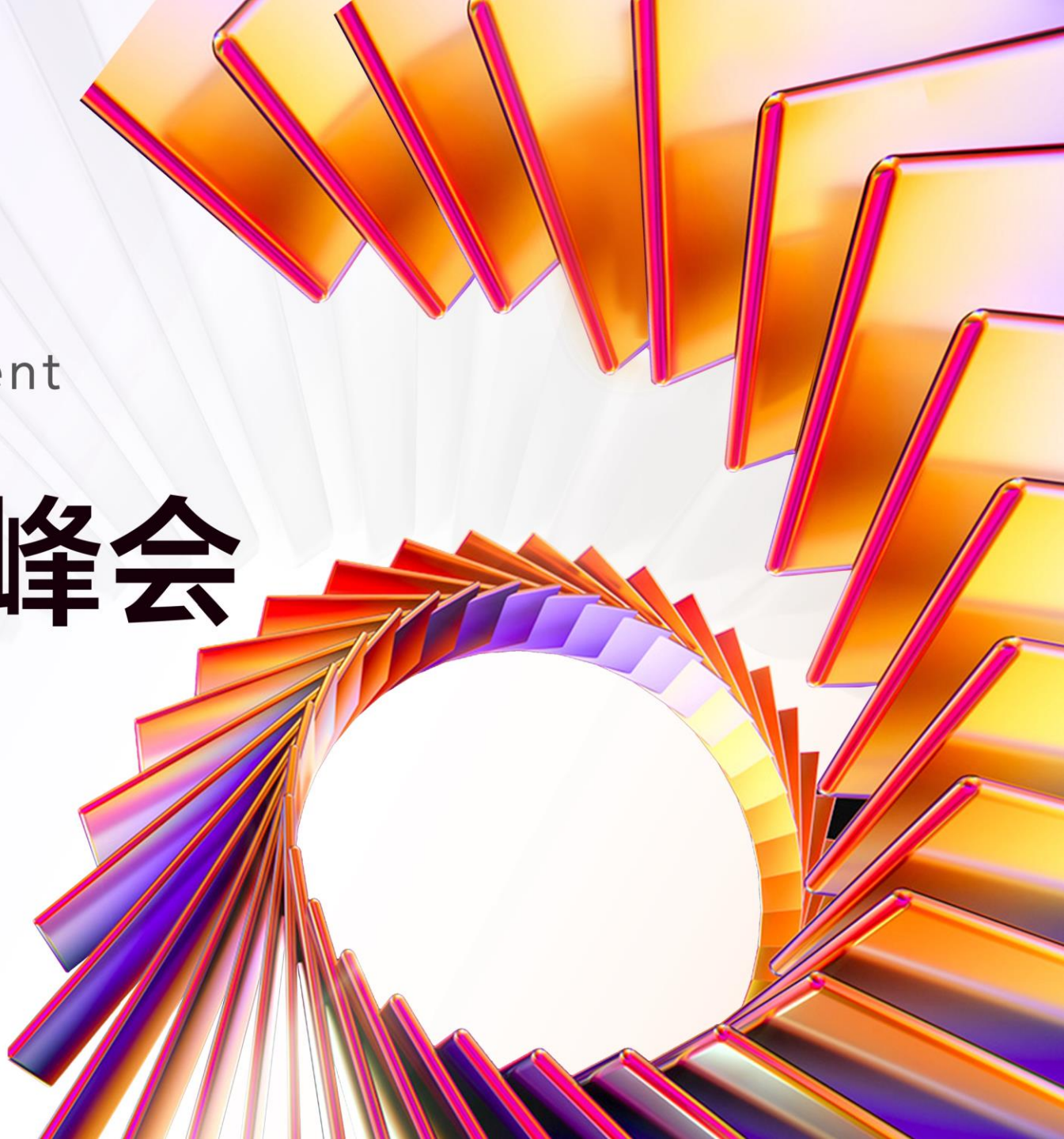




第7届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发

8月8-9日 | 北京站





第8届AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

下一站预告

11/14-15 | 深圳站

12/19-20 | 上海站



查看会议详情

深圳站论坛设置

智能装备与机器人

超越“编程 Copilot”

下一代知识工程

智能网联与汽车智能化

AI 测试工具开发与应用

AI 基础设施和运维

数据智能及其行业应用

可信 AI 安全工程

大模型和 AI 应用评测

多 Agent 协同框架

从智能测试到自主测试

大模型推理优化

多模态 LLM 训练与应用

智能化 DevOps 流水线

上下文工程

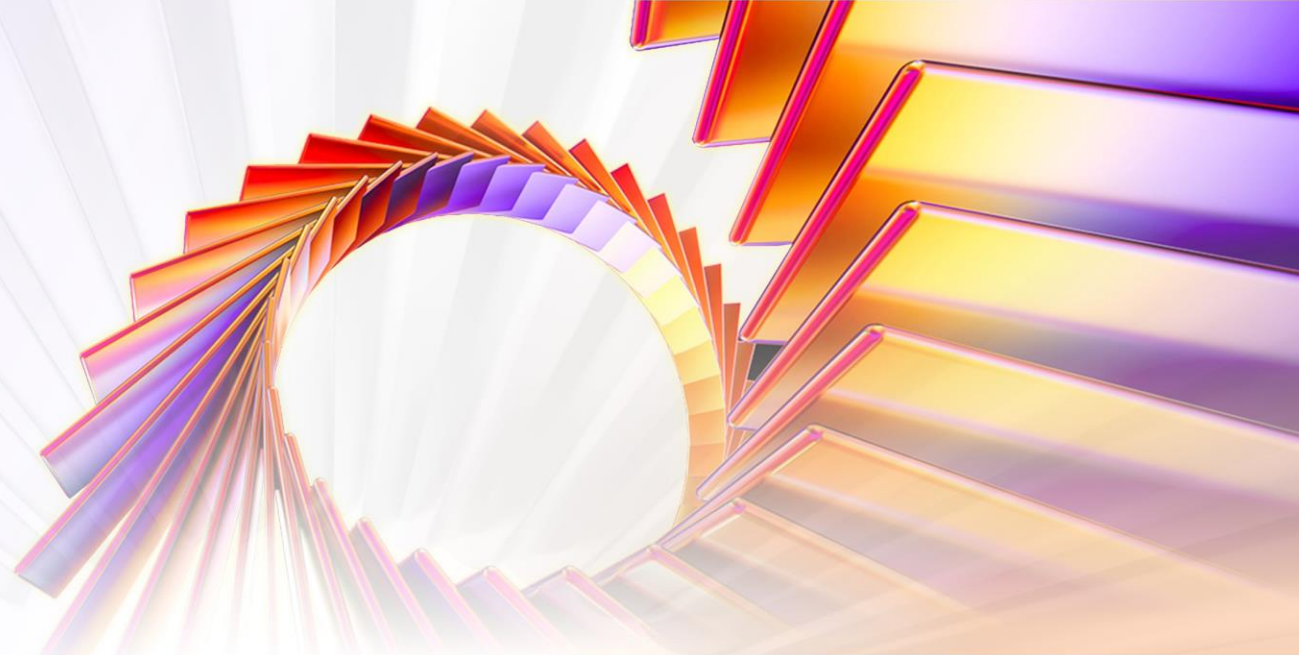


| 8月8-9日 | 北京站

第7届 AI+ Development
Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发



AI趋动下新交互生态 应用中的实践与挑战

戴冠平 | 北京缔智元



戴冠平

北京缔智元首席科学顾问

清华大学，汽车工程/计算机科学 (双学位)、本校工学硕士毕业；
具有20+年 AI 算法研究积累，不仅自建深度学习大模型框架，更将技术能力应用于实践 —— 参与国家汽车安全实验室影像分析及神舟飞船回收舱遥测跟踪等项目，以 AI 技术赋能关键场景的数据分析与精准跟踪；在中间件领域，曾任职中科院软件研究所，BEA/ORACLE（中国），深厚的行业积淀推动其自研出中间件产品 LinkMind；同时具备丰富的银行与电信系统架构经验，出版有《叱咤风云：GoldenGate 企业级运维实战》等多部企业级运维领域著作，持续深耕技术研究与应用。

目录 CONTENTS

- I. AI大模型开启的新一代人机交互机会
- II. AI大模型应用融合技术发展
- III. AI大模型应用落地实践与挑战

PART 01

AI大模型开启的新一代人机交互机会

▶ 1.1 人机交互演进历史回顾

人机交互（Human-Computer Interaction，简写HCI）：指人与计算机之间使用某种对话语言，以一定的交互方式，为完成确定任务的人与计算机之间的信息交换过程。

穿孔卡片/纸带输入，靠手工操作开关、插拔线缆输入指令，输出以指示灯或穿孔卡片呈现；
命令行界面（CLI）；

早期手工阶段

1940s-1960s

1973 施乐Alto：首个GUI（图形用户界面）；
1984 苹果Macintosh：普及鼠标+窗口操作；
Window系统推动GUI普及，以图标，窗口等元素开启大众化计算时代；

图形用户界面（GUI）革命

1970s-1980s

移动与触控兴起

1993 IBM Simon手机：触控屏；
1997 Dragon Dictate：语音识别；
2007 iPhone：电容触屏+多点触控；
手势操作（滑动、缩放）；

1990s-2000s

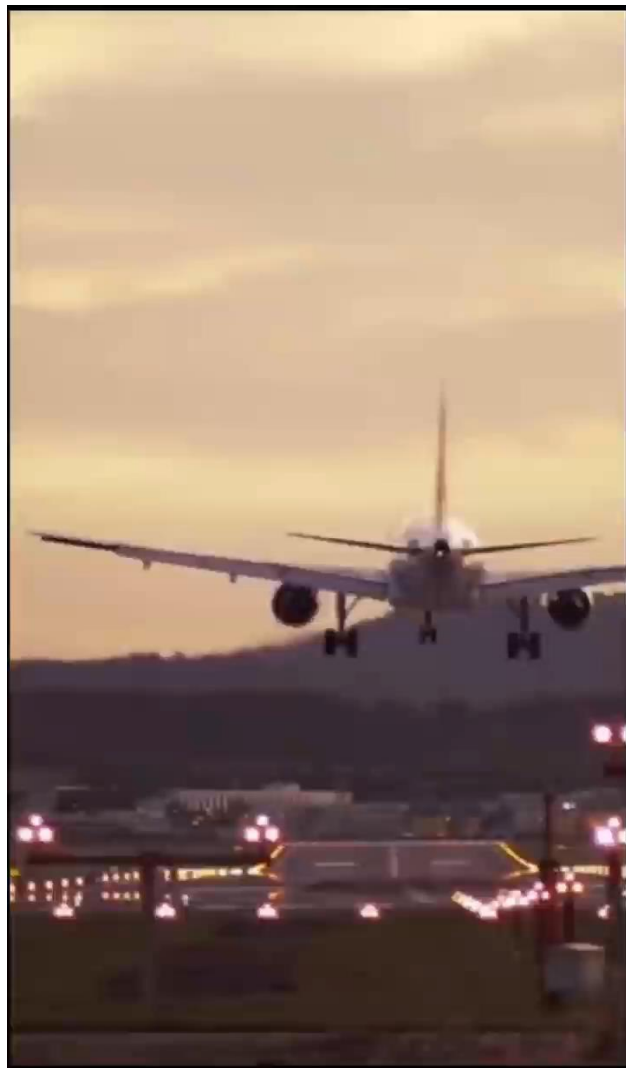
无感交互发展

语音交互（Siri/ Alexa）
计算机视觉（Face ID、手势识别）

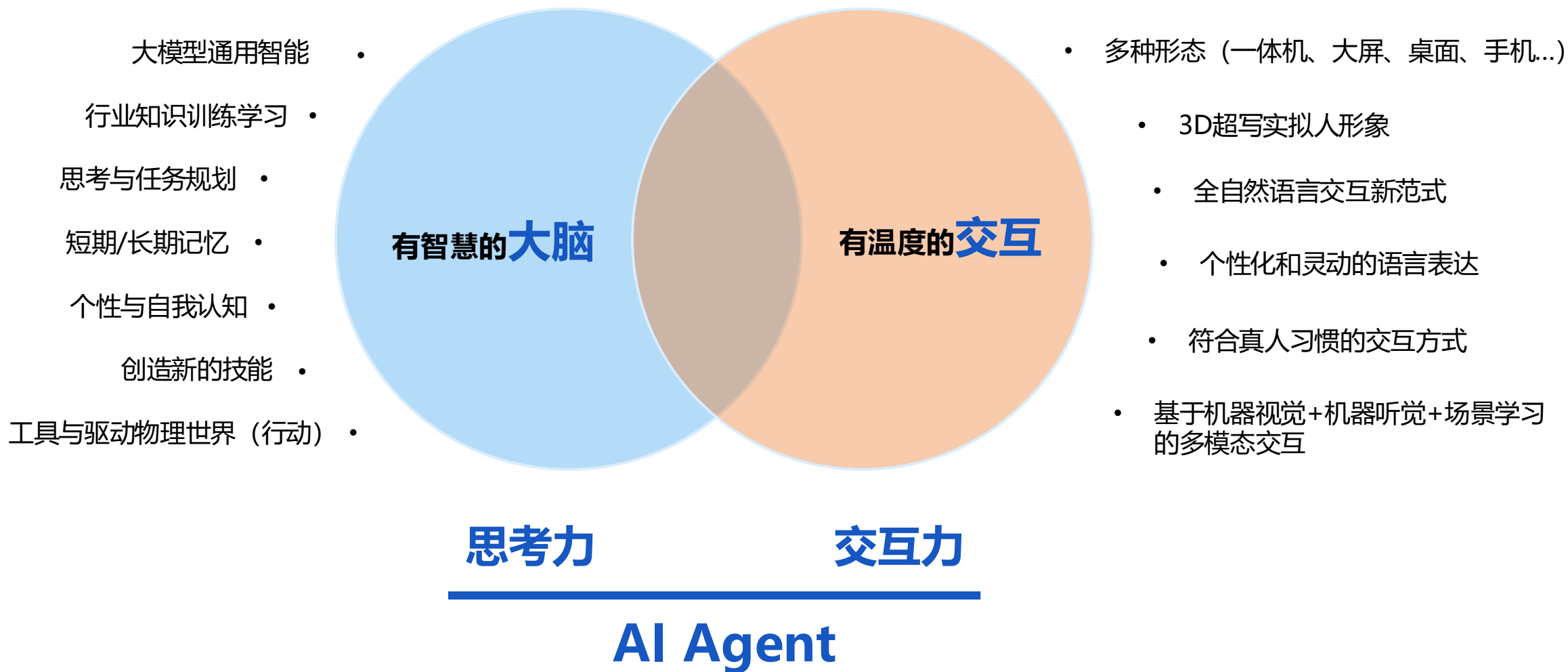
2010-至今



▶ 1.2 数字人AI原生态交互模式

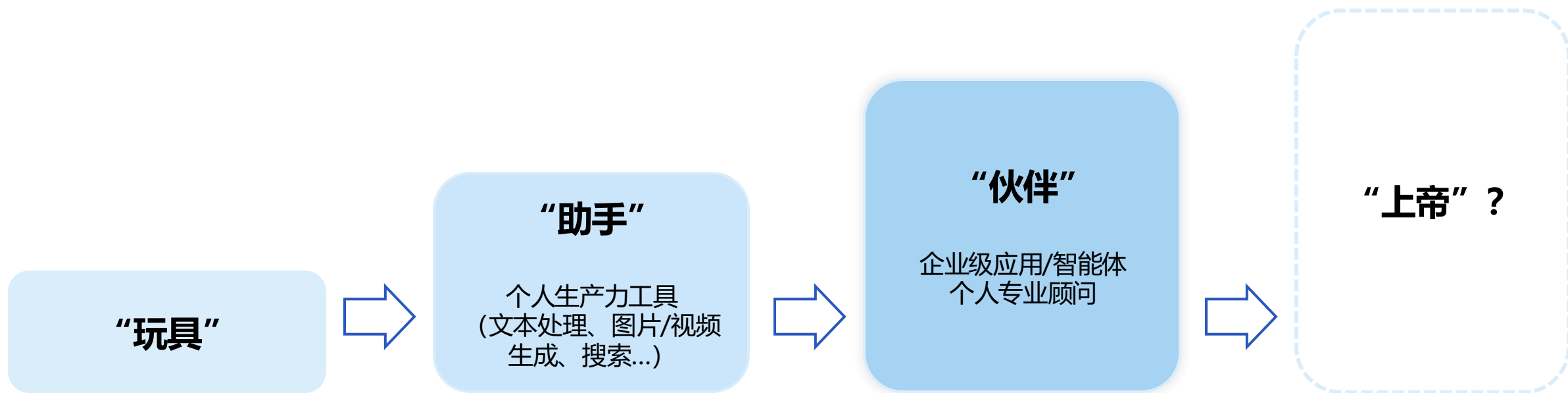


▶ 1.2 数字人AI原生态交互模式



▶ 1.3 当前行业现状、瓶颈与破局

- 高分低能：能做奥数题，但不懂基本的行业知识。
- “幻觉”问题：在严肃场合下“一本正经地胡说八道”。
- 鸡同鸭讲：细分业务场景下的意图泛化理解能力严重缺乏。
- 浅尝辄止：很难与业务流程结合而为客户创造实际价值。
- 场景限制：现有大部分应用局限在个人办公的生产力提升（基于桌面的文本处理和对话）。



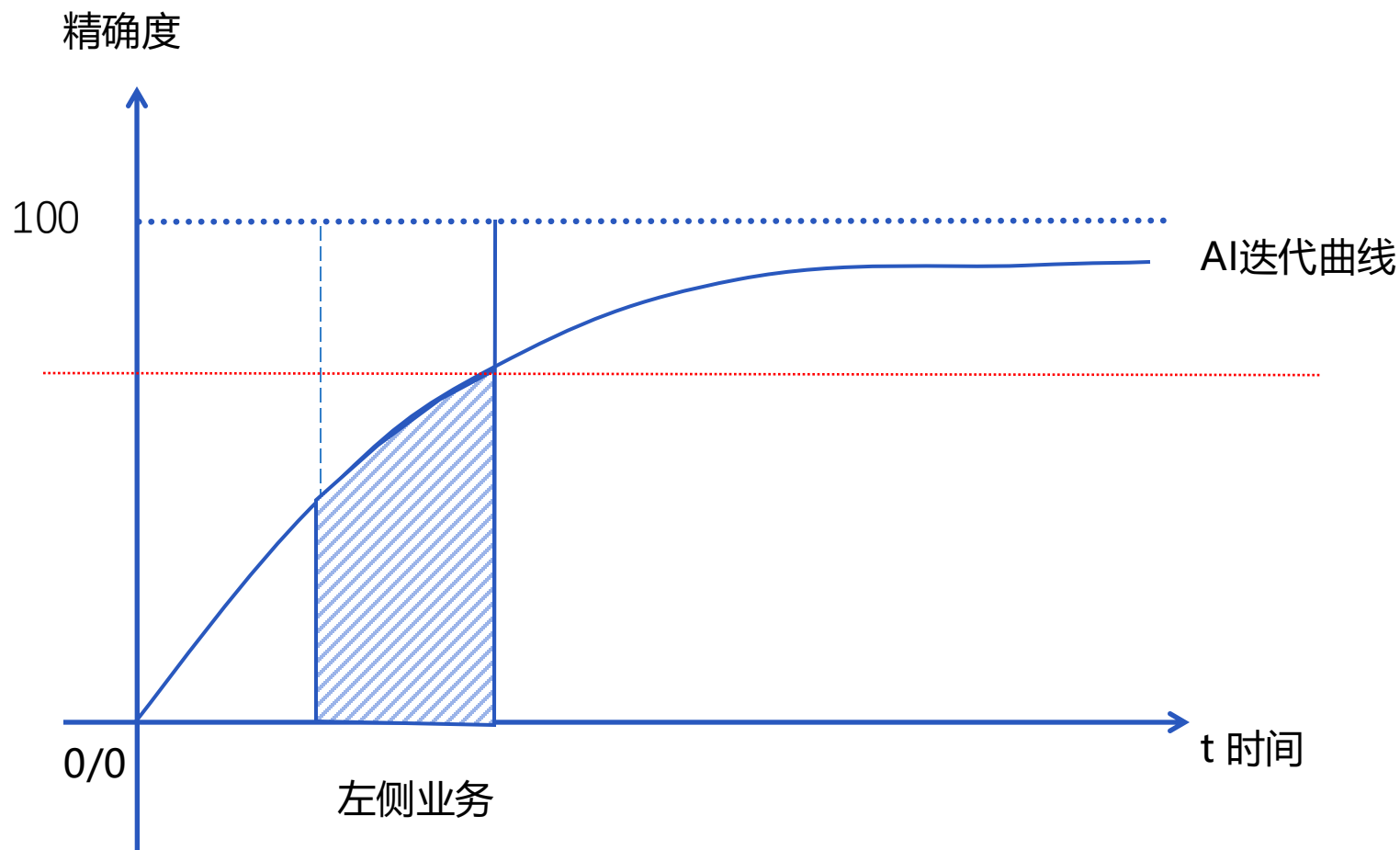
▶ 1.3 当前行业现状、瓶颈与破局

鉴于AI跟场景结合时，有时间投入边际效应递减的经验规律，建议优先选择当前AI水平易于超过行业门槛的业务切入，否则从原型到上线量产的时间成本太高。

曾经MIT人工智能实验室主任丹妮拉·鲁斯说：

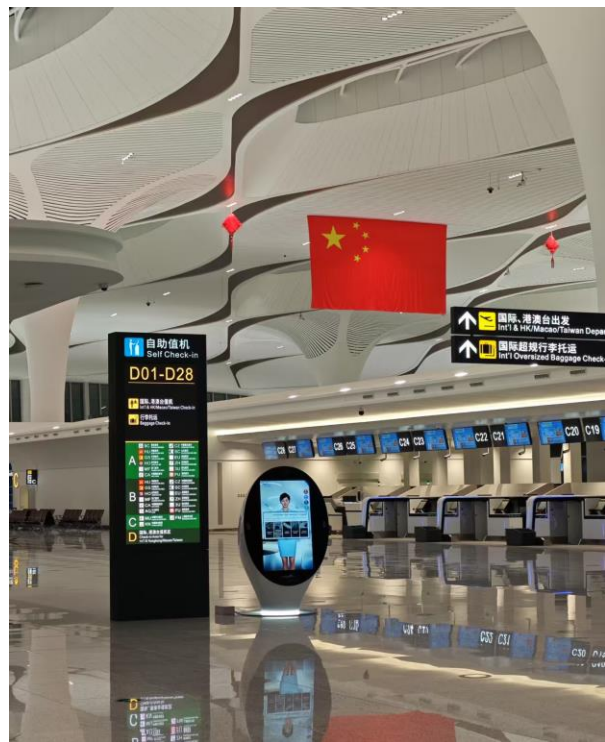
“首先是人工智能解决问题的能力没有那么好，比如说清理桌子我们觉得很简单，但是对于机器人来说它们会做得特别差。”

“第二，人工智能很难提升表现率，将90%提升到99%是特别难的”



▶ 1.3 当前行业现状、瓶颈与破局

- **信息升维**：交互的内容，从文本，到语音，到视频，维度的升级带来了信息量和体验感的指数级增长。
- **场景复杂**：交互对象不再只是电脑或手机前一个人的一问一答，而是复杂的多人交互、多轮交互。
- **消除幻觉**：比起会做奥数题，更重要的是懂行业、懂业务，不会胡说八道。
- **实时性**：在需要纯自然语言交互方式的场景下，客户对端到端的延时要求非常苛刻。
- **交互体验**：客户想要的，是解放双手，回归最符合人性和自然的视听混合体验。
- **产生行动**：与业务相关，就要求完成对环境和业务系统的控制，如何从“书呆子”变成具有行动能力的“八爪鱼”？



1.3 当前行业现状、瓶颈与破局

更准检索增强RAG

- ✓ 精细化管理和确保模型输出高准确性.
- ✓ 不断学习提升模型整体性能和准确度.

更快预读缓存Medusa

- ✓ 引入Medusa减少等待时间.
- ✓ 更高效地提升模型性能.

更稳自动切换大模型

- ✓ 在多个链路间进行备份.
- ✓ 故障无缝自动切换备用模型

大

模

型

- ✓ 精准识别用户意图提供响应.
- ✓ 快速、准确地分析用户输入.
- ✓ 实现更稳定、更易用.

更强意图探查Graph

- ✓ 一次编写便可适用多个模型.
- ✓ 减少重复工作,提高开发效率.
- ✓ 降低构建难度.

更易一次编写, 多模型通用

- ✓ 多模态内容理解.
- ✓ 丰富知识表达形式.

更多多种模态的信息来源

增强效果	更准	更快	更稳	更强	更易
大模型增强前	50%~7x%	3s-20s	使用模型发生异常导致上层应用不可用	平凡	构建一些常见应用仍需花费数月
中间件增强后	9x%	200ms~1s	当故障发生1s内完成模型切换	MoE	常见应用构建最快可缩短到周级

PART 02

AI大模型应用融合技术发展

▶ 2.1 交互生态的多维协同体系深度解析

文本模态

自然语言输入，
例如：“今天北京天气怎么样？”、“帮我查询红烧肉的做法”等



视频模态

可根据上传的视频文件，完成对
视频中人物的追踪



语音模态

语音的情绪化输入输出，根据
所选情绪可输出不同感情音频



图像模态

图像上传，可通过您上传的图
像，为您生成相关的介绍，如
图像中的人物状态等



AI大模型
中间件

AI Agent

任务执行者，具备一定的自主决策和行
动能力，能够在大型模型的支持下，灵活
应对各种任务场景



Agent Tools

能力拓展者，拓展了 AI Agent 的能力
边界，使其能够处理更广泛类型的任务



Function Calling

功能的激活器，使得大模型不再局限
于自身的知识储备，而是能够动态获
取外部实时数据和服务，增强了模型
的实用性和交互能力

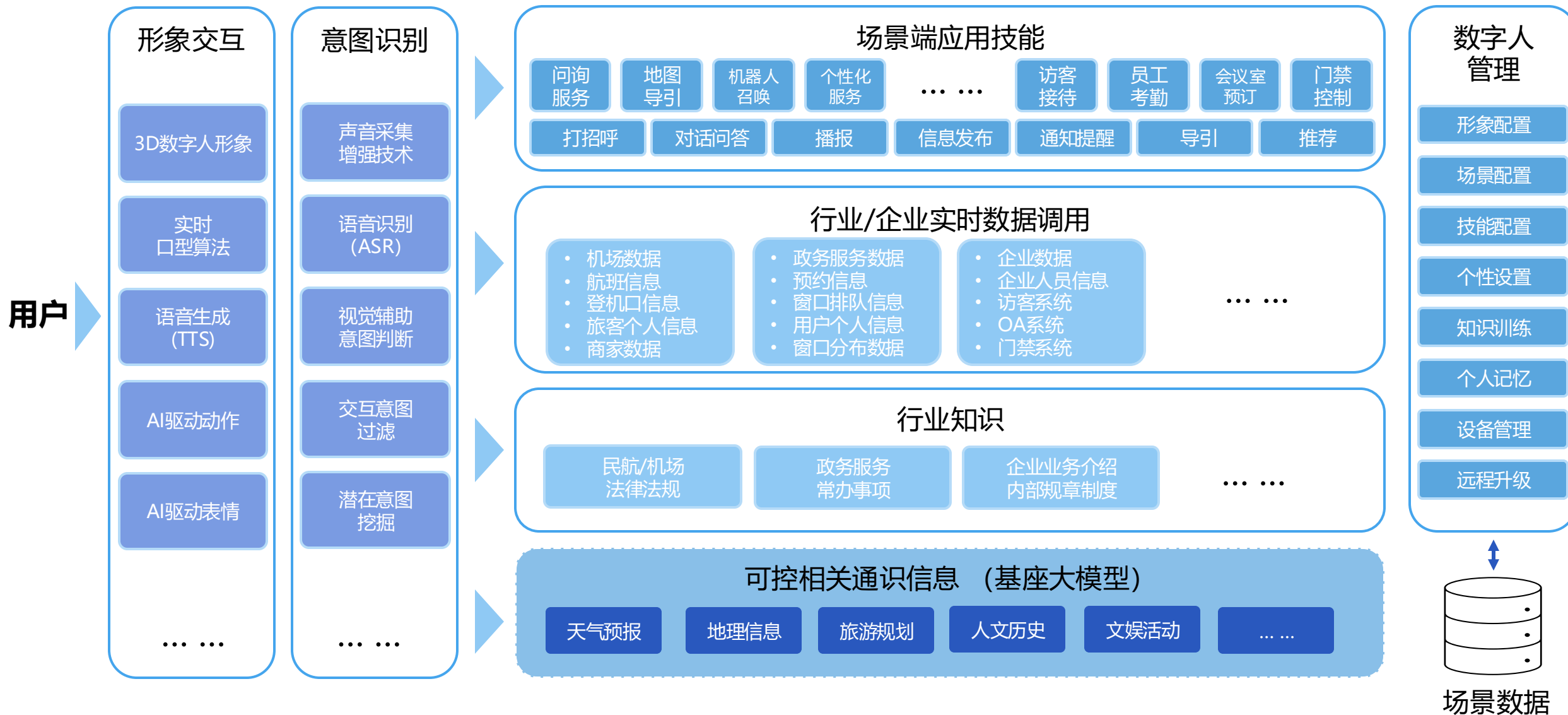


MCP

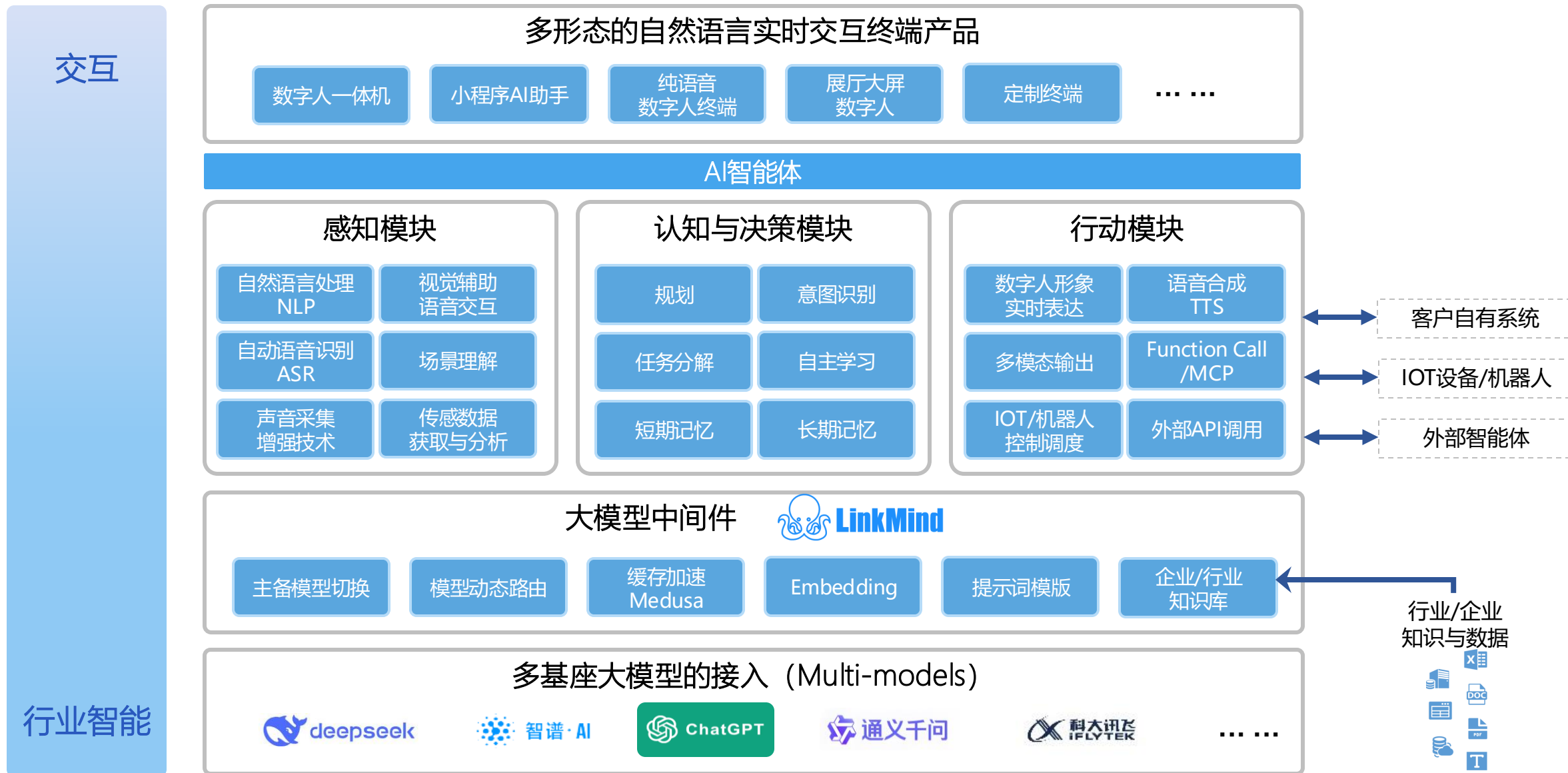
交互的规范者，确保各信息正确转换
处理



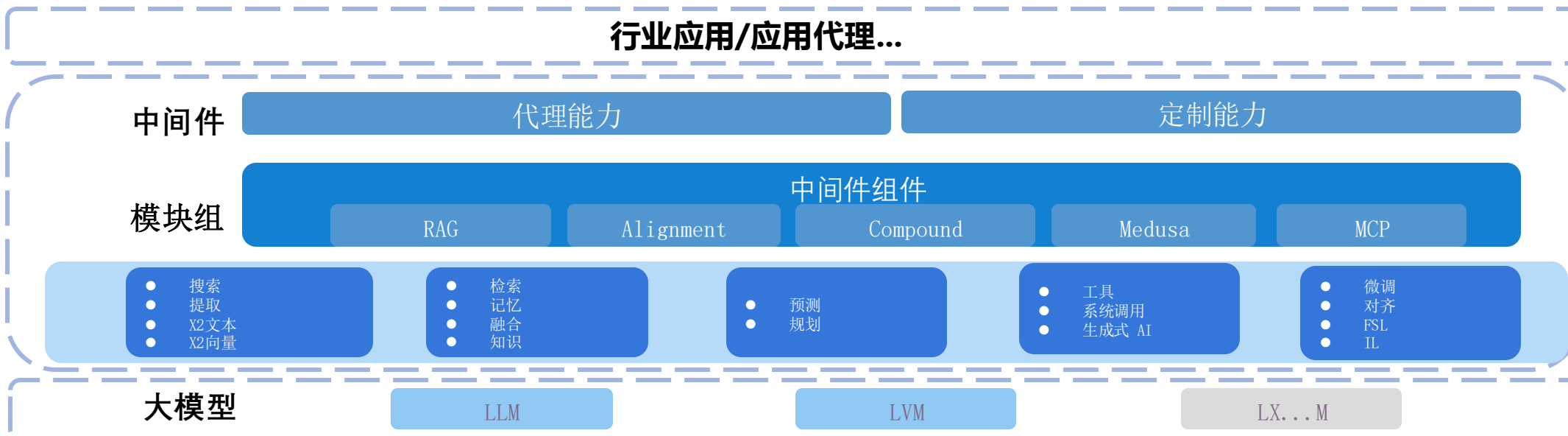
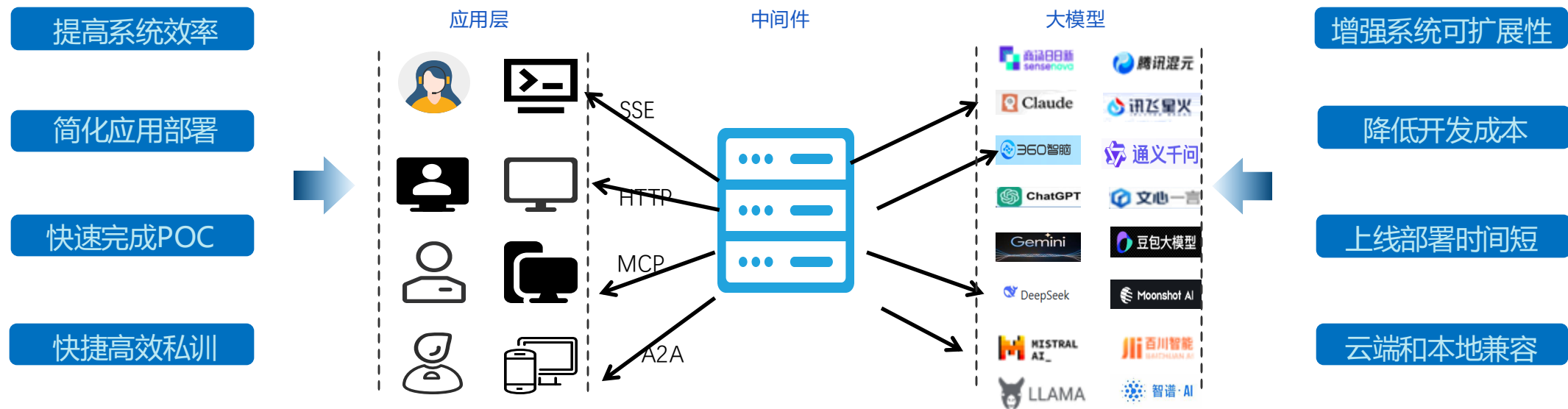
2.1 交互生态的多维协同体系深度解析



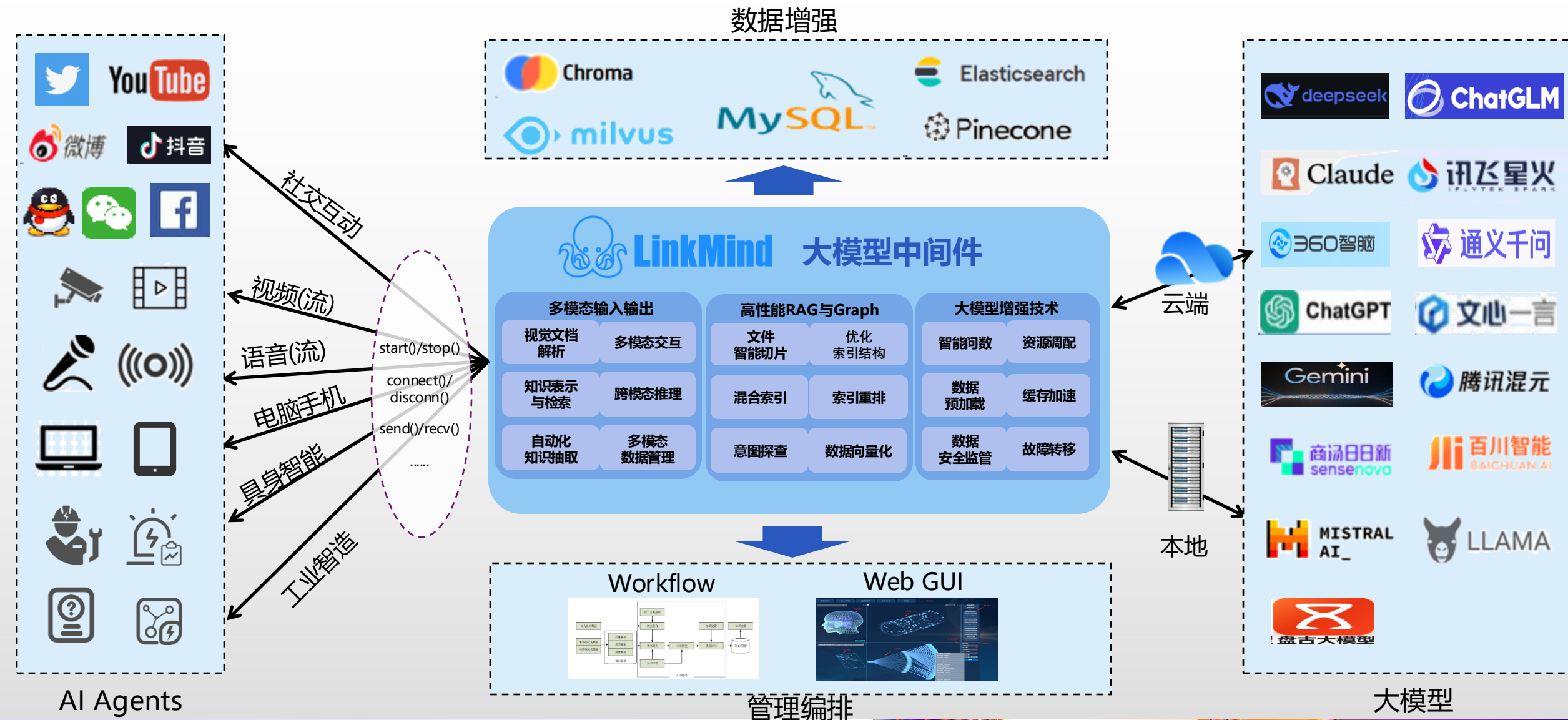
2.1 交互生态的多维协同体系深度解析



▶ 2.2 AI融合技术的“神经中枢” 中间件



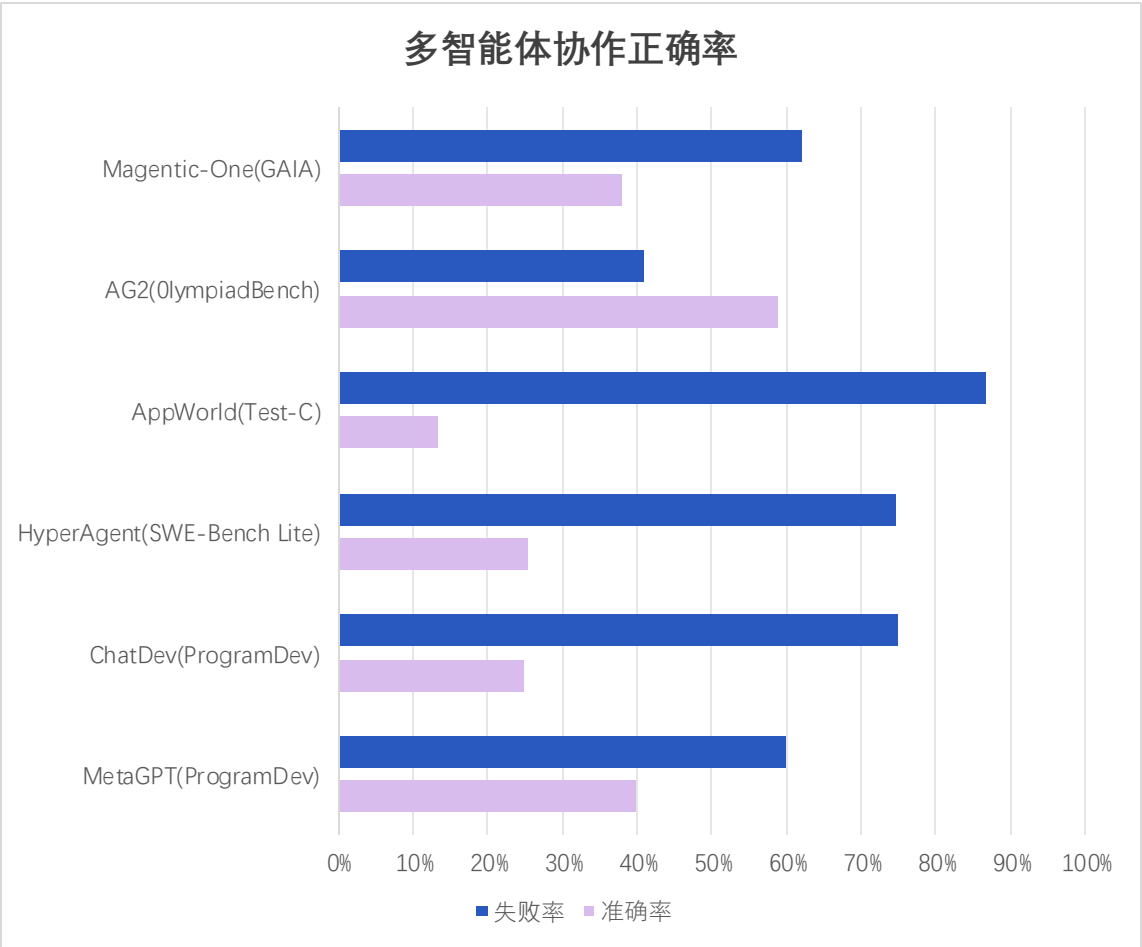
▶ 2.2 AI融合技术的“神经中枢” 中间件



2.3 Agentic AI的认知升级与协作优化

据调研，实际应用中，多智能体系统的表现难达预期。有研究团队通过分析ChatDev等当前最流行的开源多智能体系统发现，即便采用GPT-4o、Claude-3等前沿大模型，任务准确率仍远低于预期，且呈现智能体数量增加导致的性能衰减效应。

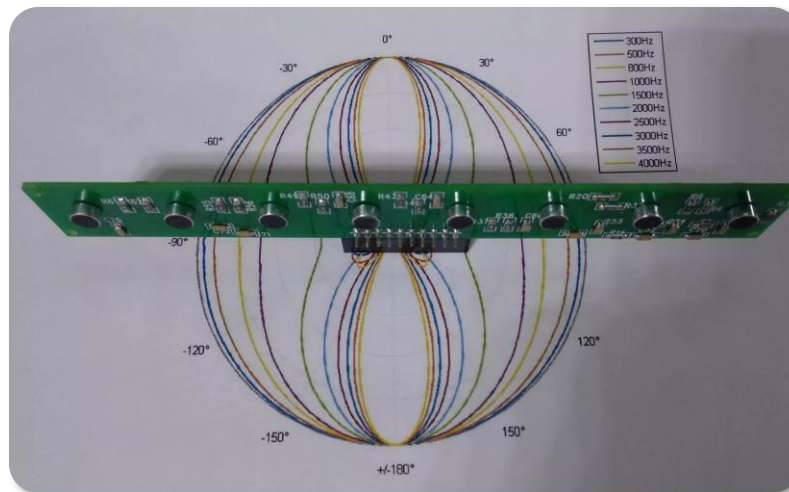
开源 MAS 框架	准确率	失败率	备注
MetaGPT(ProgramDev)	40%	60%	单智能体主导开发，侧重代码生成逻辑
ChatDev(ProgramDev)	25.00%	75%	多智能体协作模式支持复杂项目但协作效率待提升。
HyperAgent(SWE-Bench Lite)	25.30%	74.70%	软件开发评估框架，用于智能体开发能力评测，数据侧重任务完成度与效率指标。
AppWorld(Test-C)	13.30%	86.70%	聚焦软件测试自动化，测试用例生成能力较弱，缺陷定位准确率较低。
AG2(OlympiadBench)	59.00%	41.00%	数学推理专用模型，结合强化学习与符号推理，奥林匹克级问题求解表现突出。
Magentic-One(GAIA)	38%	62%	通用型多智能体框架，覆盖多领域协作场景，但各维度性能均衡性有待优化。



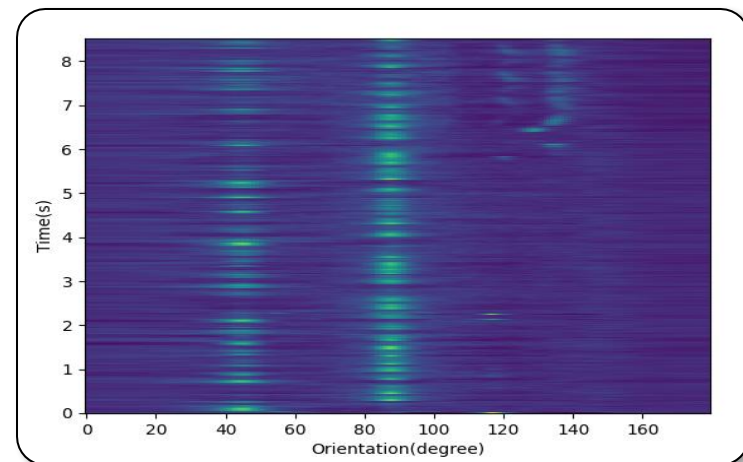
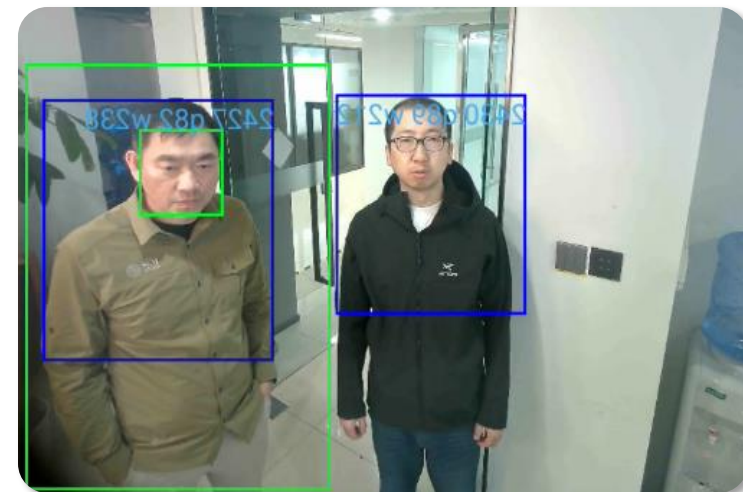
▶ 2.3 Agentic AI的认知升级与协作优化



公共场合多人场景识别，嘈杂环境准确收音，是数字人必须克服的技术门槛



缔智元自研多路麦克风阵列及算法，利用声学技术主动屏蔽背景噪音，判别声音来向。



与自研机器视觉算法联动，进行交互人员的动态跟随，实现多人交流对话,无需唤醒词，不插嘴，让数字人更机灵。

“多方向动态语音感知”技术（发明专利申请中）

未处理	方向一增强	方向二增强
		

▶ 2.3 Agentic AI的认知升级与协作优化



移动终端



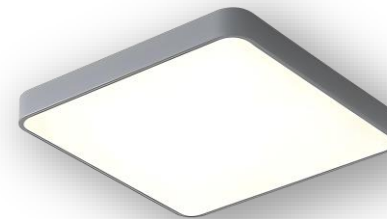
楼宇自动化控制



桌面终端



能源管理



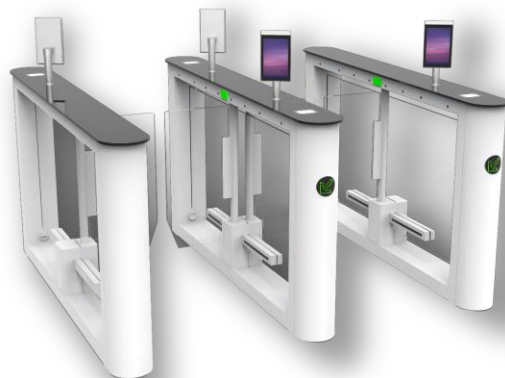
办公照明



清洁设备



智能门锁



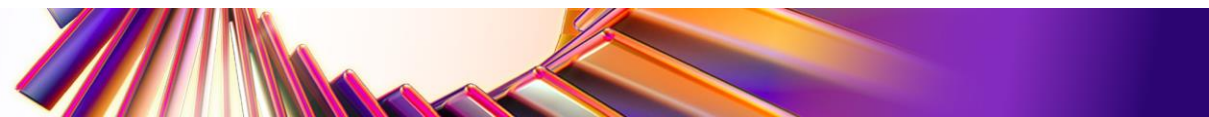
门禁闸机



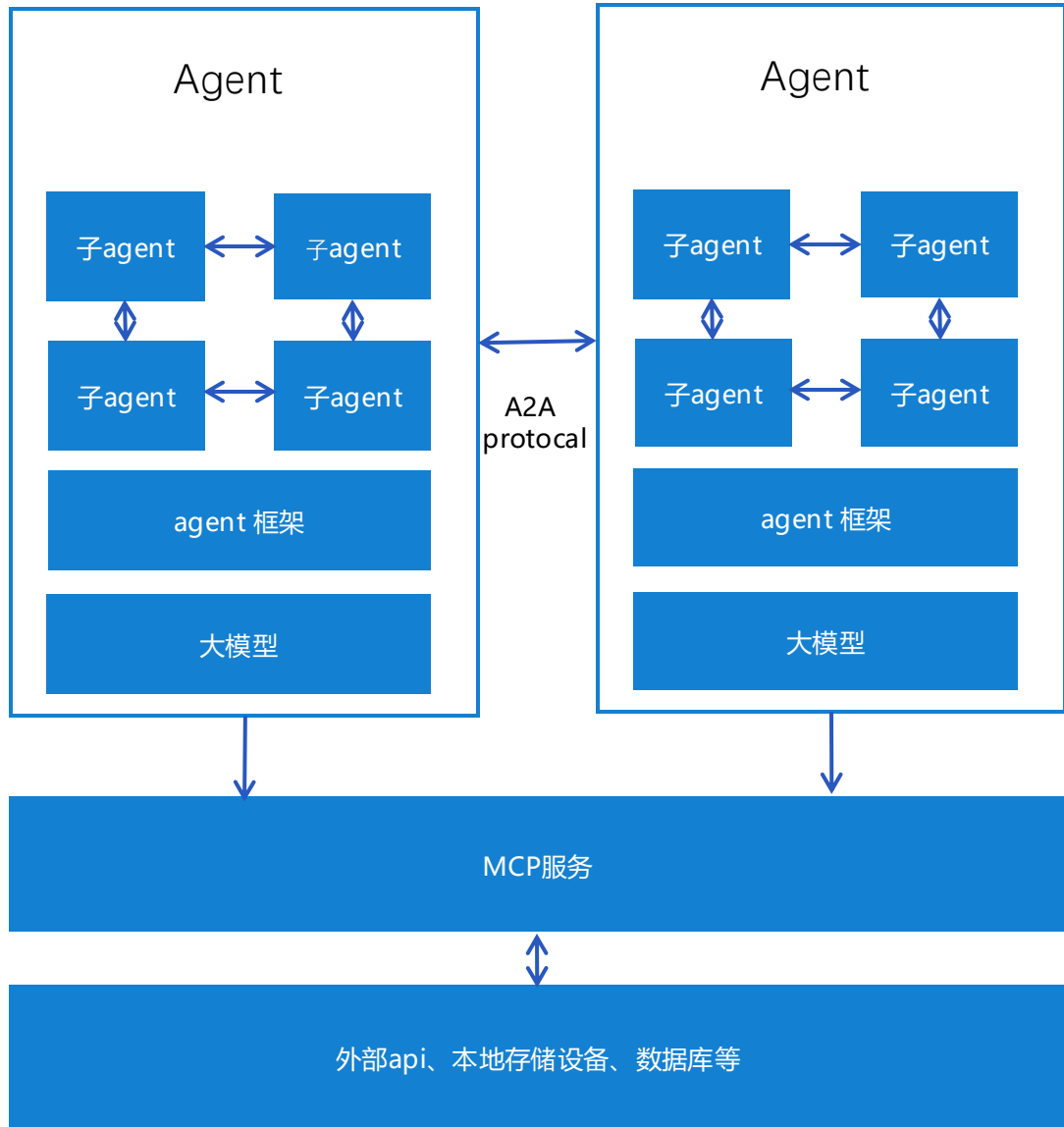
送货机器人



专用设备



2.4 热点前沿MCP、A2A技术及其他



A2A协议：

A2A (Agent-to-Agent Protocol) 是一项开源通信协议，旨在为不同框架、供应商开发的AI智能体 (Agent) 提供标准化协作方式，使其能够跨越技术壁垒，相互通信、分配任务并协同完成复杂流程。A2A协议的核心目标是解决智能体间的互操作性问题，推动多智能体系统从“孤立运行”向“开放协作”转变

核心组件：

Agent Card (智能体名片)： 每个智能体通过JSON文件声明自身能力 (如支持的任务类型、接口URL、认证方式)，供其他智能体发现。

任务管理： 定义任务生命周期 (提交→执行→完成/失败)，支持实时状态订阅 (如Server-Sent Events)。

多模态通信： 允许消息包含文本、文件、结构化数据等多种内容类型，支持用户体验协商 (如格式适配)。

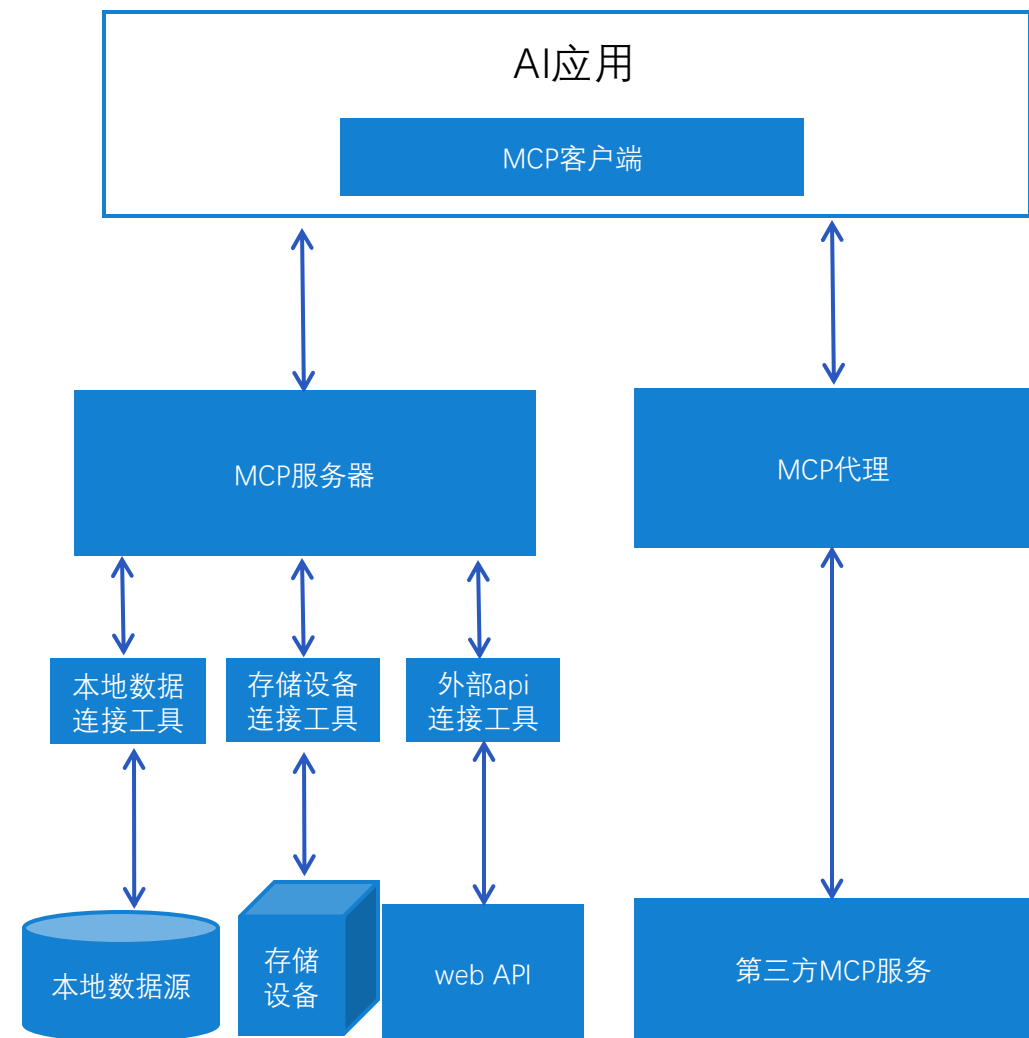
与 MCP 协议的互补性：

A2A：聚焦智能体间通信规则 (如外交礼仪)。

MCP (Anthropic)：专注模型与外部工具/数据源的连接 (如USB-C 接口)。

协同效应： A2A 管理智能体协作流程，MCP 提供外部资源访问能力，共同构建完整 AI 生态。

▶ 2.4 热点前沿MCP、A2A技术及其他



MCP:

MCP 核心采用客户端-服务器架构，它是一个开放协议，为应用程序向 LLM 提供上下文的方式进行了标准化。就像 USB-C 为设备连接各种外设和配件提供了标准化的方式一样，MCP 为 AI 模型连接各种数据源和工具提供了标准化的接口。

MCP 客户端:

MCP客户端在AI应用中扮演中间件的角色，其主要作用是通过标准化的通信协议与MCP服务器进行安全通信。它作为AI应用与MCP服务器之间的桥梁，负责将用户请求转化为MCP服务器可识别的标准化API调用，并确保数据的安全传输。并将自服务器的响应，转换为AI应用可以理解和使用形式，从而实现对各种外部工具和数据源的访问，增强AI应用的功能性和灵活性。

MCP 服务器:

MCP服务器是基于模型上下文协议（MCP）的标准化接口，主要用于让AI应用（客户端）无缝访问外部数据源、工具或服务（如数据库、存储、API等），通过统一协议打破数据孤岛，提升AI与外部系统的互操作性，简化开发并增强AI的实时数据处理与执行能力。

MCP 服务端代理:

MCP代理服务为所有MCP服务端提供了一个统一的接入点，使得客户端无需知道每个MCP服务端的具体地址。

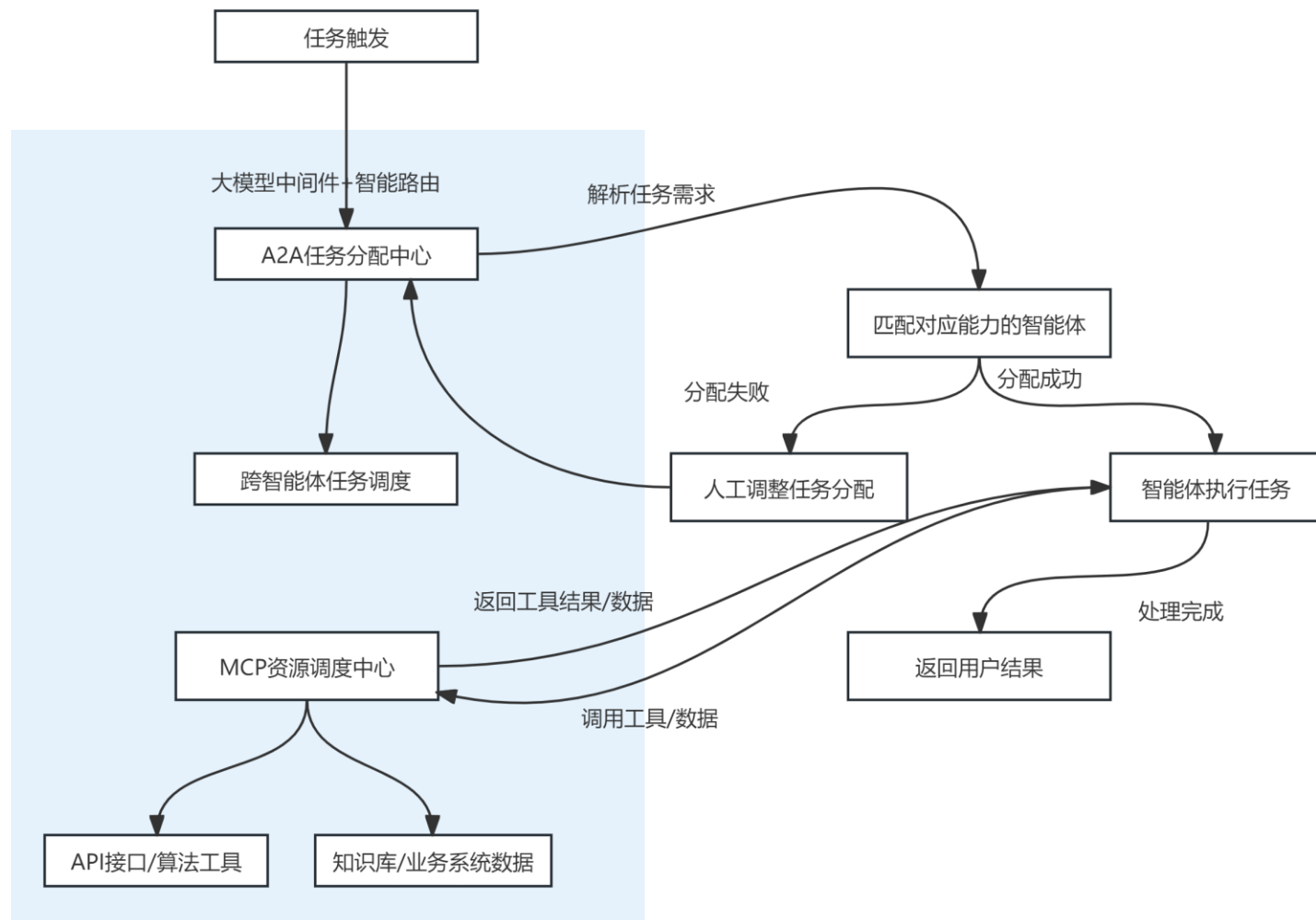


► 2.4 热点前沿MCP、A2A技术及其他

在任务处理流程上，先由 A2A 依据任务需求分配给相应智能体，智能体再借助 MCP 调用所需工具和数据。

A2A（智能体到智能体）：作为任务分配中心，负责解析需求、匹配智能体能力、协调跨智能体协作（如多轮转接场景）。

MCP（工具与数据中台）：统一管理工具和数据资源（如知识库、业务系统），支持智能体按需调用



PART 03

AI大模型应用落地实践与挑战

▶ 3.1 大模型落地应用项目的生产周期



▶ 3.2 产前私有语料融合的信息密度冲突

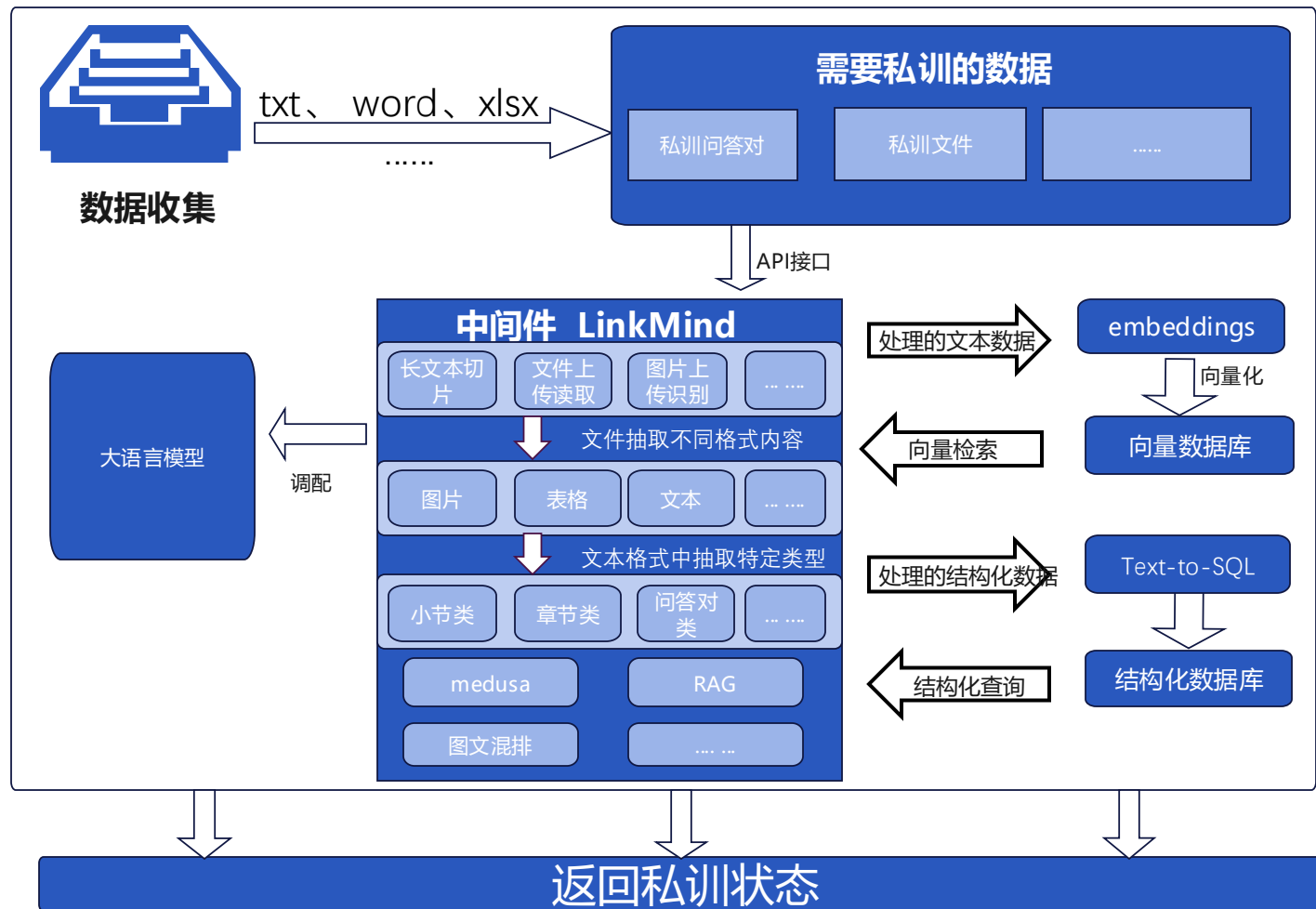
RAG 私有语料清洗与处理机制

多种类型：上传文件中有可能包含多种类型（例如：txt, word, xlsx）。

多种格式：每种类型里会有多种格式，例如word文档中可能包含图片和表格。

智能分析：通过算法对文件内容进行识别分析，判断其是否属于特定类型，例如：小节类，章节类，问答对类，并将其交给相应处理模块进行智能切片。

智能切片：针对内容与意图采取最佳处理方式对文件进行处理。



▶ 3.2 产前私有语料融合的信息密度冲突

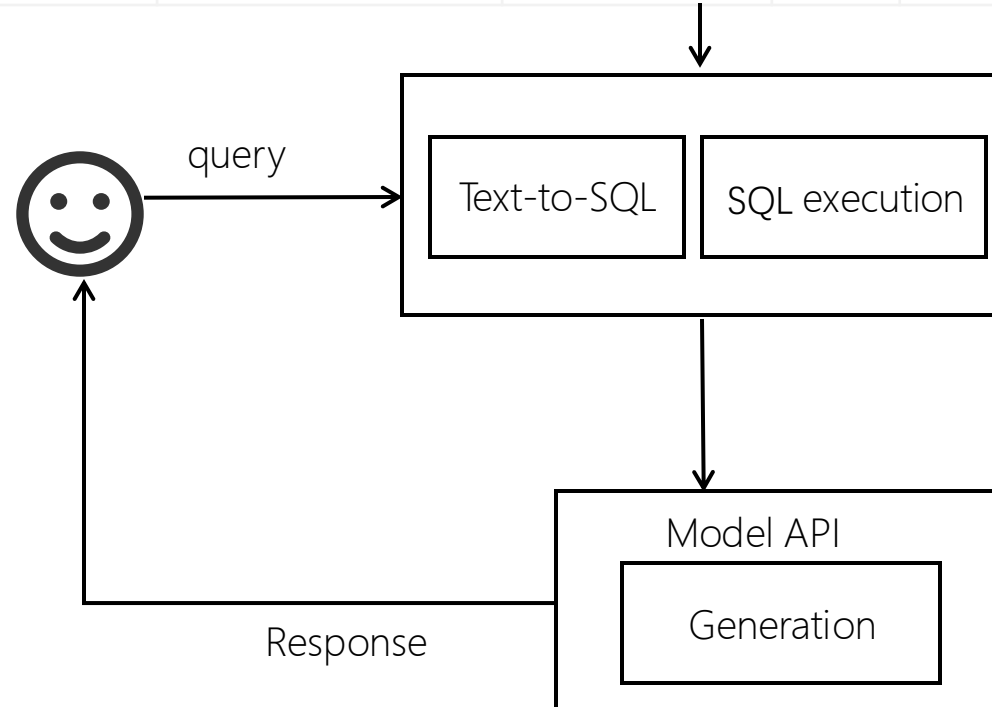
理解与响应逻辑

外部数据源可以是**结构化数据**（关系数据库表），解决大模型对数字不敏感的问题。

检索方式：

- **文本到 SQL**：根据用户查询和表模式，生成所需的 SQL 查询。
- **SQL 执行**：执行生成的 SQL 查询，获取结果。
- **响应生成**：根据 SQL 查询结果和用户原始查询生成最终响应。

Product iD	Product name	Price/unit (\$)	Units	Total
2044	Meow Mix Seasoning	10.99	1	10.99
3492	Purr & Shake	25	2	50
2045	Fruity Fedora	18	1	18
...	...	...	...	...



3.3 产前模型幻觉与准确度的矛盾调节

有监督-->自监督-->无监督

香依信息熵

(Shannon Entropy)

柯尔莫哥洛夫复杂性

(Kolmogorov Complexity)

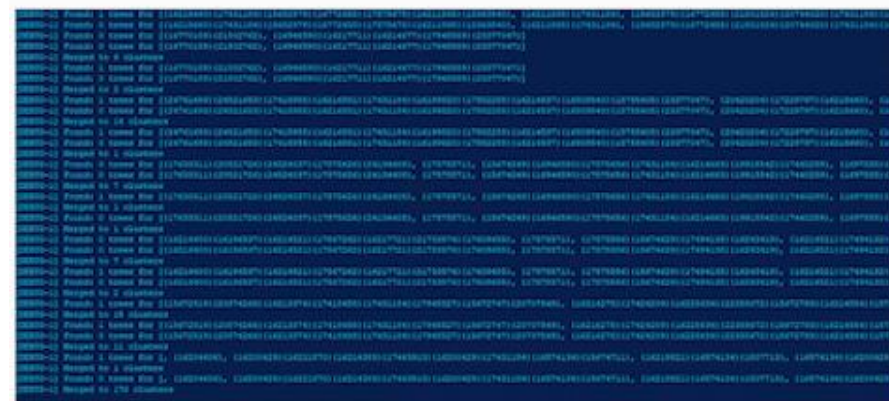
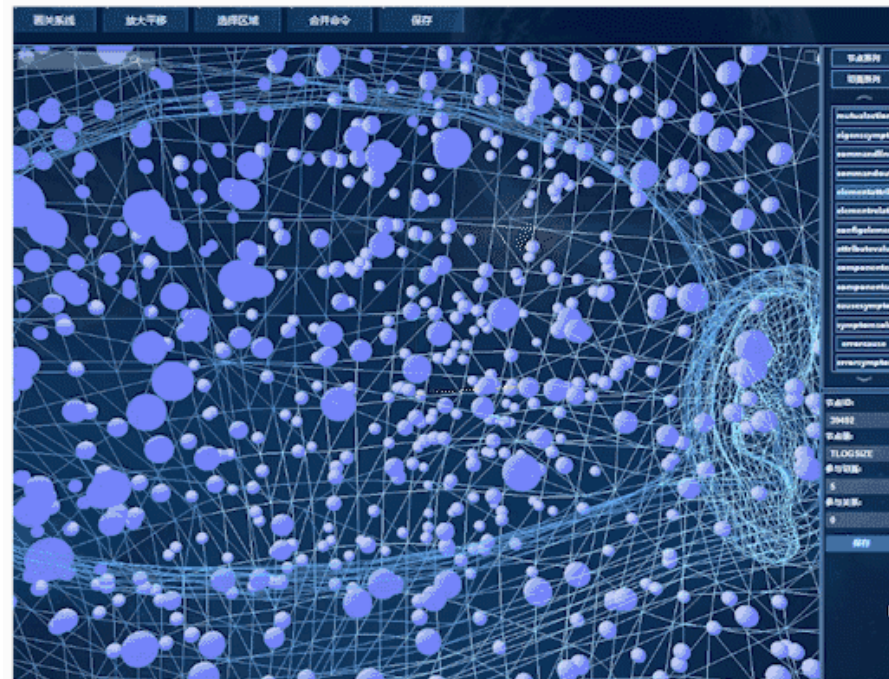
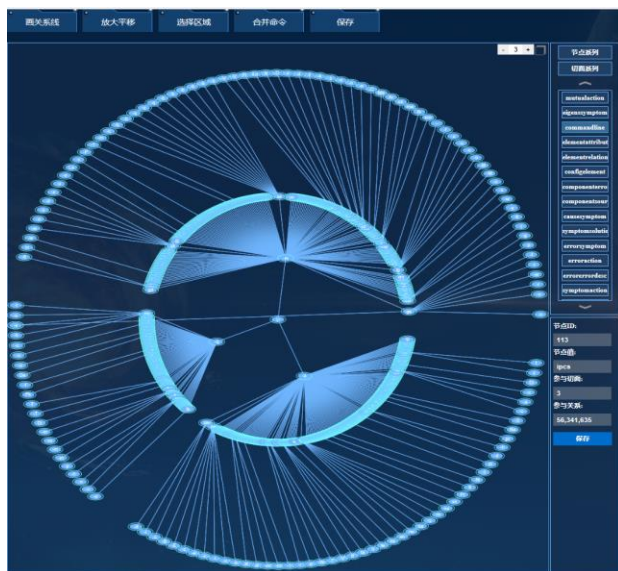
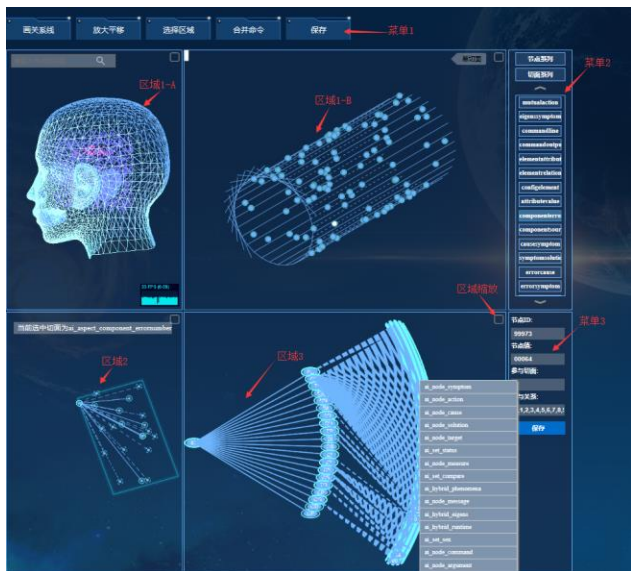
本内特逻辑深度

(Logical Depth)

$$H(X) = -\sum p(x) \log p(x)$$

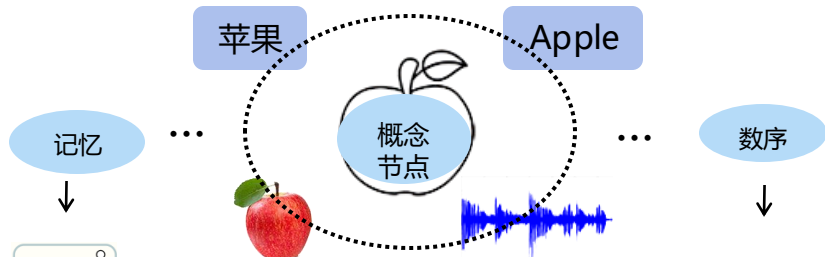
$$KC(x) = \min\{\ell(p) : U(p) = x\}$$

$$LD(x) = \min\{T(p) : \ell(p) = p^* + s \wedge U(p) = x\}$$



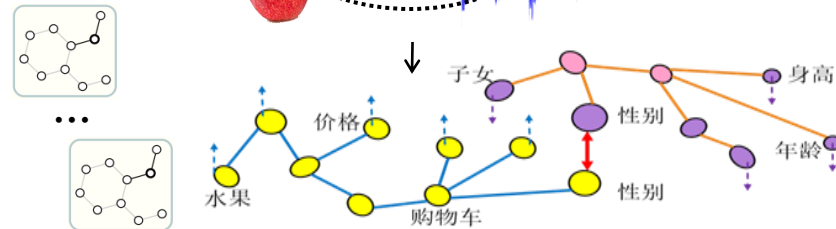
3.3 产前模型幻觉与准确度的矛盾调节

概念学习



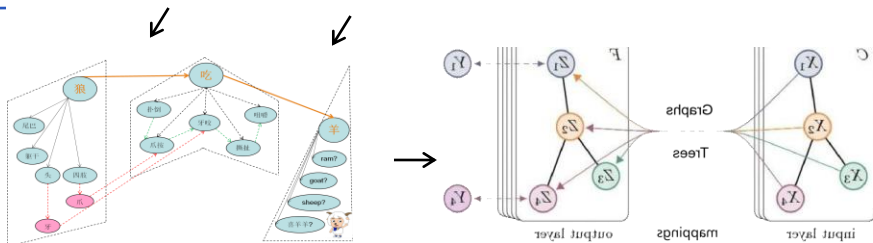
原生多模态：语言(多语种)、声音、曲线、照片、视频、传感器采集等皆可输入、输出，并不是单纯文字

有效学习



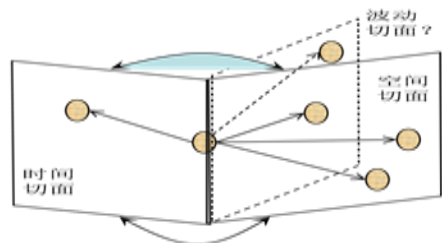
精细化调参：精准关联节点，不像传统大模型暴力求解交叉产生参数，多数本质为无效信息，过度耗散资源

增量学习



可持续生长：不必全量训练，逐个子集逐个切面一步步连续成长为超大模型，学习中模型随时持续直接可用

思维学习

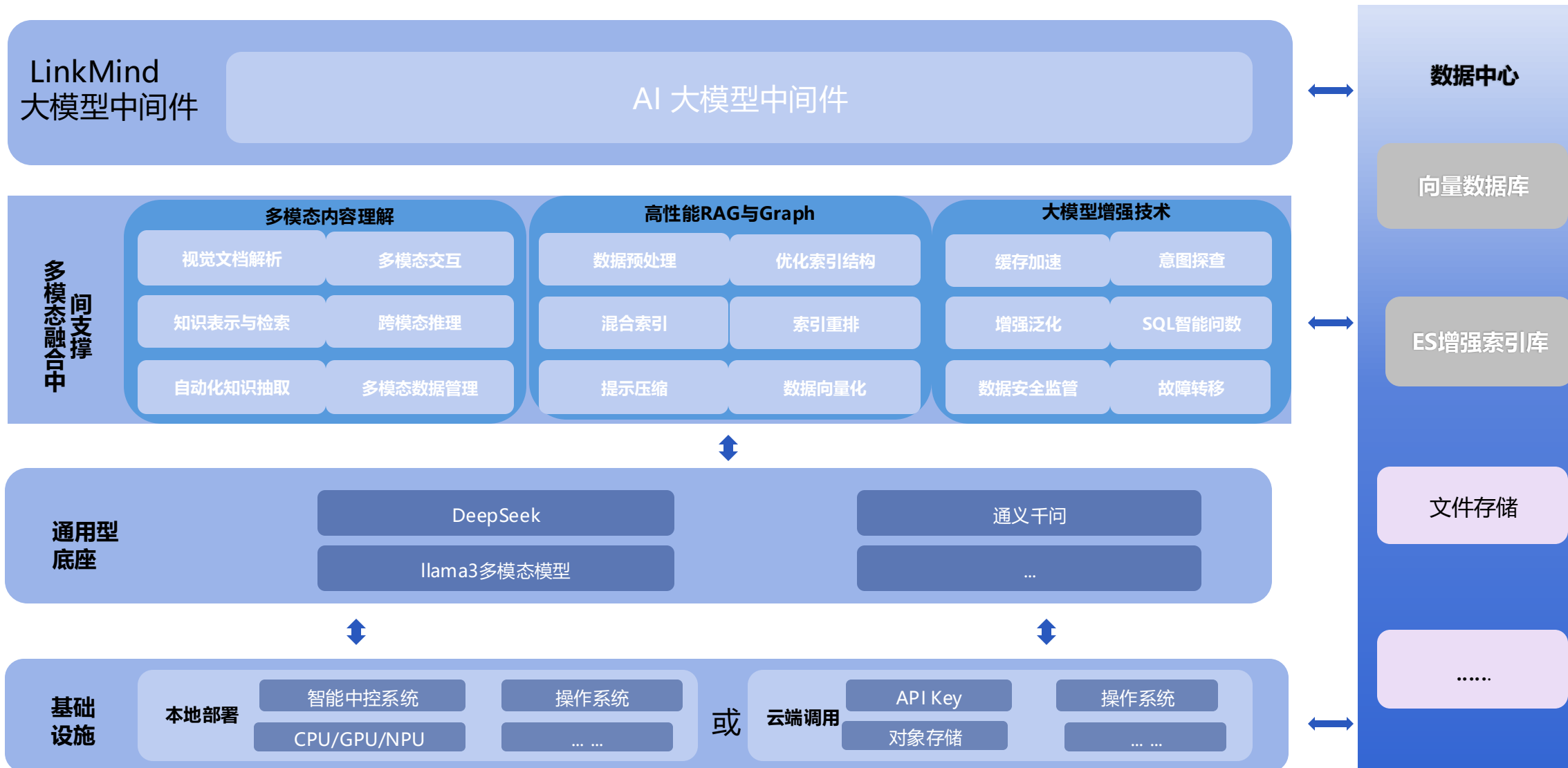


基于理解的：可解释地进行演绎、归纳和联想、抽象等智能活动，不是黑箱形态或简单形式逻辑

实践验证的一种可生长、迭代、持续、积累，逐级通向超大规模的有效路径，并且天然多模态



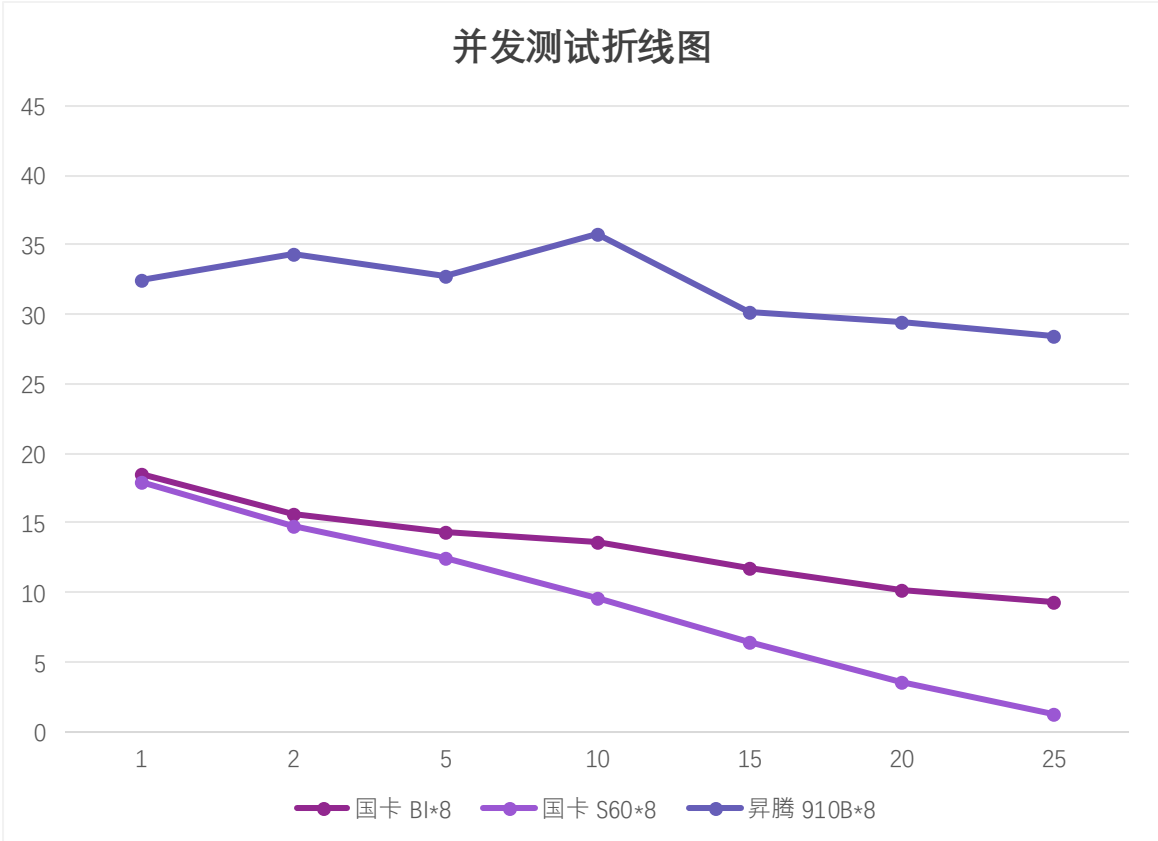
▶ 3.3 产前模型幻觉与准确度的矛盾调节



▶ 3.4 产中并发场景下性能的高衰减挑战

说明：下图展示了不同算力卡在不同并发数量下的处理速度（tps/s）情况。横轴为并发数量（个），纵轴为处理速度（tps/s）。我们选取了并发数为 1、2、5、10、15、20、25这几个关键节点进行测试。

并发数 (个)	国卡BI*8 吞吐率 (tps/s)	国卡S60*8 吞吐率 (tps/s)	昇腾 910B*8 吞吐率 (tps/s)	说明
1	18.5	17.9	32.4	轻负载下性能正常。
2	15	14.8	34.3	并发增加，性能初显损耗。
5	14	12.5	32.8	性能缓慢下降，仍可承受一定并发。
10	13	9.6	35.8	性能持续平稳下降。
15	11	6.5	30.2	并发量高时对性能造成显著影响。
20	10.2	3.5	29.5	性能下降影响突出。
25	9.3	1.3	28.5	性能下降影响加剧。



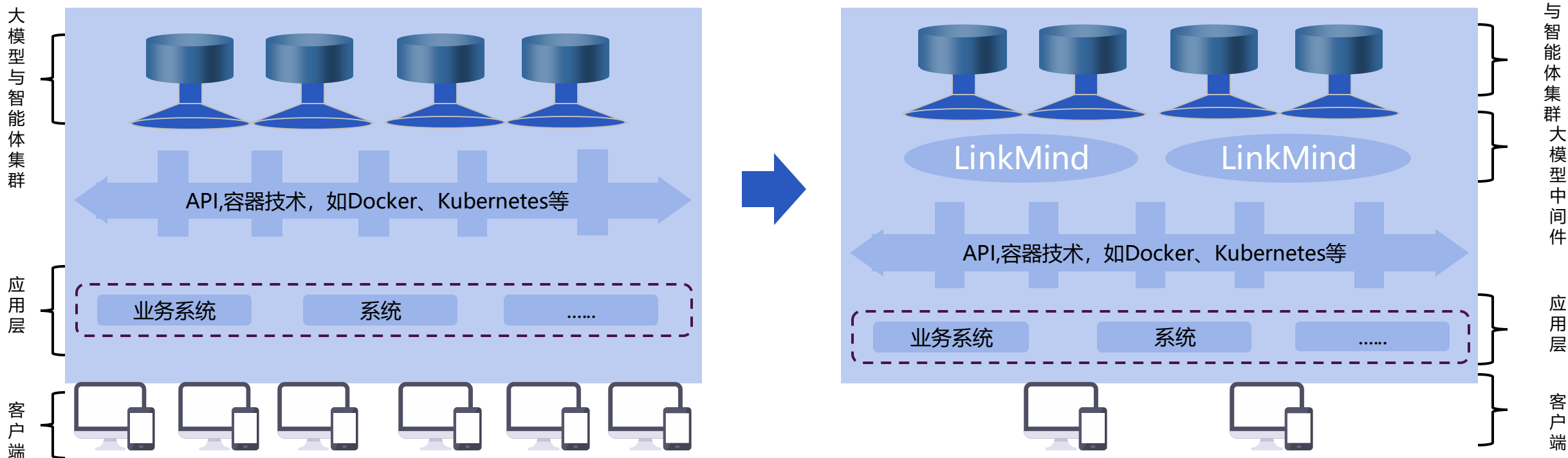
从整体折线趋势来看，随着并发数量的增加，各卡的处理速度呈下降趋势，说明并发量的增加会对该卡的性能产生不同程度的负面影响。



▶ 3.4 产中并发场景下性能的高衰减挑战

两层结构：由客户端和服务端组成，客户端直接与业务系统交互，通过 API 及 Docker、Kubernetes 等容器技术实现通信与资源管理。当并发压力增大时，处理大量请求容易出现连接超时、中断等问题，为确保系统正常运行，只能通过扩展服务器来解决，但这种扩展方式不仅成本高，还会增加复杂度。

三层结构：新增了大模型中间件组件层。在客户端与业务系统交互基础上，通过中间层的扩展，即可将业务逻辑进一步拆分和封装，提升了系统的灵活性、可维护性和可扩展性，更适应复杂业务和大规模应用场景。



总结：在高并发场景下，三层架构通过中间层的弹性扩展能力，显著降低了对服务器硬件的依赖，相对减低了成本，是比两层架构更优的选择。



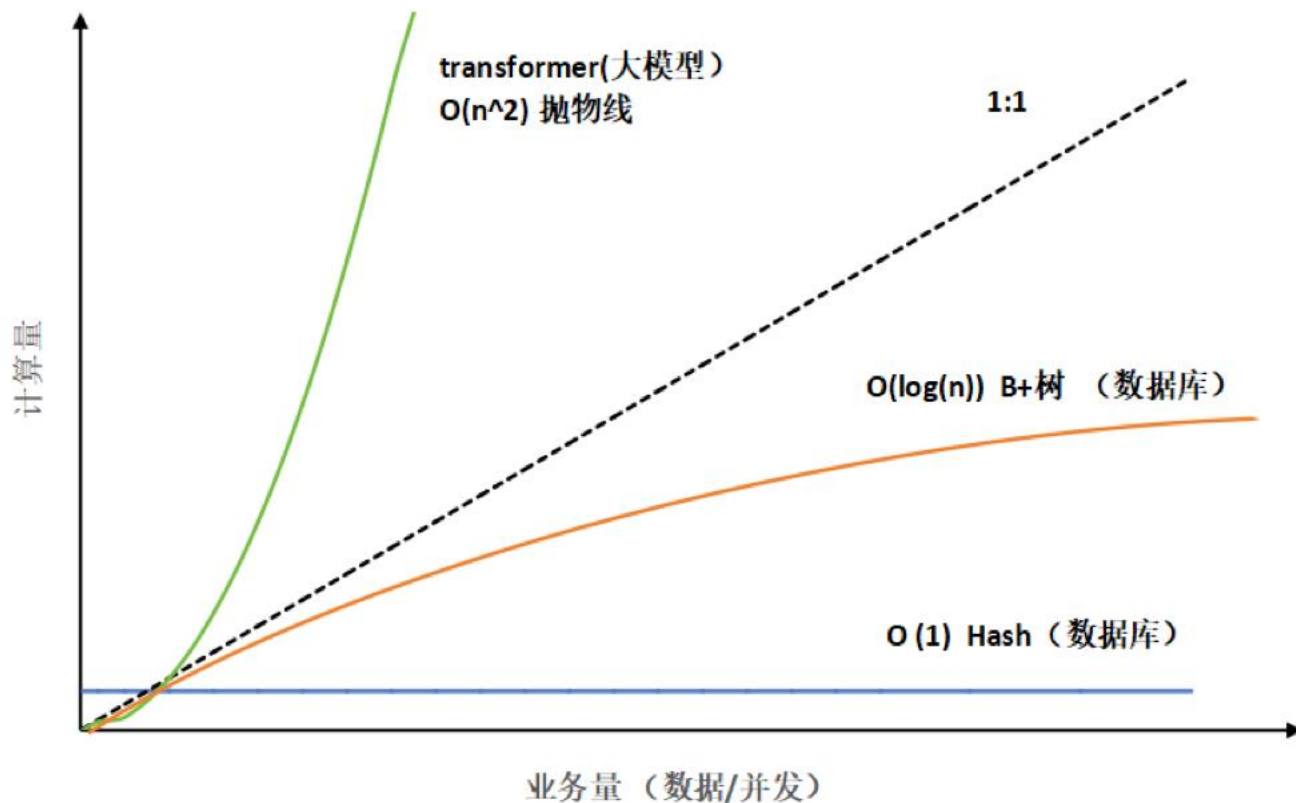
▶ 3.4 产中并发场景下性能的高衰减挑战

当前面临的核心技术挑战，是全场景一体机环境（含华为卡 / 英伟达卡）难以顶住高并发的性能瓶颈。这一问题的本质源于Transformer 数学模型的底层算法特性。

复杂度	说明
$O(n^2)$	平方级增长
Hash- $O(1)$	常数级增长
B + 树 $O(\log n)$	对数级增长
$O(n)$	线性级增长

随并发数增加，矩阵运算的计算量将呈近似平方级跃升

以此形成对算力资源的指数级压力是 DeepSeek 大模型产品天然存在的问题



▶ 3.4 产中并发场景下性能的高衰减挑战

Medusa 缓存机制

精确缓存

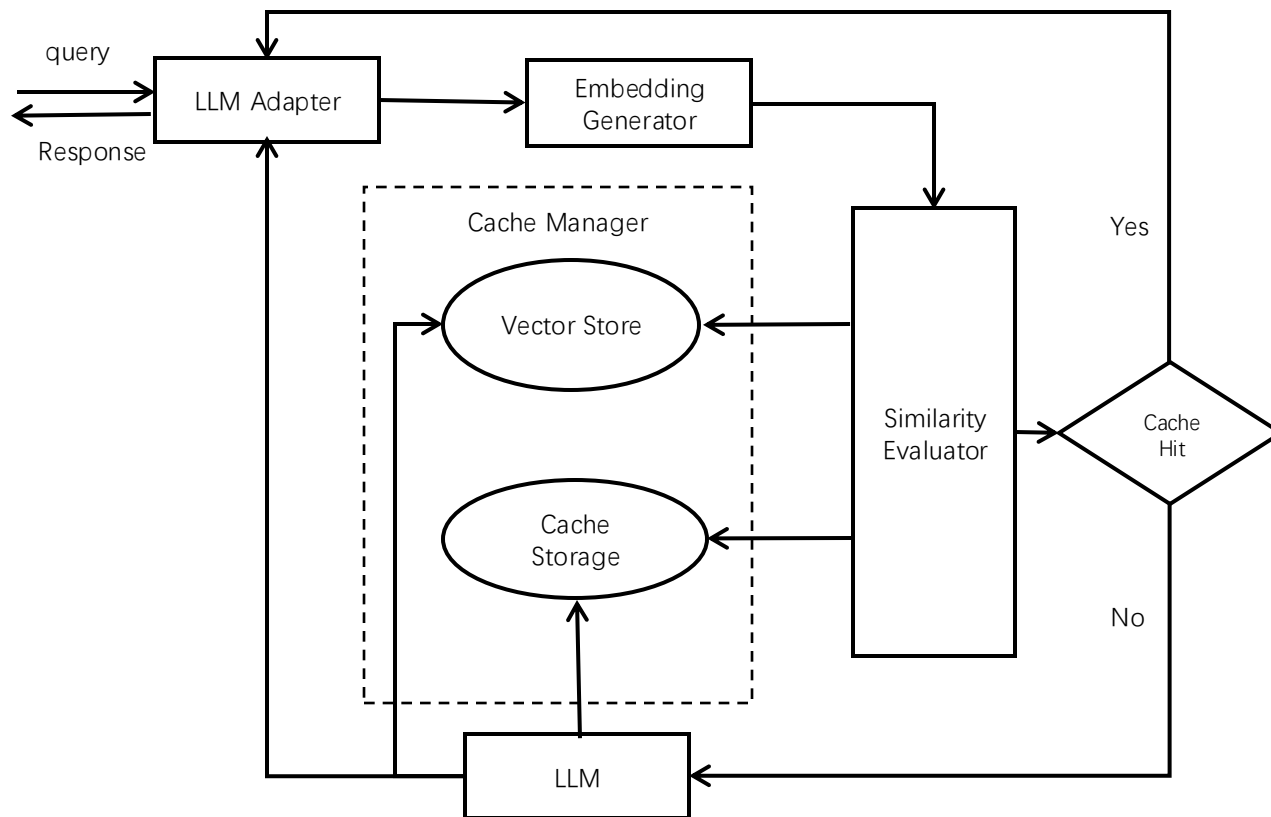
精确缓存存储已处理的项目，方便后续请求时直接复用，减少重复计算。

语义预读

通过语义预读用户下一步可能会问的问题，充分调配资源。

数据预加载

美杜莎类似于小模型，内置多轮迭代机制与预训练数据，并实现持久化存储，随着问答的进行不断迭代更新。



▶ 3.4 产中并发场景下性能的高衰减挑战

Medusa 模块训练与自动适配策略

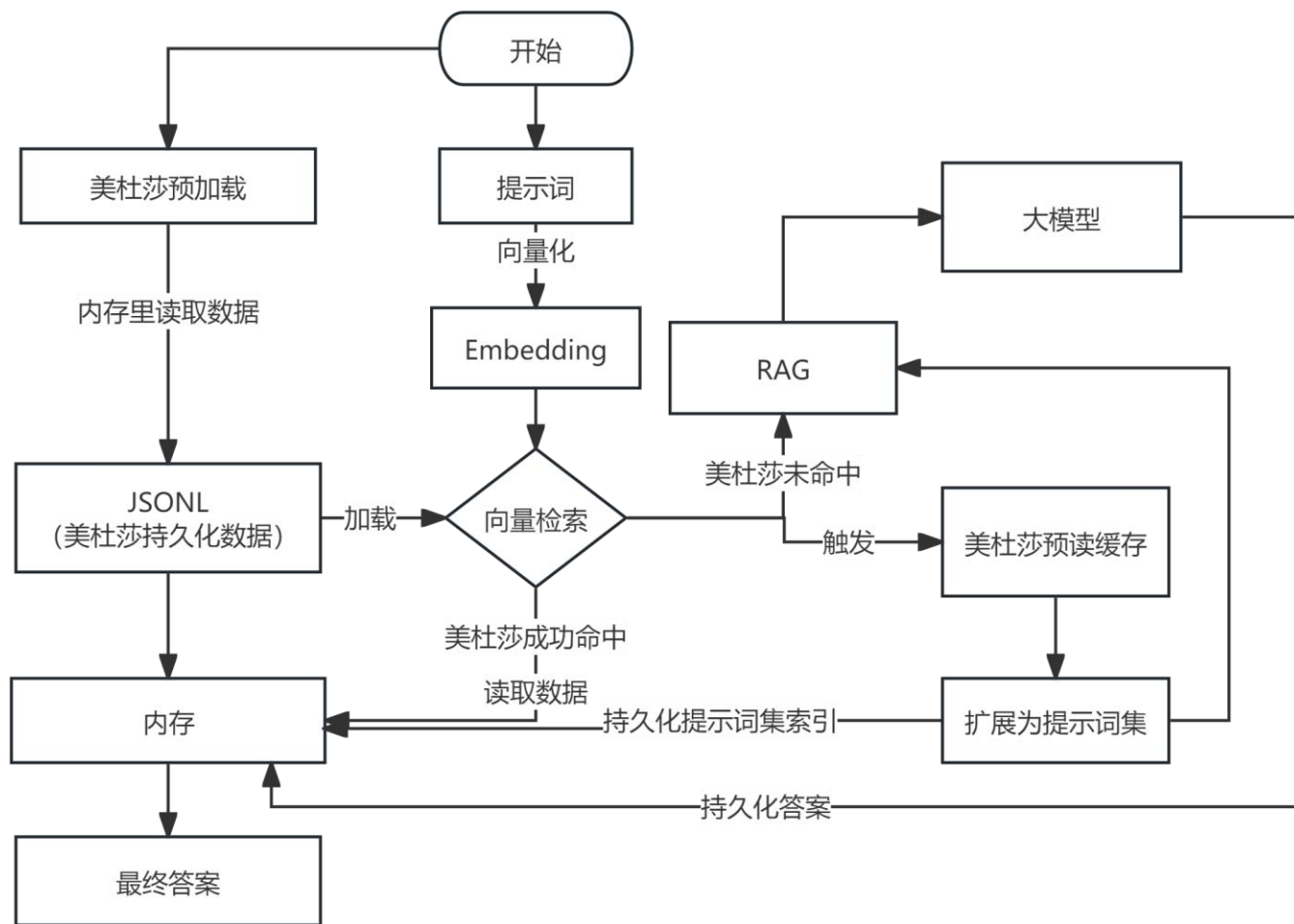
流程说明:

程序运行时，美杜莎模块自动预加载，从内存读取持久化数据。

用户访问中间件，中间件调用美杜莎模块，判断是否命中。若未命中，依据提示词预读后续问题，生成提示词集，再由中间件调配 LLM 与 RAG 模块生成答案并返回用户。

优点:

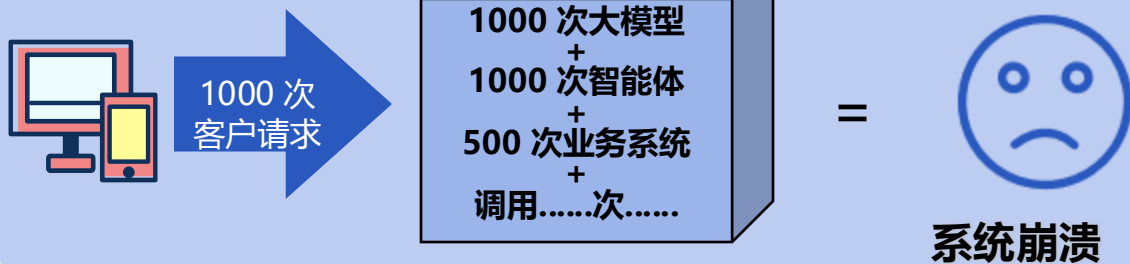
仅依赖于内存，并且内存占用较小，结构精简，减小大模型对算力的高度依赖，更容易实现向垂类大模型的转换。



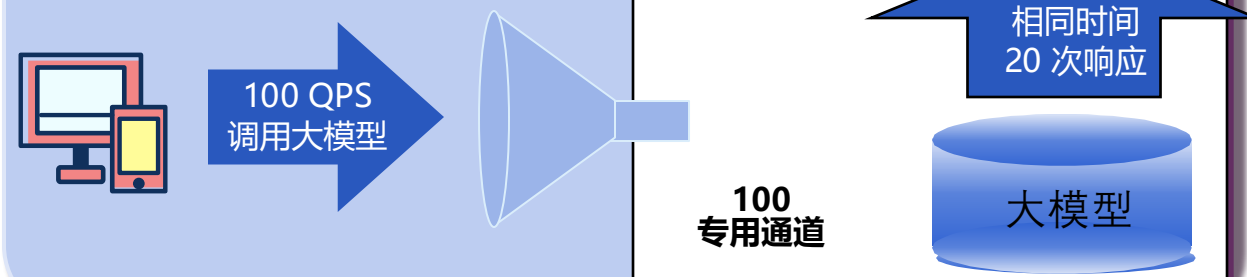
▶ 3.4 产中并发场景下性能的高衰减挑战

分发请求，负载均衡：可以依据意图和能力分布，动态将不同的业务请求分发到对应的模型、模块或智能体。智能分配请求，充分利用算力资源，提升系统性能。

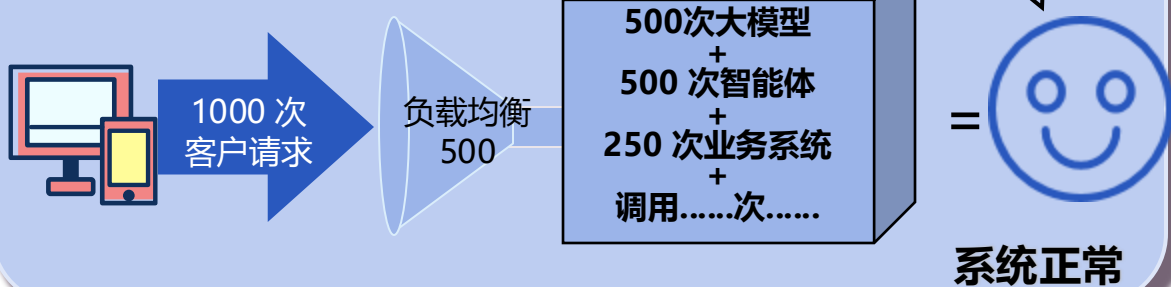
没有大模型中间件



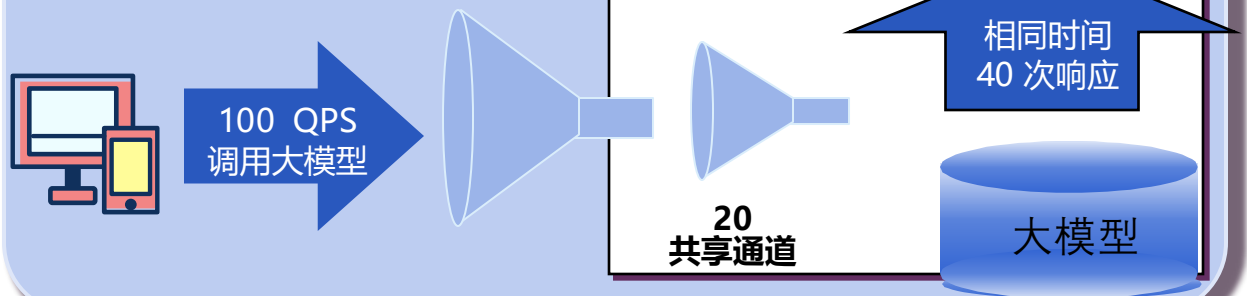
没有大模型中间件



有大模型中间件



有大模型中间件



▶ 3.5 产后的调度路由、安全与模态对齐

适配主流大语言模型能力概览

兼容主流大语言模型

支持市场上大多数主流大语言模型，帮助企业以低成本、高效率的方式实现定制与部署。

快速部署，简单配置

新增模型只需更新配置文件，配置完成后即可通过界面和接口快速调用大语言模型。

适配的大模型列表

 Landing	 Vicuna	 ChatGPT
 GPTAzure	 Gemini	 Qwen
 Baidu Ernie	 ChatGLM	 Moonshot AI
 Baichuan	 IFLYTEK	 SenseChat
 Doubao	 Deepseek	 Claude

配置示例

```
# qwen (TongyiQianwen) is an advanced large-scale language model developed by Alibaba Cloud
- name: qwen
  type: Alibaba
  enable: true
  model: qwen-turbo,
        qwen-plus,qwen-max,qwen-max-1201,qwen-max-longcontext,
        asr,
        vision,
        ocr

# Options for default, llm, voice, vision and ocr, line by line above.
driver: ai.wrapper.impl.AlibabaAdapter
# help document https://help.aliyun.com/document_detail/2712195.html?spm=a2c4g.2712576.0.0.733b3374np40s0
api_key: your-api-key # url address https://bailian.console.aliyun.com/#/home
access_key_id: your-access-key-id
access_key_secret: your-access-key-secret
```



▶ 3.5 产后的调度路由、安全与模态对齐

智能请求路由策略与动态分配机制

意图识别

根据用户输入判断其意图联合评分机制，将查询路由至适合的模型或智能体处理。

故障转移

在系统异常时，自动切换至备用模型或服务，确保稳定性。

负载均衡

智能分配请求，优化资源利用，提升系统性能。

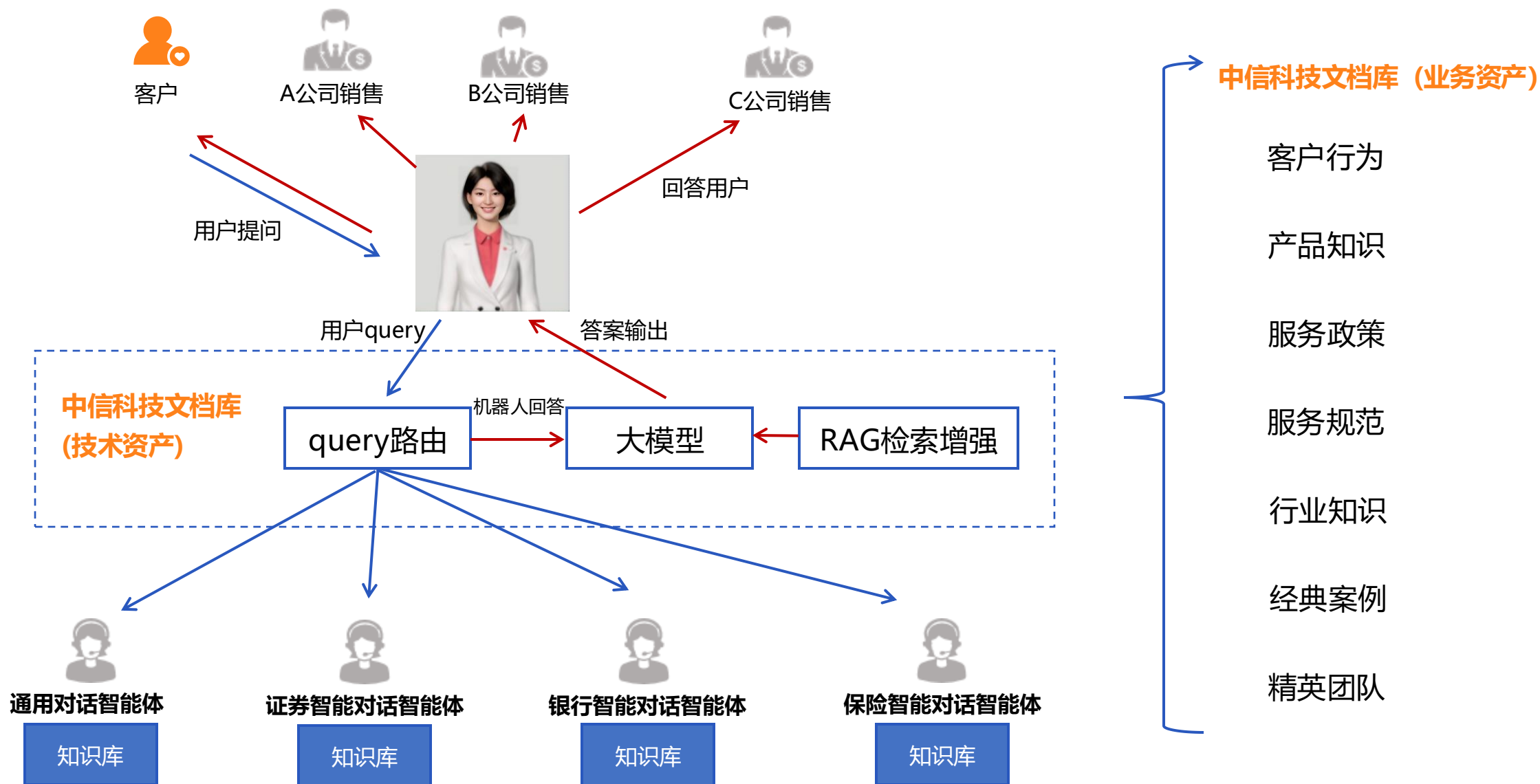
routers:

- name: best
rule: (|,&)
A|B -> polling, A or B, means randomly round-robin within A and B;
A,B -> failover, first A, if A fail, then B;
A&B -> parallel, call A and B concurrently, choose the proper only one result;
The rule could be combined like ((A&B&C),(E|F)) which means first call ABC simultaneously, if fail, then call E or F randomly.

- name: pass
rule: (%)
% is a wildcard to represent agent name.
if specified, invoke this agent.
if just % is given, % would be decided by parameter at calling time.



▶ 3.5 产后的调度路由、安全与模态对齐



▶ 3.5 产后的调度路由、安全与模态对齐

各个领域涉及的敏感词不同，通过以下方法，从中间件层灵活调整对话返回结果，优化业务场景应用：

敏感词过滤

在 sensitive_word.json 中添加需过滤的关键词。

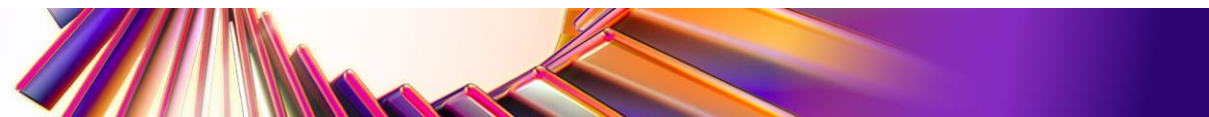
优先关键词

在 priority_word.json 中指定需优先回答的关键词。

停止关键词

在 stopping_word.json 中设置对话停止的关键词。
这些配置可引导对话朝特定方向发展，并在必要时自动停止对话。

```
filters:
# Sensitive words filter
sensitive:
mask: "*"
# 1:replace all returns with blank a string, 2:Replace the matched
strings with a mask string, 3:replace the matched strings with an
empty string
level: 2
rules:
- rule: 'openai'
# Vector index words with high priority
priority:
- car
- weather
# Conversational stop word
stopping:
- bye
# Conversational continuation words
continue:
- about
```



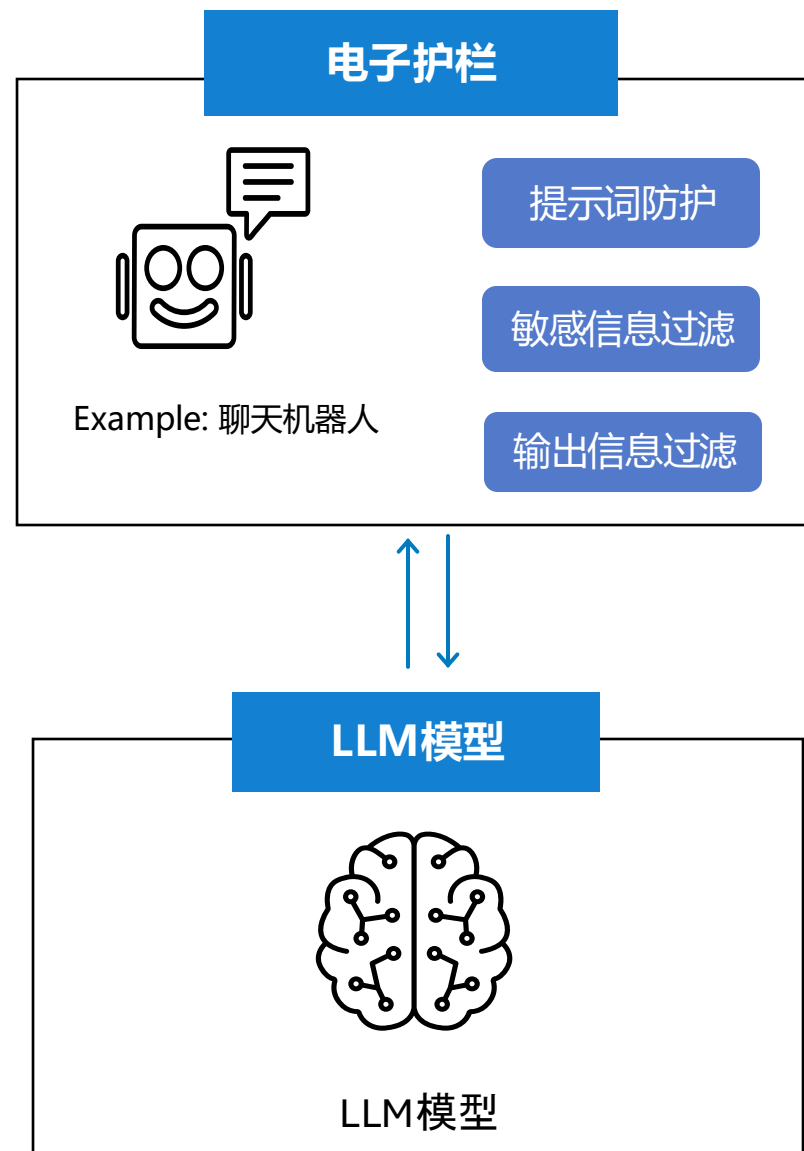
▶ 3.5 产后的调度路由、安全与模态对齐

- 本安全框架由 LinkMind 中间件支撑，统一协调四大安全模块。

- 构成闭环的模块：

- 防御性训练：模型训练阶段的预防策略
- 后安全过滤：模型输出的风险筛查
- 前安全保护：用户输入的风控机制
- 权限分级：身份和角色控制机制

- 构建多层次、全流程、可插拔的 AI 安全治理能力。



▶ 3.6 协同应用集成案例及其开发者价值

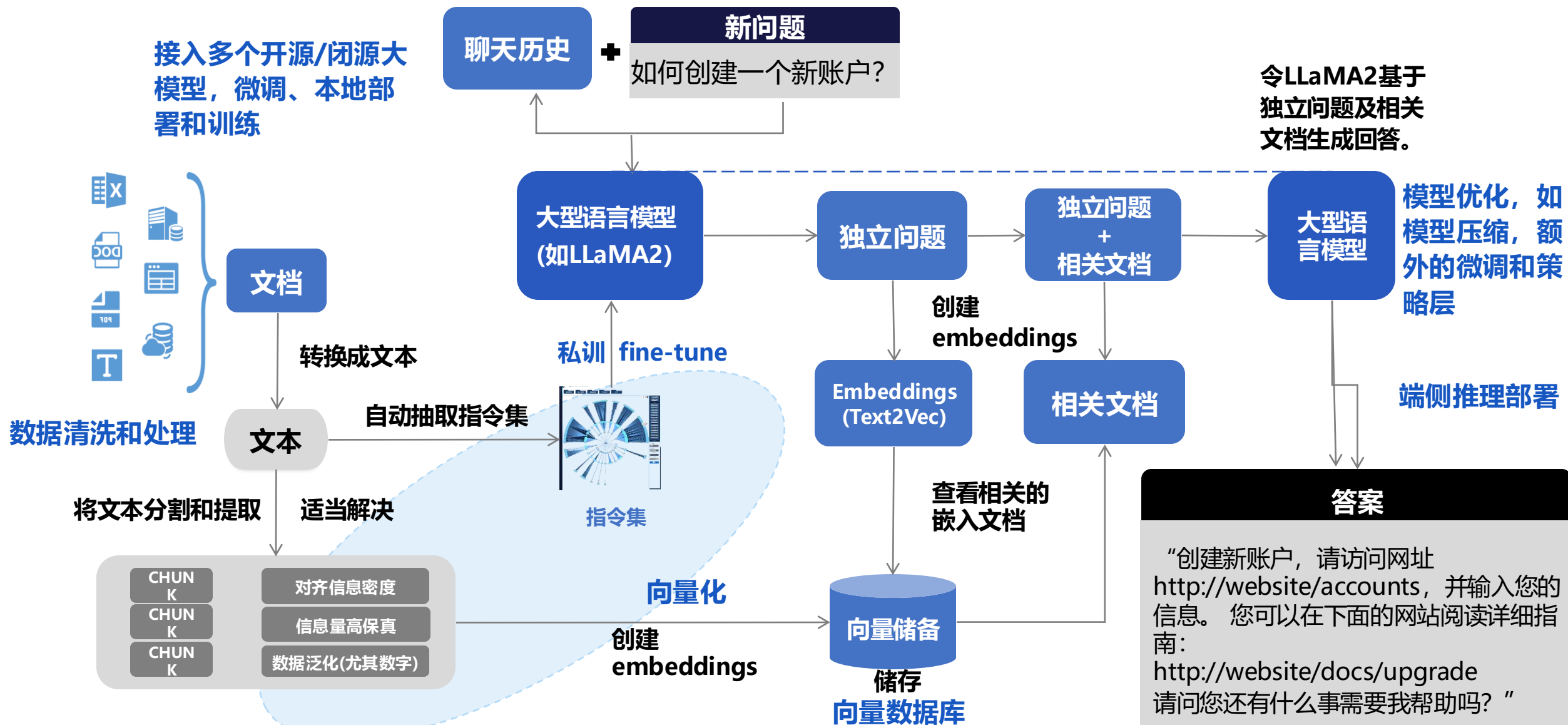
资料归集—辽宁联通内部文档系统

企业定制专属大模型，契合企业内部办公需求，个性化服务。

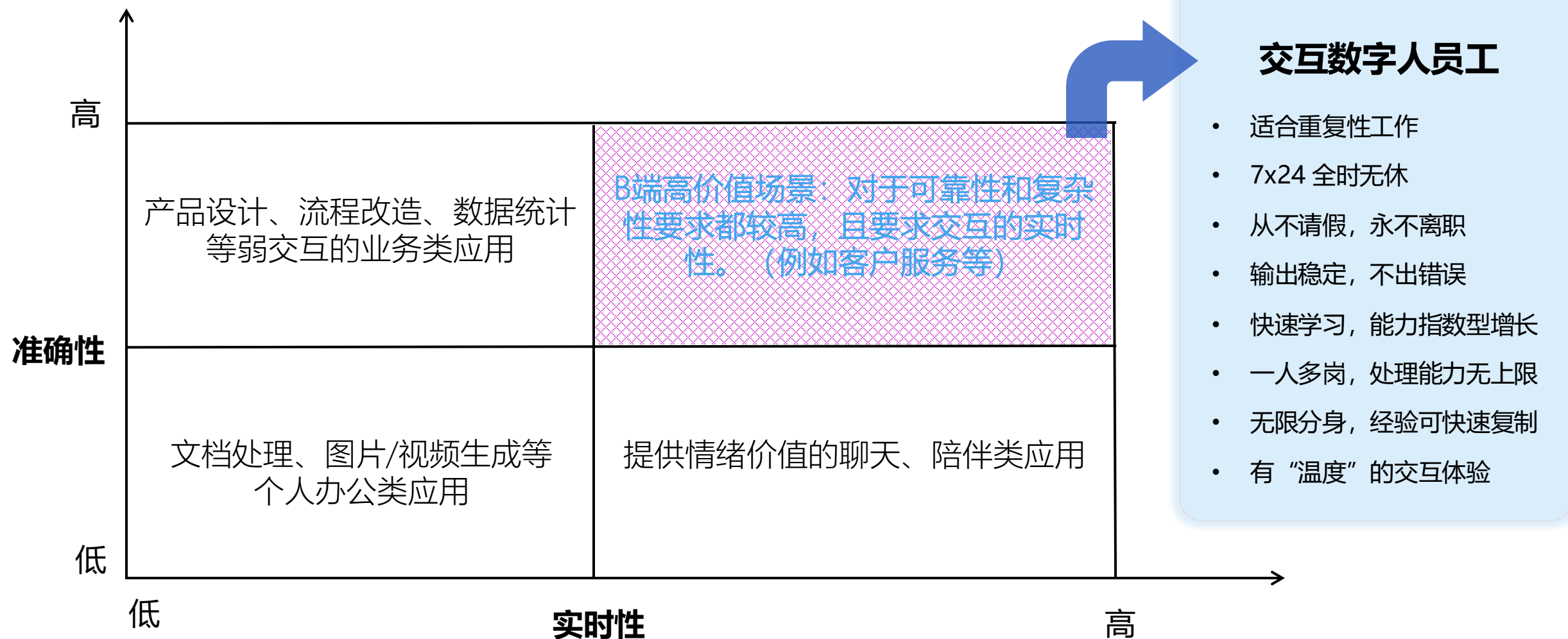
- 通过模型控制不同级别用户的文件权限，**提高文件管理安全性**；
- **智能问答**，根据权限不同，对相同问答作出不同的回答，用户获取的附件内容也不一致；
- **情景微调**，对用户不满意部分进行二次优化。



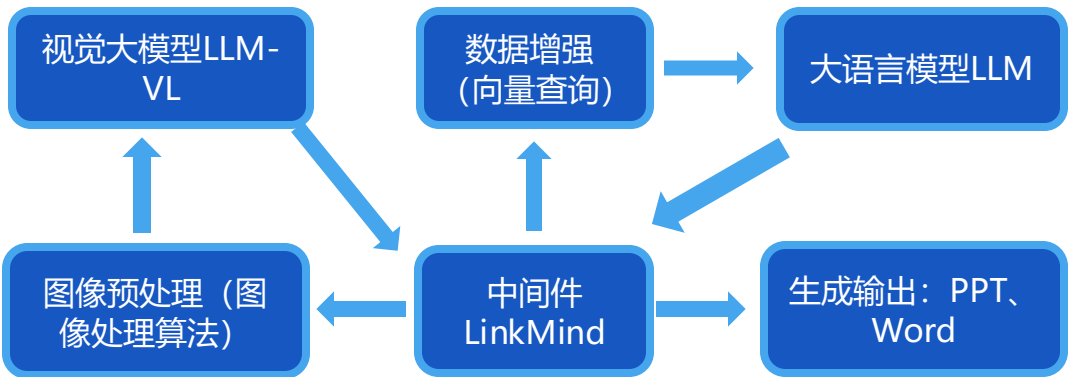
▶ 3.6 协同应用集成案例及其开发者价值



▶ 3.6 协同应用集成案例及其开发者价值



3.6 协同应用集成案例及其开发者价值

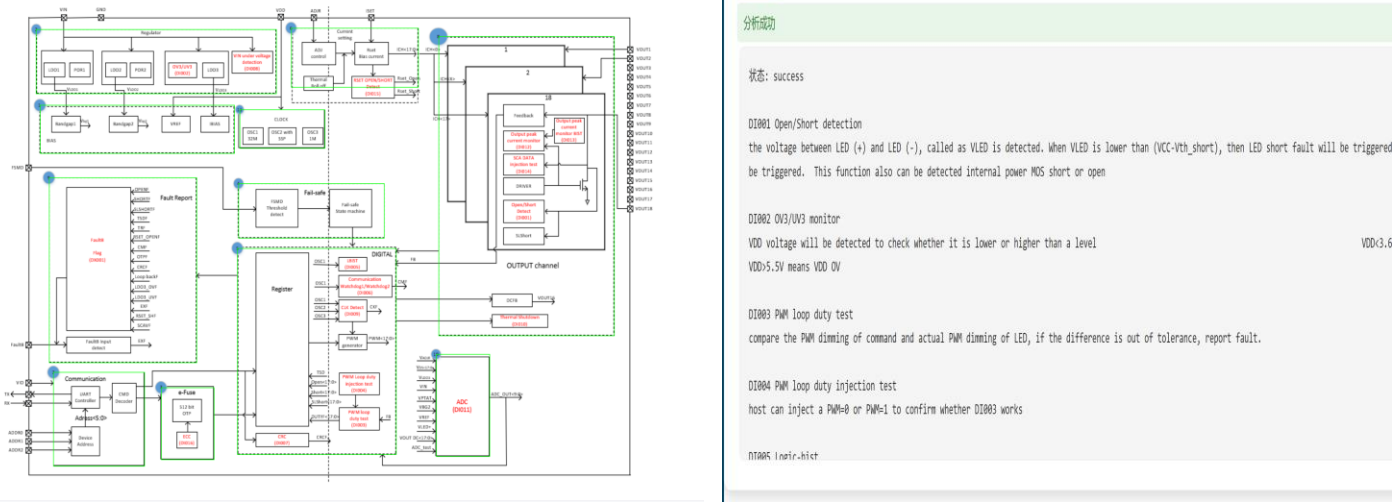


技术架构图

模块	功能与技术	功能描述
图文识别	LinkMind图像算法+LLM-VL+ OCR	识别BD图模块编号（子Block、DI/DX），提取结构信息
内容理解与匹配	LinkMind增强检索+LLM+Embedding	从library检索模块描述，支持模糊匹配和上下文理解
文档生成（Word）	LLM + 模板引擎 + 表格处理	自动填充Word表格（参数设置、HSR、Mapping表）
PPT生成	LLM + 文档解析 + 模板驱动	解析datasheet生成PPT（Product Overview、Block Diagram等）

交付成果与价值:

- 效率提升：批量生成Word/PPT，人工校对，节省80%时间
- 高准确性：大模型多轮匹配，避免字段遗漏或错配
- 强扩展性：适配多类datasheet，拓展至汽车电子、工业控制
- 标准化交付：输出符合客户模板，直接用于审查



▶ 3.6 协同应用集成案例及其开发者价值

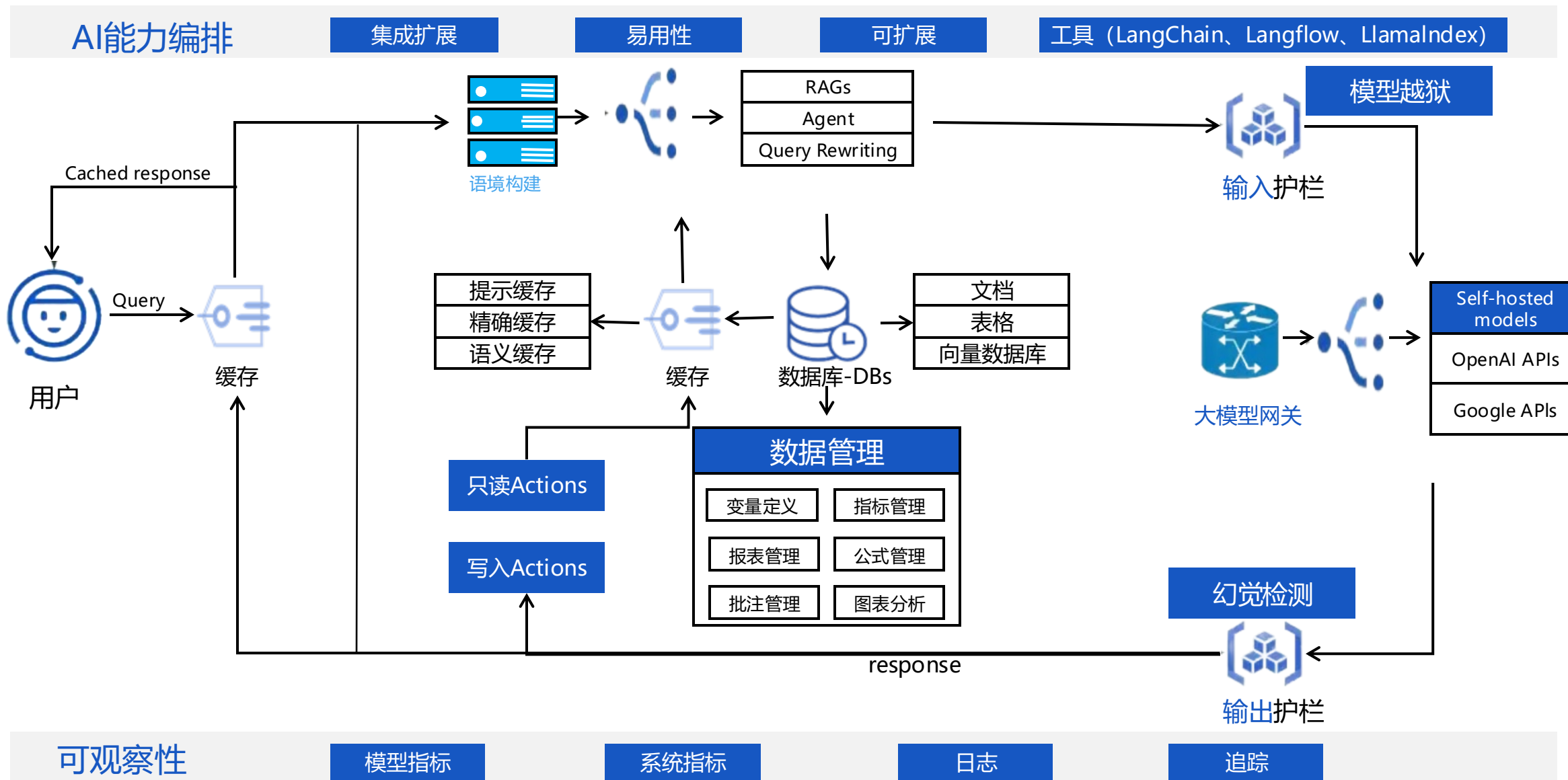
成功案例 - 兰州中川机场

兰州中川国际机场三期T3航站楼总建筑面积约40万平方米，是现有T1及T2航站楼面积总和的4倍，于2025年2月底投入使用。

为新航站楼提供了端到端的智能客服解决方案，包含数字人客服，问询机器人，引路机器人，洗地机器人与消杀机器人。行李机器人，是数字人客服最新引进的小助手，在数字人客服的调度下，能够独立将旅客的随身行李运送到指定地点等待领取，业内首个机场行业大模型，为设备装上了统一的“大脑”。



▶ 3.6 协同应用集成案例及其开发者价值



▶ 3.6 协同应用集成案例及其开发者价值

竞争产品:



LangChain



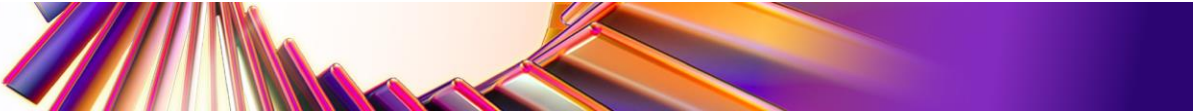
LlamaIndex



LinkMind

框架轻重	中	轻	轻	轻	中+
主要客户	框架级	开发级	开发级	AI开发者	企业级
检索增强框架	平凡RAG	平凡RAG	平凡RAG	非Graph的RAG	Graph+RAG
多模态对齐框架	无	无	无	无	有
公/私复用框架	无	无	无	无	有
软件语言	Python	Python	Java	python	Java
加速引擎	无	无	无	无	Medusa
向量模型	无	无	无	无	自研Embedding

结论：市场上还没有真正诞生满足中间件定义的产品(特别是在中国)





第8届AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

下一站预告

11/14-15 | 深圳站

12/19-20 | 上海站



查看会议详情

深圳站论坛设置

智能装备与机器人

超越“编程 Copilot”

下一代知识工程

智能网联与汽车智能化

AI 测试工具开发与应用

AI 基础设施和运维

数据智能及其行业应用

可信 AI 安全工程

大模型和 AI 应用评测

多 Agent 协同框架

从智能测试到自主测试

大模型推理优化

多模态 LLM 训练与应用

智能化 DevOps 流水线

上下文工程

AiDD

「深行 · 浅智」

Walk Deep, Think Light.

2025.11.16

AiDD首届麦理浩径徒步





科技生态圈峰会 + 深度研习

——1000+ 技术团队的选择



AiDD峰会详情





第7届AI+研发数字峰会
AI+ Development Digital Summit

感谢聆听!

扫码领取会议PPT资料

