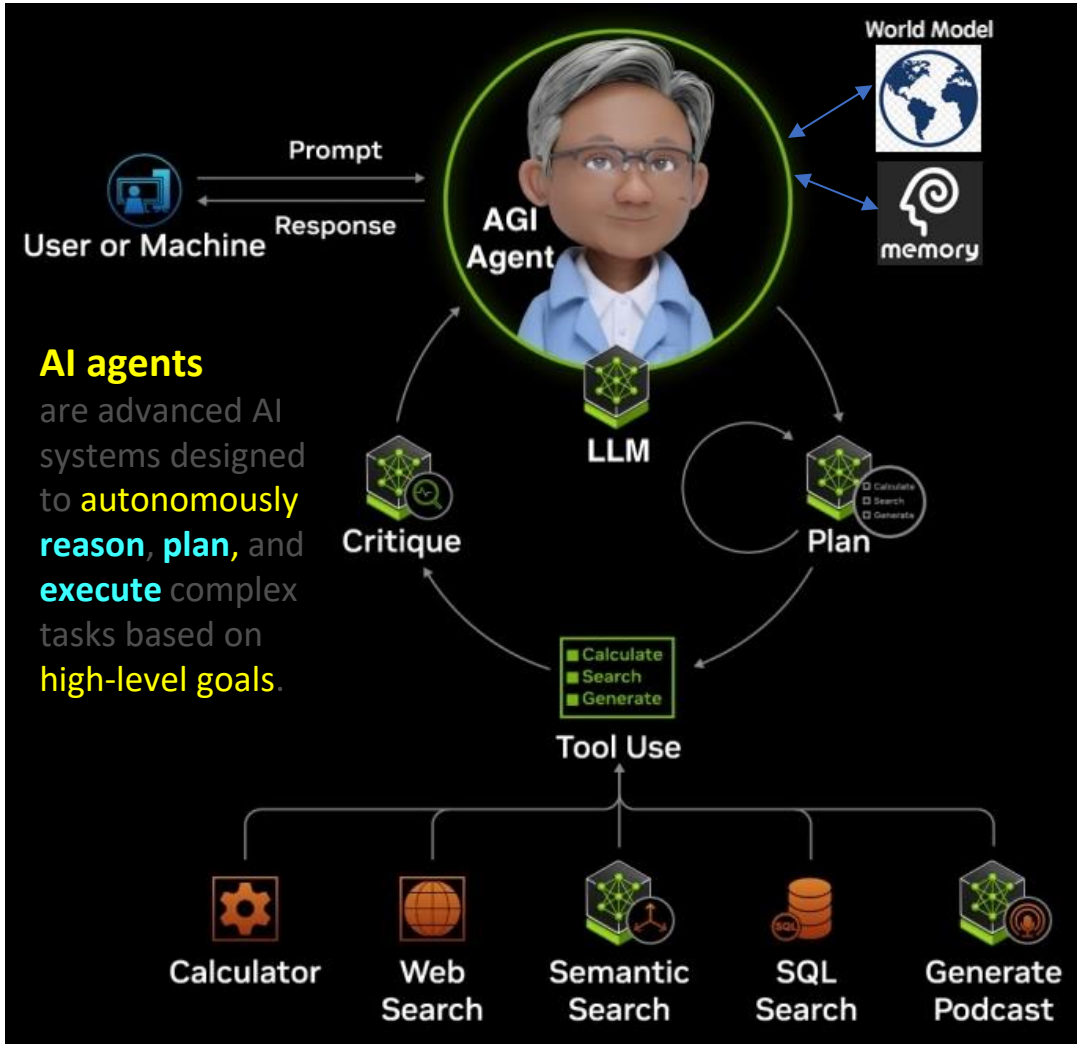
- ✓ 基础层：CoT/ToT
- ✓ 核心层：RLVR 等
- ✓ 框架层：Agent

### 2、多模态融合

- ✓ 原生多模态
- ✓ 语义对齐多模态

### 4、世界模型/具身智能

- ✓ 大脑–小脑 分层协作
- ✓ 端到端 VLA 模型
- ✓ 仿真–现实 迁移



# AGI 从“大模型能力”走向“智能体系统”竞争

大模型(LLM)是基础，也是成本投入，要真正创造价值，在于解决实际问题的**智能体**(Agent)系统。

从能力到价值

深入场景，打造“**大脑(LLM) + 五官四肢**”**智能体**(Agent)协同体系，解决实际问题，才能跨越产业化“最后一公里”。

躬身入局，解决问题

如何将通用 LLM 能力，转化为专业领域**高能力、高可靠、低成本**的专家级智能体(Agent)，是推动产业升级的关键。

突破瓶颈，升级产业

# 智能体创造价值的核心挑战

1、实用，解决实际问题，引发效率质变，突破能力上限

2、可靠，克服“幻觉”——不合逻辑或时宜的“涌现”

3、普惠，成本可控，可规模化应用

# 02 专家级 Agent 技术架构

—— 通用为基，专业破局

# 2. 智能体演进三模式

智能体打造关键：

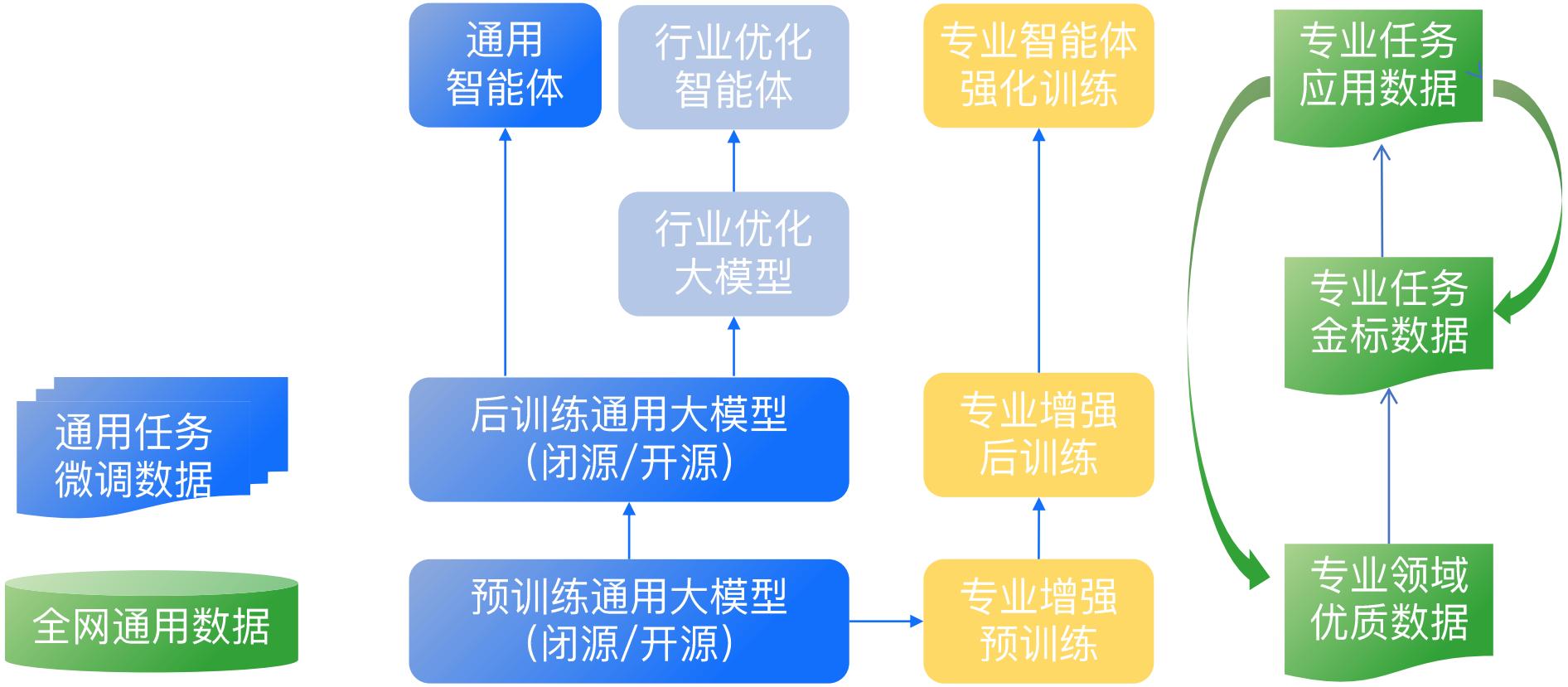
1、完整技术能力

2、优质数据闭环

3、场景应用优化

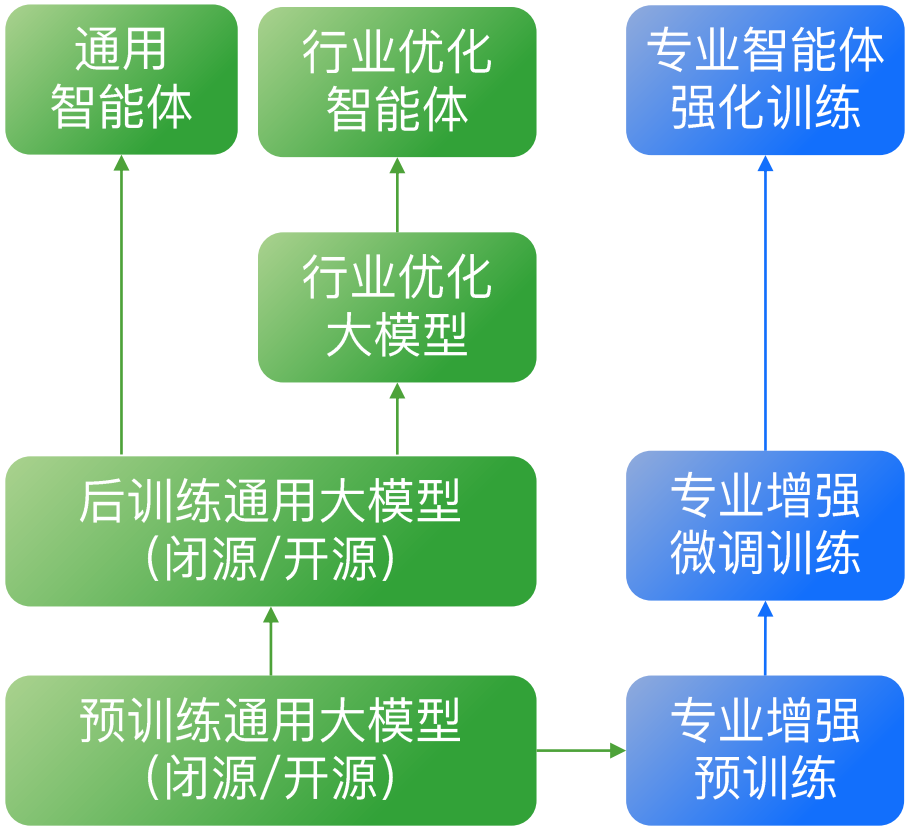
4、工程成本优化

5、高效算力平台



# 2.1 打造专业Agent模型

**需要在不同训练阶段注入不同知识**




| 能力扩充 | 数据域 | 训练阶段 |
| --- | --- | --- |
| 内化记忆能力 | 1. 专业任务金标数据<br>2. 专业任务应用数据 | 智能体强化训练 |
| 增强工具调用能力 | 1. 专业任务金标数据<br>2. 专业任务应用数据 | |
| 强化任务规划能力 | 1. 专业任务金标数据<br>2. 专业任务应用数据 | |
| 丰富感知输入能力 | 1. 专业领域优质数据<br>2. 专业任务金标数据 | 增量预训练+微调训练 |

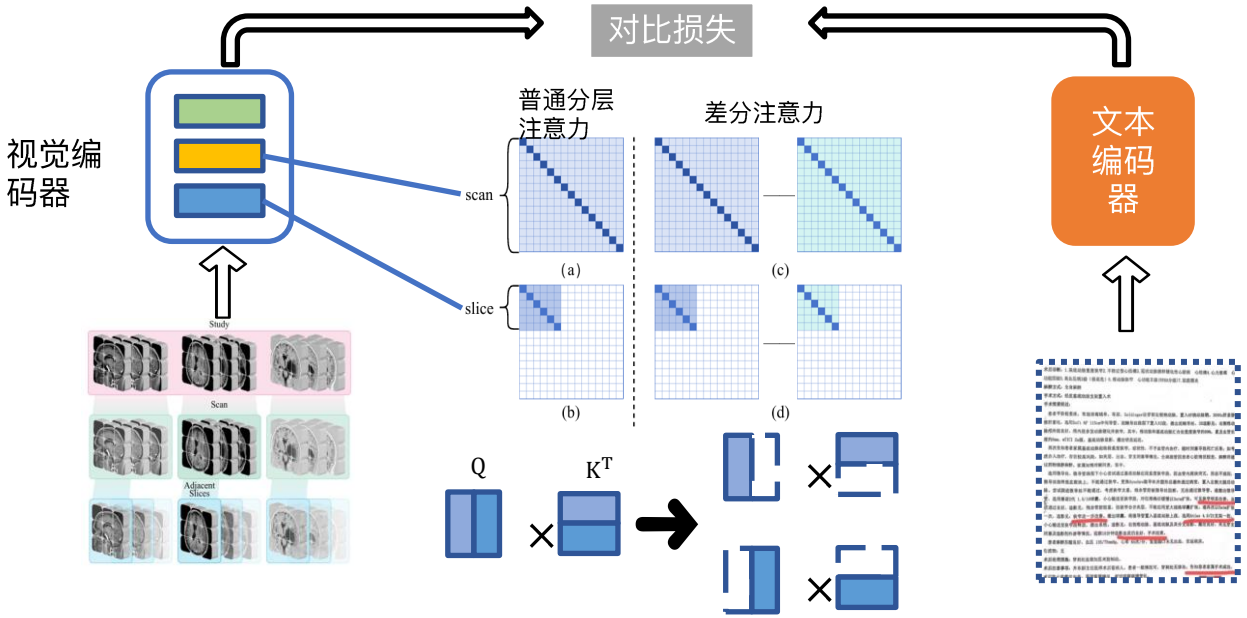
# 2.1.1 专业基座增强：专业数据扩充+模态扩充

## 通用增量训练

1. 现有通用基模，专业知识会被海量通用语料“稀释”，现有模型没有经过高质量专业数据训练

2. 现有基模基本是文本数据，部分多模态模型更多建模的是自然图像或者闲聊语音，对比专业领域比如医学影像、工程图纸、医学语音等存在明显的分布差异，所以现有的对齐策略不完全适用

基于行业多模态数据特点，
定制化增量训练策略和多模态对齐策略




基于差分注意力的视觉文本模态对齐算法

# 2.1.2 专业智能体强化训练：任务规划能力

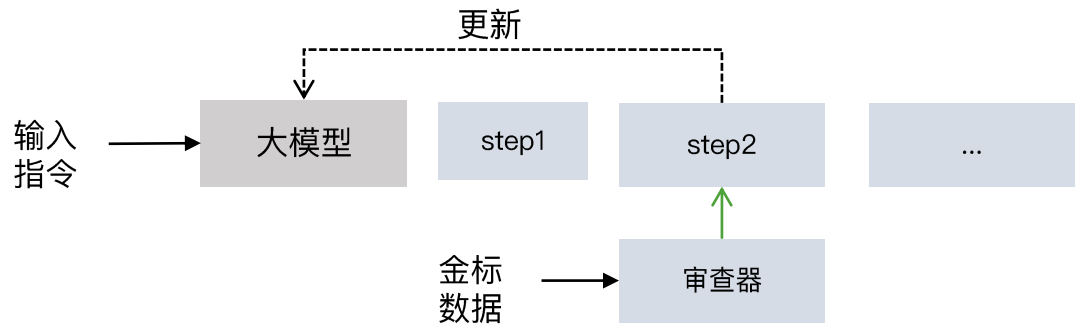
别人走向“聪明”，我们走向“可信”。

### 现有训练方案

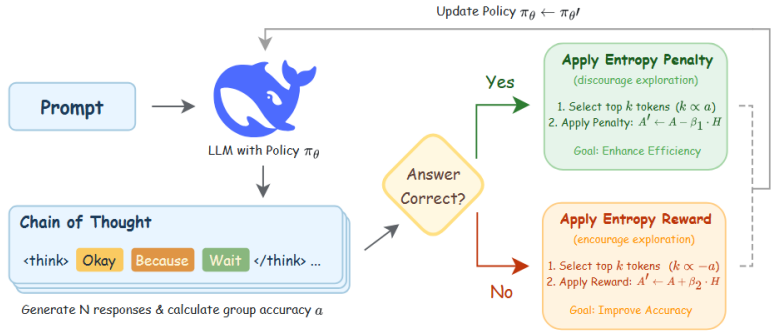
1. 通用模型在数学、代码任务上追求形式逻辑上的绝对正确，而真实专业场景在过程中充满不确定性

2. 专业规划数据量少，知识密度高，现有算法样本学习效率低下，利用不够充分

RLVR + 行业“审查”器，增强在不确定性规划上的可用性




通过条件熵针对性强化专业任务规划能力，完成少量数据的高效学习



# 2.1.3专业智能体强化训练：工具调用能力

现有工具调用方案

1. **交互模式**：以单轮生成结果为主，根据结果做只做一次反馈

2. **工具箱**：大量无法使用的通用工具集，缺乏专业工具箱来应对专业问题

通用工具：模型通过单次调用工具结果反馈
专家级工具：在动态环境中持续多轮交互反馈

通用Tools → 专家级Tools

单次调用工具反馈

多轮环境交互行为数据

通用知识　　隐性专业知识

**通用工具**

计算器　网页查询　天气查询　地点查询　机器翻译

专家级能力沉淀

**医学专家级工具**

诊断术语标准化　医学命名实体识别　诊断一致性检查　用药合理性检查　诊断自动编码

# 2.1.4 专业智能体强化训练：任务相关记忆优化

利用强化学习训练使得模型可以动态地、自主地控制这些结构化记忆的构建、演化和修剪

## 通用智能体的记忆

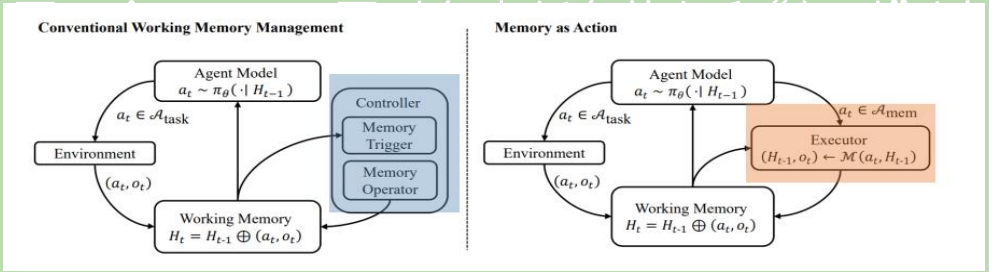
1.1 短期记忆依赖上下文工程，基本是扁平的token序列数据

1.2 对于长期记忆：用传统的文档库，普遍召回效率差，幻觉高

## 知识萃取：无结构化知识的降噪提纯

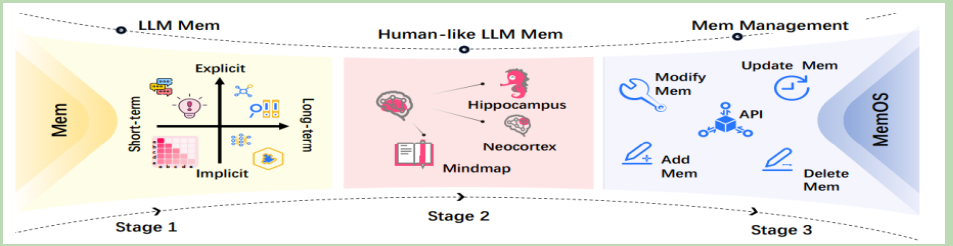

## 短期记忆内化：萃取精炼相关信息

历史精炼和萃取结构化数据作为上下文

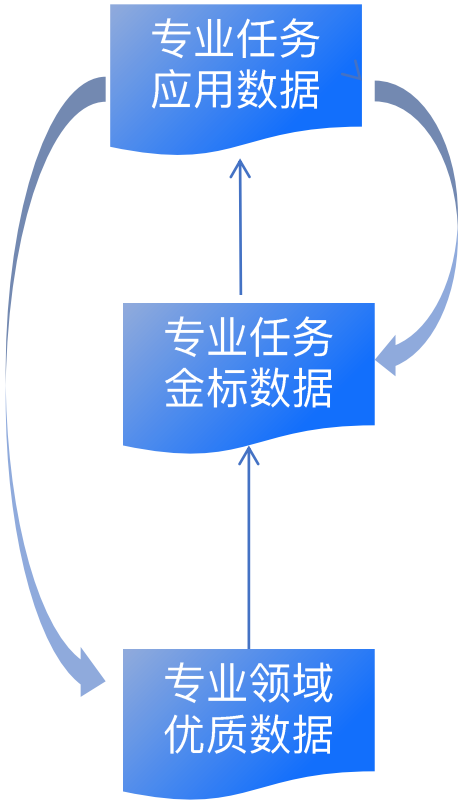


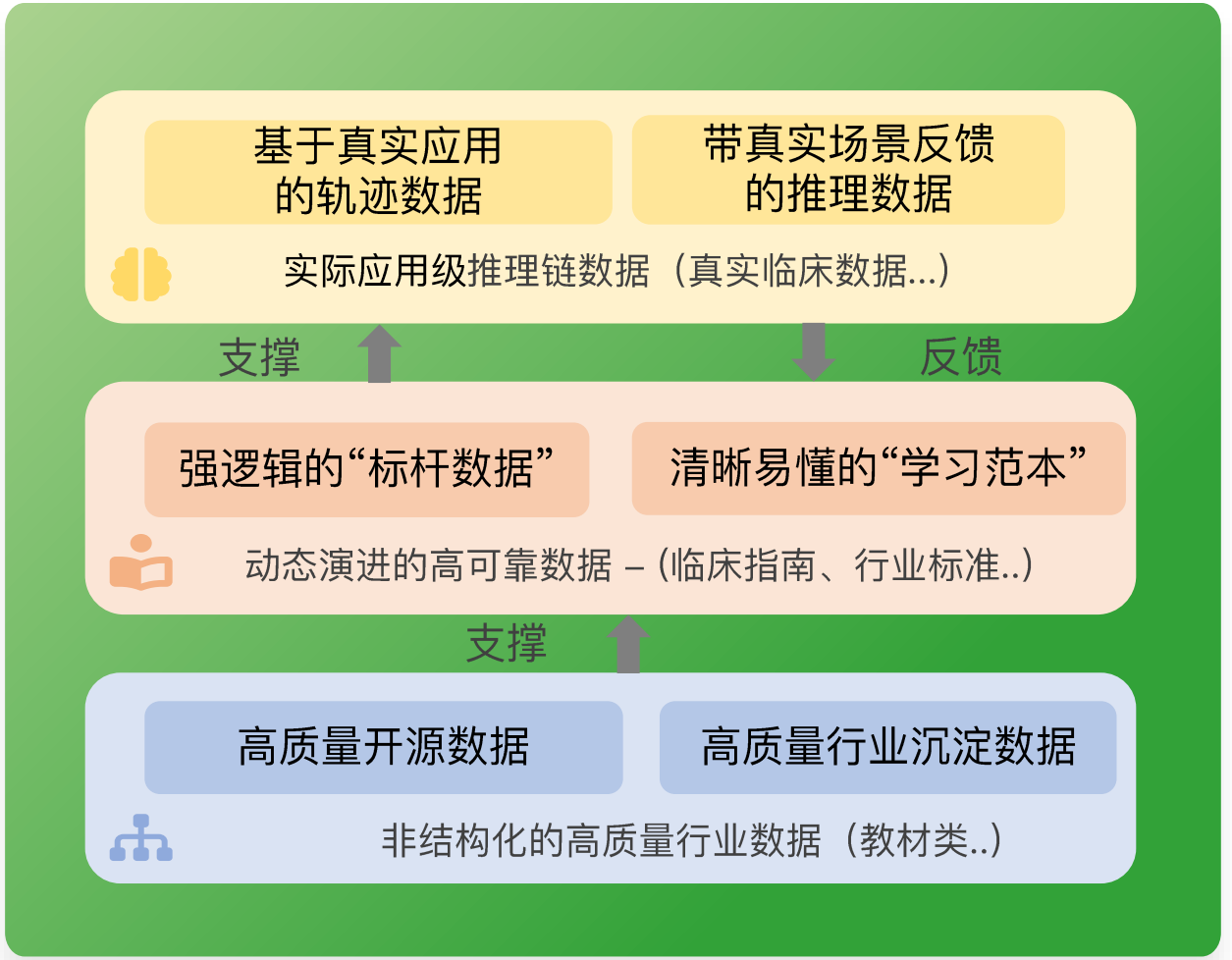

## 长期记忆内化：动态加载可靠信息

根据信息的重要性和稳定性决定注入哪些信息

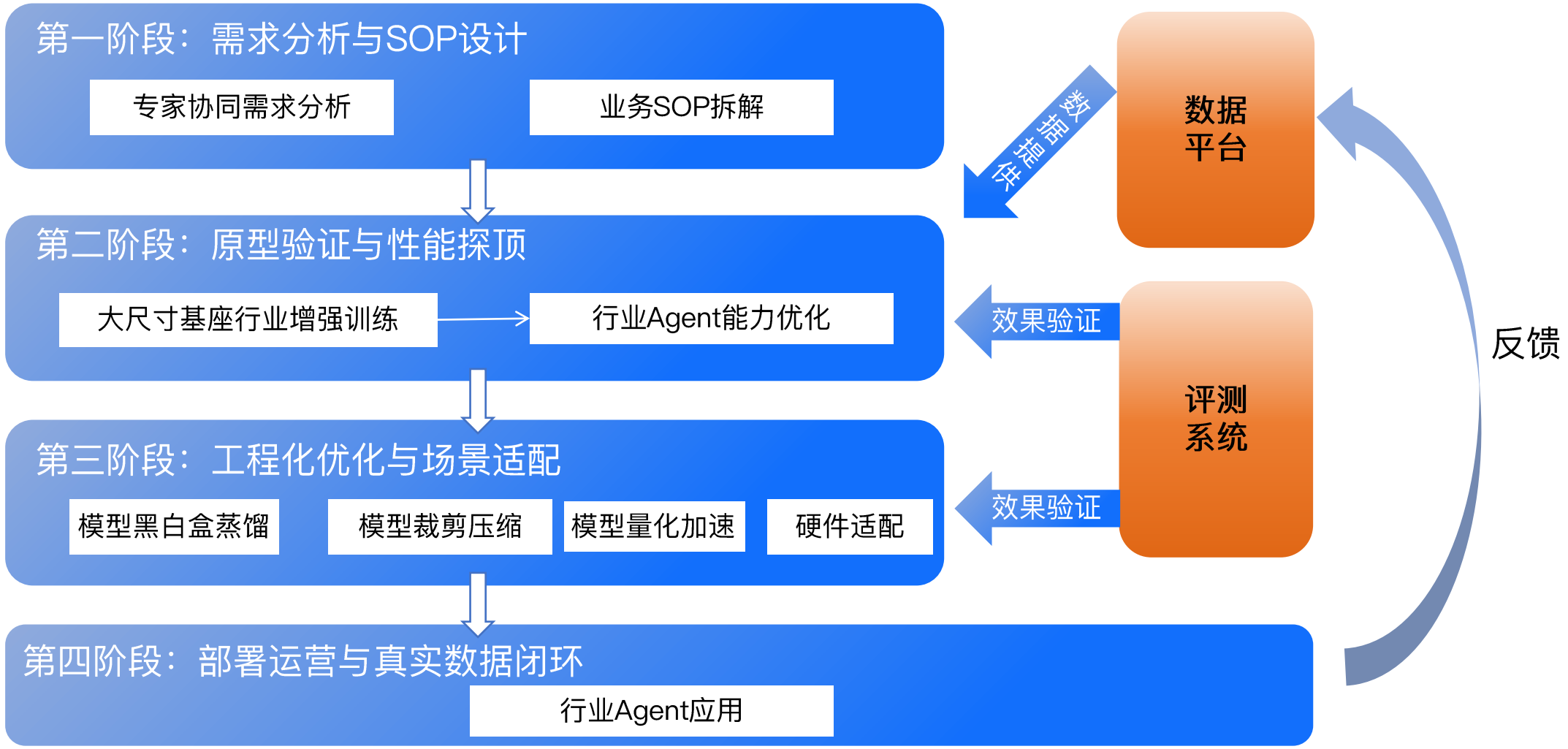


# 专业任务数据构建闭环


专业任务
应用数据
专业任务
金标数据
专业领域
优质数据


基于真实应用
的轨迹数据
带真实场景反馈
的推理数据
实际应用级推理链数据（真实临床数据...）
支撑
反馈
强逻辑的“标杆数据”
清晰易懂的“学习范本”
动态演进的高可靠数据 –（临床指南、行业标准..）
支撑
高质量开源数据
高质量行业沉淀数据
非结构化的高质量行业数据（教材类..）

# 工程化整合


第一阶段：需求分析与SOP设计
专家协同需求分析
业务SOP拆解
第二阶段：原型验证与性能探顶
大尺寸基座行业增强训练
行业Agent能力优化
第三阶段：工程化优化与场景适配
模型黑白盒蒸馏
模型裁剪压缩
模型量化加速
硬件适配
第四阶段：部署运营与真实数据闭环
行业Agent应用
数据提供
数据平台
效果验证
效果验证
评测系统
反馈

# 03 实践出真知

## ——跨越产业化门槛

# AGI 技术产业化格局

| 产业层次 | 核心要素丨焦点 | 主要挑战 |
| --- | --- | --- |
| 基础设施层 | 算力芯片、智算集群、数据服务、电力网络<br>NVIDIA、AMD、Google，华为、寒武纪等 | 1.算力瓶颈：中国算力规模约为先进国家的15%,高端芯片（如GPU）供给受制于人 |
| | | 2.能源消耗：超大规模智算集群部署面临能源瓶颈，需推进“算电协同” |
| | | 3.生态依赖：国产软硬件全栈技术生态仍在构建中，需突破CUDA等固有生态壁垒 |
| 模型层 | 基础大模型、多模态、世界模型、开源生态<br>OpenAI、Google、Meta，DeepSeek，千问、豆包、混元等 | 1.技术代差：在视频生成、多模态、世界模型等前沿领域与国际顶尖水平尚有差距 |
| | | 2.数据制约：高质量、专业化数据集缺乏，数据流通和治理体系不完善 |
| | | 3.可信与可靠：模型存在“黑箱”不可解释性，引发安全、隐私和伦理担忧 |
| 中间层/工具层<br>专业化 | AI 智能体、开发工具链、行业模型微调<br>Anthropic 的 MCP 协议、各厂商智能体平台<br>行业模型专业化、Prompt 工程、RAG 技 术等 | 1.技术复杂性：智能体在复杂环境中的感知、规划与可靠执行仍是技术难点 |
| | | 2.标准化缺失：工具调用、交互协议等标准尚未统一，开发效率与interoperability受影响 |
| | | 3.落地门槛：将大模型能力与特定行业知识、业务流程深度结合的成本高、难度大 |
| 行业应用层 | 千行百业智能化解决方案、AI 硬件等：<br>• 金融：智能投顾、风险控制<br>• 工业：智能工艺设计、能耗优化<br>• 医疗：AI辅助诊断、药物研发<br>• 汽车：AI座舱、端到端自动驾驶<br>• 机器人：具身智能 | 1.场景深度：应用从“单点尝试”迈向“全局重构”,对业务理解和技术适配要求极高 |
| | | 2.人才缺口：兼具AI技术和行业知识的复合型高端人才严重短缺 |
| | | 3.投入产出评估：部分场景的商业模式需持续探索，难以快速验证商业价值并规模化 |

# 云知声 AGI 技术及产业化布局


## 2012 创立入局

- 语音交互+深度学习+智慧物联
- 2014“云端芯”一体化战略
- 福布斯成长最快科技企业

## 2016 全栈升级

- Atlas 平台+知识图谱+智慧医疗
- 吴文俊、北京市科技进步一等奖等
- 2018 年起入选全球 AI 独角兽

## 2022 AGI 升级

- 山海大模型：通用一流+医疗顶尖
- 北京市大模型伙伴，十大应用案例
- 2025 港交所 AGI 第一股上市

# 聚焦技术产业化的四大支柱

**“高素质”：通用大模型基座**

覆盖广度：见多识广，能言善辩，能思善学
顶级院校本科生，高素质、高潜力

**“进化力”：Atlas 基础设施**

高效训练：动态调度扩展，高效模型迭代
数据飞轮：技术优势转化为数据与模型壁垒


**“高水准”：专业级大模型**

攻克深度：做得到、做得好、低幻觉
百万年薪专家水准，引领行业突破与变革

**“低成本”：端侧芯片优化**

端云协同：端侧大模型与芯片级优化
规模化应用：AGI普惠，深入千行百业

# 智慧医疗实践：从效率工具到决策支持

三医协同管理
- 态势感知
- 医疗管理
- 医保管理
- 医药监管

高端医疗提升（三甲医院） ↔ 普惠医疗升级（基层医疗）

智慧医院 SaaS 平台
- 山海智枢 → 山海智医

大健康产业
- 医药器械
- 跨境医疗
- 高端康养
- AI 出海


智能病历质控


医疗质量监管


医保支付管理



居家养老智能健康监测及数字伴侣

**智慧医疗体系：由浅入深**

- 语音电子病历、病历自动生成
- 医院智能化：导医/分诊/随访/辅诊…
- 病历质控、医疗质量监管
- 医保控费、商业保险
- 数字医生：助手/同事/专家/导师
- 三医协同：医疗/医保/医药

*技术驱动，应用牵引，闭环迭代！*

# 智能体实践：如何从幻觉角度选择场景

## 幻觉率

大模型能力成熟度

- 文本理解
- 文本生成
- 医疗知识
- 临床推理
- Agent能力（规划，记忆，反思，工具使用等）

## 幻觉
## 容忍度

应用场景能否容错

- 大模型的鲁棒性
- 大模型的可靠性

## 幻觉
## 可检测性

用户能否识别出幻觉

- 大模型输出的可溯源性
- 大模型输出的校验成本

**理想的应用场景： 幻觉率低，幻觉容忍度高，幻觉可检测性高**

# 医疗智能体应用场景分析(示意性)

| 典型场景分析 | 幻觉率 | 幻觉容忍度 | 幻觉可检测性 |
| --- | --- | --- | --- |
| AI 健康咨询（非诊断） | 低 | 较高 | 低 |
| AI 问诊（面向患者） | 较高 | 低 | 低 |
| 辅助诊断（面向高年资医生） | 较高 | 较高 | 高 |
| 辅助诊断（面向低年资或基层医生） | 较高 | 较低 | 低 |
| 病历生成（书写辅助） | 低 | 较高 | 高 |
| 病历质控 | 低 | 较高 | 高 |
| 医保审核 | 低 | 较低 | 高 |

# 案例1：门诊病历生成系统

利用AI大模型可快速生成符合模版规范逻辑清晰、内容表达丰富、易读性好、可解释性强的结构化医疗文书。


在门诊场景下，医生对患者进行问诊并查体，给出初步诊断和处理方案，这些内容需要录入门诊电子病历，但该工作会挤占宝贵的门诊时间。门诊病历生成系统基于医患对话，自动生成门诊病历，大幅提升了该项工作的质量和效率。

## 应用流程


## 功能特点

### 对话识别

门诊医患对话语音实时采集，并识别转写

### 角色分离

采用麦克风定向+声纹识别+大模型语义级理解进行角色分离

### 信息摘要

基于山海大模型对医患对话文本进行分析、理解、标化，从而完成摘要

### 病历撰写

结合《电子病历书写基本规范》，应用山海大模型的文本生成能力，自动生成满足要求的门诊电子病历

### 实时质控

门诊同时进行问诊质控，及时提醒问诊遗漏

# 案例2：智能病历质控系统

病历环节质控 | 病历终末质控 | 病案首页质控 | 门诊病历质控 | 单病种质控

1. **质控点完整覆盖**（内涵质控、重点专业质控）
2. 对国家上报病种、填报明细、上报要求**采取同步处理机制**，动态满足最新要求
3. **多模态医疗信息统一解析与关联分析**，实现交叉验证，自动发现潜在的不一致与异常
4. 病历审查覆盖度**从2-5%提升至100%**
5. 检出缺陷的**查准率达90%+，查全率达85%**

# 案例3：智能医保审核系统

云知声
股票代码:09678.HK

## 医院端：智能提醒客户端

嵌入HIS系统，诊疗过程中提醒医生可能违规的收费项目，从源头避免违规情况的发生



## 医保监管端：智能监管平台

提供多维度的统计分析图表，层层下钻，精准掌握医院运营和医保违规情况


医保自动审核控费率较当前人工抽审方式提升4倍至8%以上，即每100元医保支出可节约8元。

# 产业化实战分享

## 1、战略布局

- ✓ 技术与产业趋势
- ✓ 赛道与路径选择

## 2、应用导向

- ✓ 应用场景与价值
- ✓ 研发与复制成本
- ✓ 壁垒与风险

## 3、技术驱动

- ✓ 关键点与能力边界
- ✓ 数据闭环构建
- ✓ 成本优化

## 4、组织优化

- ✓ 架构、流程、规范
- ✓ 价值观与文化

# 04 展望未来

——认知与产业升级

# AGI 三维度拓展


## 工程维度

1、需求提炼，系统集成与优化，解决实际问题
2、优质高效数据闭环构建，持续高效迭代
3、工程优化及成本控制，支持规模化应用

## 产业维度

1、能力输出到价值创造，人机协同机制
2、可信数据平台构建，数据资源参与产业分配
3、标准规范及生态构建，治理体系完善

## 科学维度

1、原创算法提升能力上限，完善模型能力
2、从“炼丹术”到“化学”，提升可解释、可控和可靠性
3、AI4S 探索新知识，新一轮认知与哲学革命

# 迎接超级智能与人机协同时代到来


**这是最好的时代，也是最坏的时代，谨慎乐观推进：**

1、当确定目标的有效可计算框架取得突破，机器超越人类只是时间问题，超级智能的出现不可避免

2、机器要在与人互动中成长，人类会拥有智能伙伴，并在互动中重新调整定位与模式

3、复杂世界并非单一目标，平衡内在矛盾与非理性因素，仍需人类参与并负责

4、机器有不同于人的“意识”，人类认知也将迎来变革

# 极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

## 4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

## 6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

## 8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

## 10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

## 12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

InfoQ 极客传媒

# THANKS

## 探索 AI 应用边界

Explore the limits of AI applications

AiCon

全球人工智能开发与应用大会