《企业AI能力中枢—AICoreNexus》

顶层设计(愿景·目标·蓝图)

张子彪

2025-08-10

价值主张

各类模型 统一调度与治理

智能资产 持续沉淀与演进

定位与使命

企业AI能力的
"总线与中枢"

01 使命

企业级AI能力的统一治理平台

- 统一入口：标准化API，屏蔽底层差异
- 智能调度：多维度路由，自动负载均衡
- 集中管控：统一监控、成本、权限治理

02 定位

让AI能力集成从"复杂分散"变为"统一智能"

- 解决：多源异构、难以治理、难以度量、难以复用
- 实现：标准化接入、智能化调度、统一化治理、生态化扩展

03 愿景

成为企业AI治理的生态化平台

- 对内：AI能力统一治理与运营
- 对外：AI生态开放与能力输出
- 共建：插件生态与能力市场

背景与痛点

能力来源多元且异构

1.上游包括三类：多家 LLM 供应商、自研 AI 组件（抽取、分类、生成等）、自研或外部 传统 ML 模型服务。

2.协议/鉴权/数据格式/流式能力不一致，集成成本高、上层改造重。

复用沉淀不足

1.提示词工程、行业模板、业务插件难以沉淀与共享，重复造轮子、经验不可迁移。

成本与效果难平衡

1.Token/调用成本差异大且波动，缺乏按用户/应用/模型的细粒度成本归集与预算控制。

2.需要在“成本、延迟、效果”间动态权衡与自动化路由。

可观测性与运维割裂

1.健康、QPS、延迟、错误率、命中率、成本等指标分散在各处，缺少端到端可观测与统一告警。

产品全景

统一入口API

对上REST，对下多源适配

模型池与路由

标签/等级/健康/QPS/成本

插件化与服务注册

热插拔与动态路由

提示词模板中心

集中管理/热加载/复用

监控与成本治理

健康/QPS/命中/成本聚合

前端控制台

仪表盘/配置/发现/历史

六大亮点

01 **多模型统一管理**

元数据标准化/热加载

02 **智能路由**

多因素评分/Fallback

03 **插件生态**

@plugin_api/双入口/并发

04 **服务生态**

注册即入池/动态代理

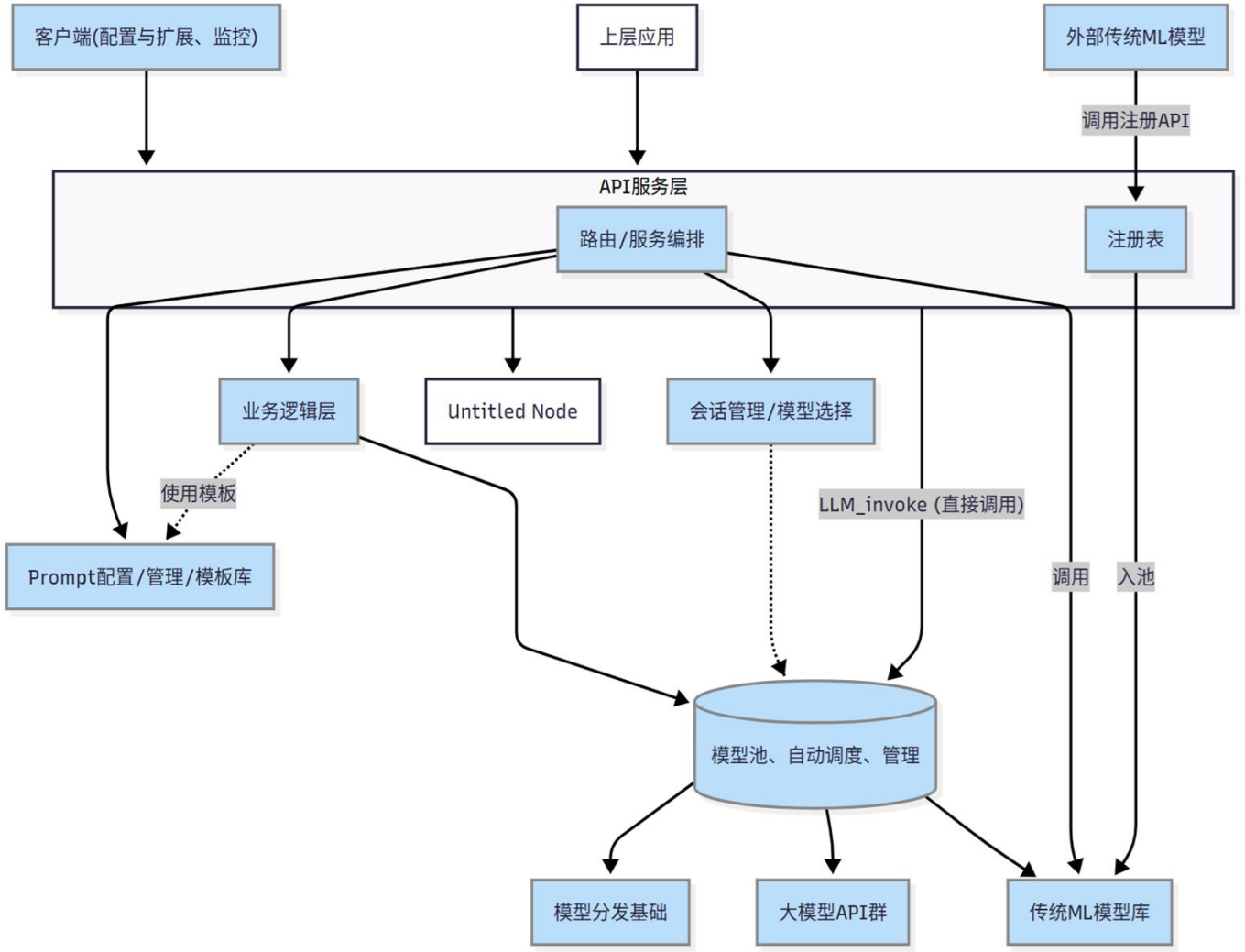
05 **Prompt中心**

模板沉淀/A-B/热更新

06 **可观可管**

QPS/健康/成本/历史

总体架构图


客户端(配置与扩展、监控)
上层应用
外部传统ML模型
调用注册API
API服务层
路由/服务编排
注册表
业务逻辑层
Untitled Node
会话管理/模型选择
使用模板
LLM_invoke (直接调用)
Prompt配置/管理/模板库
调用
入池
模型池、自动调度、管理
模型分发基础
大模型API群
传统ML模型库

关键模块

API服务层

入口/参数/路由/异常

业务逻辑层

会话/选择/批并发

LLM池与路由

状态/选择/限流/健康

插件及服务注册

扫描/注册/热加载

Prompt中心

模板加载/接口/前端

统计与状态

命中/成本/历史/偏好

差异化与优势

中枢定位

对上治理，对下适配

成本-效果平衡

可配置/可度量

开发者友好

插件/并发/Prompt中心

运维友好

全链路可观测

生态开放

自研+第三方

可演进

策略/架构扩展

路 线 图

一期

服务发现/插件生态/ /自适应路由

二期

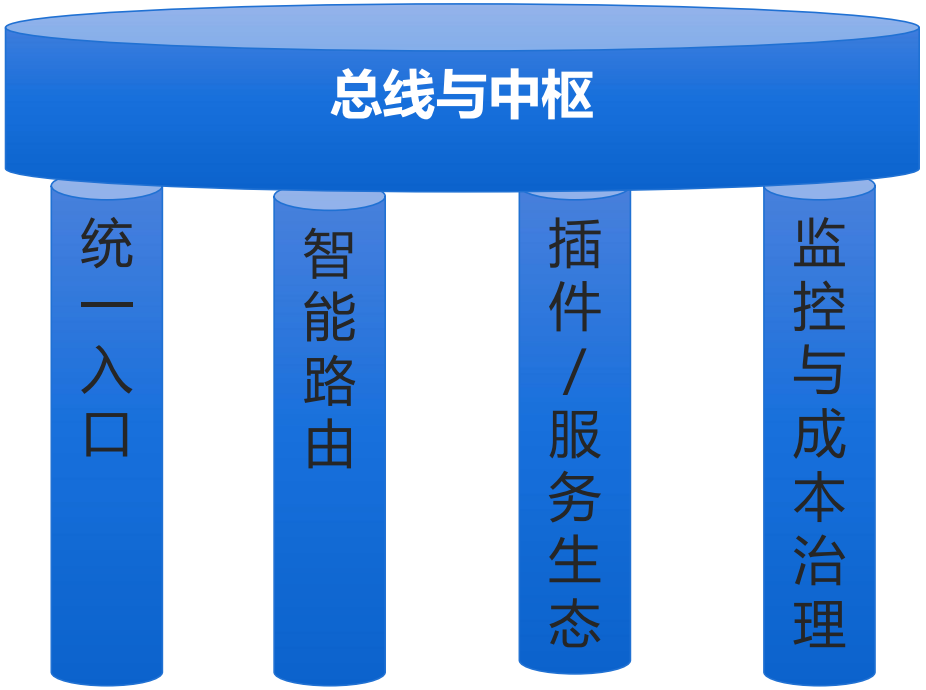
生态市场/安全审计/多租户

总 结

1、四根支柱

构建企业AI的“总线与中枢

企业AI演进的不息之道

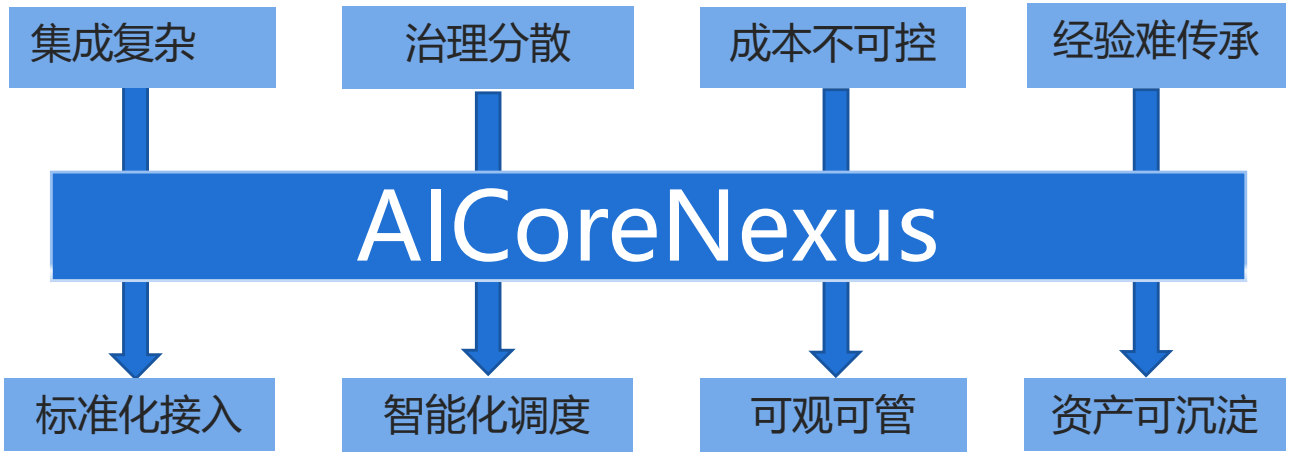


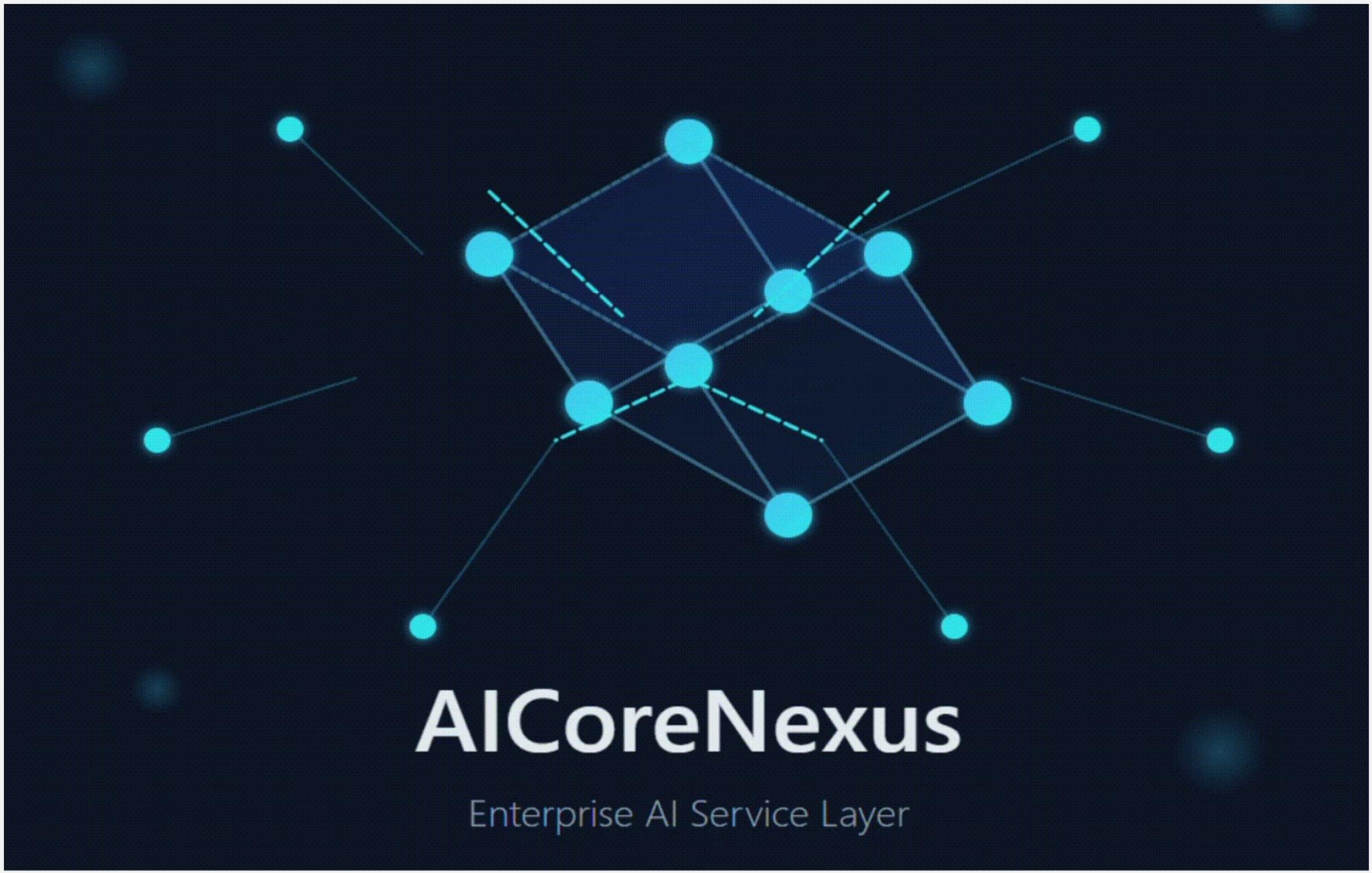
总 结

2、能力中枢

实现从问题到能力的转变

企业AI演进的不息之道




AICoreNexus
Enterprise AI Service Layer