东吴证券 SOOCHOW SECURITIES

房地产行业深度报告

地产+AI工具系列报告之六：面向 Agent 时代的工程师训练路径与国产大模型实践

增持（维持）

2026 年 07 月 17 日

证券分析师 姜好幸
执业证书：S0600525110001
jianghx@dwzq.com.cn
证券分析师 刘汪
执业证书：S0600526030001
liuwang@dwzq.com.cn

投资要点

- **Agent 应用落地的核心约束正从模型能力转向工程化能力：**据 Gartner 预测，到 2028 年，33% 的企业软件应用将包含 Agentic AI；预计到 2027 年底，超过 40% 的 Agentic AI 项目将被取消，主要原因包括成本上升、商业价值不清晰和风险控制不足。两组判断共同表明：企业侧对 Agent 应用的需求仍在提升，但真实落地过程中，可靠性、成本控制、风险治理和系统运维等工程问题将成为重要约束。我们认为，Agent 项目的关键挑战正在从“模型能否调用”逐步转向“组织是否具备将模型能力嵌入稳定业务链路的工程能力”。
- **切入 AI 教育“中间层”，补上系统化工程训练：**当前 AI 学习资源较为丰富，但更多分布在入门科普与论文 / 学术课程两端。前者有助于快速跑通 Demo，后者能够解释模型原理，但二者之间仍缺少将理论、代码、部署和业务可靠性串联起来的系统化工程训练。“AI 大模型学习之路”以 4 阶段、32 模块、10 个动手实验、8 个 Capstone 业务场景案例和 4 个毕业项目组织内容，覆盖数学基础、模型原理、推理优化、RAG、Agent、Agent Infra 与生产架构等主题，核心目标是帮助具备编程基础的开发者从“API 使用者”迁移至“AI 系统构建者”。
- **从零构建+失败排查，补足 Demo 到生产的关键断层：**平台设置 4 个毕业项目，形成“理解架构—训练模型—构建应用—部署优化”的能力链，具体包括从零实现 Transformer、预训练小参数模型、搭建完整 RAG 系统以及使用 vLLM 部署开源模型并进行压测。与此同时，平台在部分实验中引入调试挑战，覆盖注意力分数未缩放、梯度未裁剪、KV Cache 更新异常、RAG 切分策略导致召回污染等典型问题。该训练方式不只强调“跑通正确路径”，更强调问题定位、原因解释和修复能力，指向真实 AI 系统上线后常见的可靠性、成本、评估与运维挑战。
- **国产大模型深度适配，降低生态迁移成本：**国产模型生态正在从能力追赶逐步进入工程落地深化阶段。平台将 DeepSeek、GLM、Kimi、混元等国产模型相关实践嵌入 AI 助教、推理部署、Agent 开发、长上下文、模型评估和量化优化等模块，关注 Function Calling 差异、推理框架适配、量化后的精度—吞吐取舍、长上下文工程策略以及成本曲线管理等问题。其价值并非简单证明某类模型更优，而是将分散的工程经验沉淀为可复用知识，降低工程团队采用国产模型时的隐性迁移成本。
- **平台自身即 AgentInfra 教学样本：**平台并非仅通过课程讲解 Agent Infra，而是将学习过程本身组织为一个简化版 Agent 系统：AI 助教承担推理服务，边缘函数完成上下文注入，知识地图记录学习状态，间隔重复系统保存知识点掌握情况并触发复习，D1 承载学习进度与复习队列，边缘架构支持按需调用和低固定成本部署。学习者使用平台的过程，本身也在体验一个包含推理、上下文、状态、记忆、触发和成本治理等要素的最小化 Agent 系统样本。
- **工程知识基础设施是国产大模型下一阶段的关键补位：**从产业视角看，值得持续关注三类环节：一是围绕国产模型的推理框架、量化、Agent 编排等工具链完善程度；二是工程知识资产（课程、评估、案例）能否形成标准；三是相关能力认证、企业培训与人才供给链的对接。这些环节或比“模型参数榜单”更接近实际落地。
- **风险提示：**内容过时风险、成本口径风险、国产模型生态变化风险。

行业走势

相关研究

《房企投资谨慎，核心城市房价平稳》
2026-07-16

《房地产行业 2026 年中期策略——底部信号渐明，核心资产先行》
2026-07-09

请务必阅读正文之后的免责声明部分

内容目录

1. 为什么是现在：Agent Demo 到生产系统之间，隔着一整条工程链 4
2. AI 教育供给存在结构性缺口：工程师能力迁移路径仍待完善 5
2.1. 系统化断层：碎片较多，完整能力路径相对不足 5
2.2. 工程化断层：跑通 Demo 不等于具备生产级系统能力 5
2.3. 国产模型适配断层：模型能力提升后，工程知识体系需要同步完善 6
3. 平台定位：从 API 使用者到 AI 系统工程师 6
4. 实践体系：以实验与毕业项目补齐从认知到交付的能力闭环 9
5. 教学方法：既训练直觉，也训练排错 11
5.1. 直觉优先：降低理论学习门槛，建立工程化理解框架 11
5.2. 渐进式代码构建：从最小实现逐步过渡到完整系统 12
5.3. 引入调试挑战：从展示正确路径转向训练排错能力 13
5.4. 间隔重复机制：以记忆、状态和触发机制提升长期掌握效果 13
6. 国产大模型实践：降低从“可调用”到“可落地”的迁移成本 14
7. 平台架构具备 Agent Infra 教学样本属性 16
8. 技术架构：边缘优先、低固定成本、弱中心化依赖 17
8.1. 认证与访问控制：以边缘网关完成轻量化权限校验 18
8.2. 数据模型：以 D1 承载学习状态、复习队列和社区互动 19
8.3. 计算范式：V8 Isolate 有助于降低轻量 API 的部署与扩缩容复杂度 20
9. 竞争定位：差异化切入 AI 工程实践与国产模型适配场景 20
10. 发展规划：从学习平台走向工程知识基础设施 21
11. 风险提示 22

请务必阅读
更多免费行业报告 并购家 www.ipoipo.net

图表目录

图 1: 平台首页示意图 7
图 2: 知识地图 | 5 层视图（基础/核心/训练/应用/前沿）+ 前置依赖 + 跳过测试 8
图 3: 动手实验 | 10 个实验列表，含难度筛选器与技术栈标签 9
图 4: Capstone 项目库 | 8 个业务场景案例 11
图 5: 模块阅读页 | 对比表 + 数学推导 + 直觉理解框 + 代码实现的四位一体教学 12
图 6: 渐进式代码 | 完整实现 + 详尽中文注释 13
图 7: 间隔重复系统数据流示意图 14
图 8: AI 助教 | 浮窗式对话，实时解答与代码调试建议 15
图 9: AI 助教请求链路示意图 16
图 10: 系统架构总览 17
图 11: 平台登录页示意图 18
图 12: 认证流程时序示意图 19
图 13: D1 数据模型示意图 19
图 14: 边缘计算 vs 传统容器 | 冷启动、内存占用、扩缩容与月成本对比 20

表 1: 平台 4 阶段学习内容概览 8
表 2: 平台五层技术架构 18

请务必阅读正文之后的免责声明部分

1. 为什么是现在：Agent Demo 到生产系统之间，隔着一整条工程链

以典型企业试点场景为例，一名后端工程师可以在较短时间内基于 LangChain + RAG 搭建基础 Demo：完成 PDF 上传、文本切分、向量库写入、模型接入，并增加简单的工具调用能力。在演示阶段，系统往往能够较顺畅地完成问答交互，初步验证方案可行性。

但进入真实业务链路后，系统性问题会逐步显现。随着文档规模扩大，检索结果可能混入旧版本资料或低相关内容；在多轮对话场景下，上下文持续累积，推理成本和响应稳定性面临压力；工具调用出现偶发异常时，系统需要具备重试、降级或向用户确认的机制；当底层模型切换至不同国产模型接口时，Function Calling 的参数格式、错误返回、兼容性和稳定性也可能存在差异。更具挑战的是，AI 系统并非总是稳定出错，而是可能在少数情况下生成表面合理、但事实并不准确的回答，这对生产环境下的可观测性、评估机制和风险控制能力提出了更高要求。

由此可见，**跑通 Demo 只是 AI 工程的起点。真正的难点在于，如何将模型能力嵌入一条稳定、可控、可观测、可维护的业务链路中。**

这也是当前 AI 教育体系中较容易被低估的断层。多数教程能够帮助学习者快速完成模型调用、聊天界面搭建和向量库接入，但较少系统训练以下能力：

- 如何判断模型错误来自 Prompt、检索、上下文、工具调用还是数据源；
- 如何控制长上下文和多轮对话带来的成本压力；
- 如何设计 Agent 的状态、记忆、回退和异常恢复机制；
- 如何选择推理框架、部署方式与量化策略；
- 如何将海外框架经验迁移至国产模型生态；
- 如何从“能跑”进一步走向“能上线、能压测、能维护”。

据 Gartner 预测，到 2028 年，33% 的企业软件应用将包含 Agentic AI；与此同时，其亦预计到 2027 年底，超过 40% 的 Agentic AI 项目将被取消，主要原因包括成本上升、商业价值不清晰和风险控制不足。两组判断并不矛盾：一方面，企业侧对 Agent 应用的需求持续提升；另一方面，Agent 项目在真实落地过程中会更加集中地暴露工程化能力短板。

我们认为，**Agent 落地的核心约束正从“模型能力是否足够强”，逐步转向“组织是否具备足够的工程人才、系统架构能力和运维治理能力”。**

请务必阅读

更多免费行业报告 并购家 www.ipoipo.net

2. AI 教育供给存在结构性缺口：工程师能力迁移路径仍待完善

当前 AI 学习资源供给较为丰富，但工程师在学习和落地过程中仍容易出现能力断层。其核心原因并非信息不足，而是不同知识点之间缺乏系统化结构和可迁移路径。

入门内容供给较为充足，例如快速理解 Transformer、短时间跑通 RAG、搭建基础 Chatbot 等。这类内容降低了学习门槛，有助于学习者完成第一步验证，但也可能带来一定误区，即将 AI 工程简化理解为 Prompt、API 和若干应用框架的组合。

学术内容同样较为丰富。论文、公开课和开源实现可以较完整地解释模型结构、训练方法和评估指标，但通常默认学习者具备一定数学与工程基础，同时并不直接覆盖生产部署中的具体工程问题。

因此，真正稀缺的是位于两者之间的“工程实践中间层”：**能够将理论、代码、部署和业务可靠性串联起来的系统化训练。**

2.1. 系统化断层：碎片较多，完整能力路径相对不足

工程师从零学习大模型，通常需要在数学基础、深度学习、Transformer、Embedding、RAG、微调、推理优化、Agent 和 Infra 等主题之间频繁切换。每个主题都能找到大量资料，但能够将其连成一条完整能力路径的内容相对有限。

这容易导致“看过很多，但难以串联”的问题：学习者可能了解 Attention 的公式，但未必清楚其与 KV Cache、推理效率之间的关系；可能知道 RAG 的基本流程，但对召回污染、上下文压缩和结果评估理解不足；可能知道 Agent 可以调用工具，但缺少处理多步失败、状态漂移和异常恢复的实践经验。

系统化学习的价值不仅在于节省时间，更在于帮助工程师建立一张可迁移的认知地图。只有理解各类知识点在工程链路中的位置，学习者才更容易在面对新模型、新框架和新问题时快速定位问题来源，并形成相对稳定的解决路径。

2.2. 工程化断层：跑通 Demo 不等于具备生产级系统能力

AI 工程能力的核心不只是“首次跑通”，而是当系统出现问题时，能够判断问题来源并选择合适的排查路径。在真实业务中，AI 系统常见问题包括：

- 检索系统召回错误文档，模型却生成较为自信的回答；
- 用户连续追问后，历史上下文对当前判断形成干扰；
- 工具调用超时或失败，Agent 缺乏合理降级机制；
- 模型输出格式不稳定，导致下游解析失败；
- Token 成本在低流量阶段不明显，业务放量后快速上升；

- 量化后系统吞吐提升，但部分中文任务质量出现下降。

上述问题难以仅通过复制示例代码解决，需要学习者同时理解模型机制、系统边界和工程取舍。因此，AI 工程教育不应只展示“正确路径”，也需要系统呈现“失败路径”，帮助学习者在调试、评估和优化过程中形成工程判断力。

2.3. 国产模型适配断层：模型能力提升后，工程知识体系需要同步完善

国产大模型能力仍在持续演进。DeepSeek、GLM、Kimi、混元等模型在推理、代码生成、长上下文、中文理解和成本效率等维度不断优化，在部分场景中已具备较强的应用价值。

但在实际落地过程中，工程师面临的问题往往不再是“能否调用模型”，而是更具体的工程适配问题，例如：

- 模型在 vLLM 等推理框架上的部署方式、吞吐表现和显存占用情况；
- 长上下文场景下如何进行切分、压缩、缓存和成本控制；
- Function Calling 与 OpenAI 接口的兼容程度及差异边界；
- W4A8、INT8 等量化策略在具体任务中的质量损失和性能收益；
- Agent 多轮规划过程中，工具调用的稳定性和异常处理机制；
- 不同模型在中文知识问答、代码生成和复杂推理任务中的能力边界。

如果上述经验长期分散在个人实践记录、社群讨论和零散博客中，将在一定程度上增加整个生态的隐性迁移成本。国产模型要实现更大范围的产业落地，除了模型厂商持续提升能力上限外，也需要围绕模型调用、推理部署、Agent 编排、量化优化和评估体系形成更加系统化的工程知识沉淀。

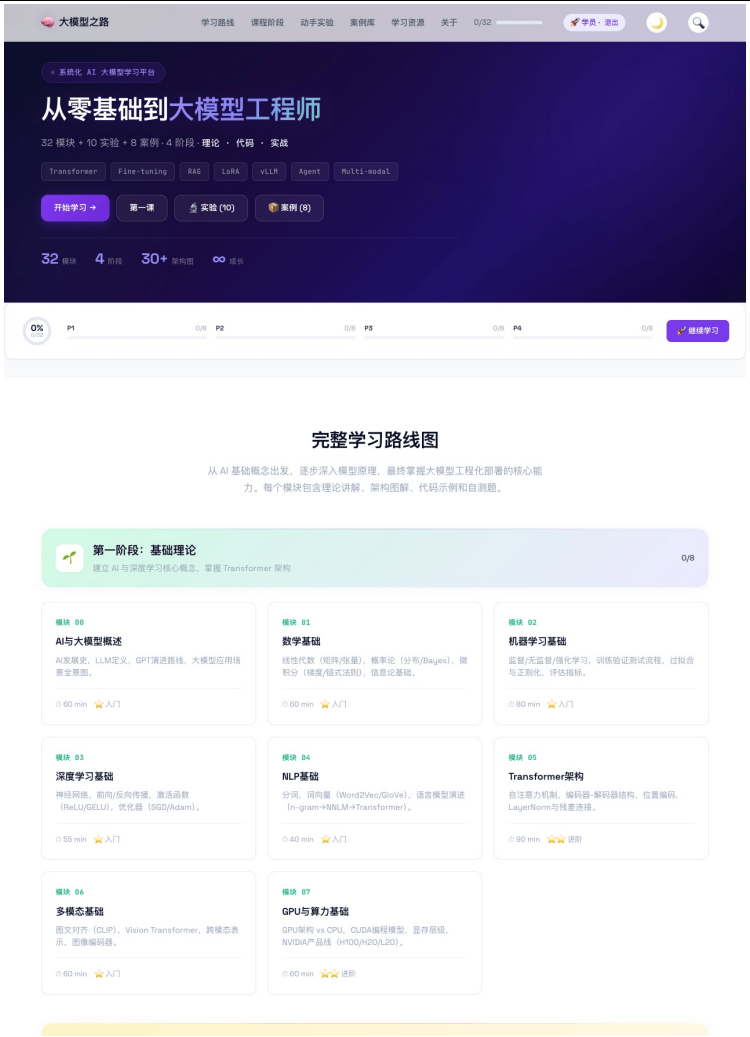
这正是“AI 大模型学习之路”所选择的切入位置。

3. 平台定位：从 API 使用者到 AI 系统工程师

AI 大模型学习平台（地址：https://ai-llm-learning.pages.dev/）主要面向已经具备一定编程基础、希望系统进入 AI 工程领域的开发者和工程师，而非泛 AI 兴趣用户。其核心定位在首页即有所体现：登录后首先呈现“从零基础到大模型工程师”的学习主张，并通过“32 模块、4 阶段、30+ 架构图”等信息强化内容体系的完整性。随后，页面展示第一阶段基础理论的 8 张模块卡片。整体来看，Hero 区、内容规模和模块阵列形成递进式表达，帮助学习者在第一屏即判断课程面向对象、内容深度和学习路径。

请务必阅读
更多免费行业报告 并购家 www.ipoipo.net

图1：平台首页示意图



数据来源：AI 大模型学习平台，东吴证券研究所

平台的训练目标并非仅帮助学习者更快搭建一个演示性应用，而是推动其理解 AI 系统从底层机制到生产部署的完整链路。其交付重点不是单个课程模块的堆叠，而是一条面向工程师的能力迁移路径：

- 从“会调用模型 API”，转向“理解模型输出机制”；
- 从“能搭建 RAG Demo”，转向“能判断召回、上下文和生成质量的问题来源”；
- 从“能接入 Agent 框架”，转向“能设计工具编排、状态管理和错误恢复机制”；
- 从“知道国产模型名称”，转向“理解不同模型在部署、成本、长上下文和工具调用等维度的工程差异”。

请务必阅读正文之后的免责声明部分

内容体系以 4 个阶段、32 个模块组织，覆盖从基础理论到工程实践、再到前沿架构的完整学习路径。

表1：平台 4 阶段学习内容概览

阶段	核心主题	训练目标
Phase 1	数学基础与神经网络	建立线性代数、微积分、概率、优化、神经网络的工程直觉
Phase 2	Transformer 与大模型核心技术	理解 Attention、Embedding、分词、预训练与模型结构
Phase 3	应用、优化与 Agent	掌握微调、RAG、推理优化、量化、Agent 与 Agent Infra
Phase 4	实战与前沿	进入 MoE、分布式训练、安全对齐、边缘部署与生产架构

数据来源：东吴证券研究所整理

在模块设计上，平台多数内容按照“直觉理解—公式推导—数字案例—代码实现—实验验证—失败排查”的节奏展开。该设计的核心价值在于，AI 工程师既需要理解抽象理论，也需要具备代码实现和问题排查能力，稳定的工程能力往往来自理论认知与工程实践之间的反复连接。

为降低长周期学习中的路径迷失，平台在首页底部设置交互式知识地图。该地图按 5 层组织，包括基础层（数学 / ML / 深度学习）、核心层（NLP / Transformer / GPU）、训练层（预训练 / 分词 / 注意力深入）、应用层（微调 / 推理部署 / RAG）和前沿层（Agent）。节点之间通过虚线展示前置依赖关系，节点自身则通过颜色区分“已完成 / 可学习 / 需先修”等状态。对于已经具备相关基础的学习者，平台还提供“跳过测试”功能，以支持其更高效地进入后续内容。

图2：知识地图 | 5 层视图（基础/核心/训练/应用/前沿）+ 前置依赖 + 跳过测试

数据来源：AI 大模型学习平台，东吴证券研究所

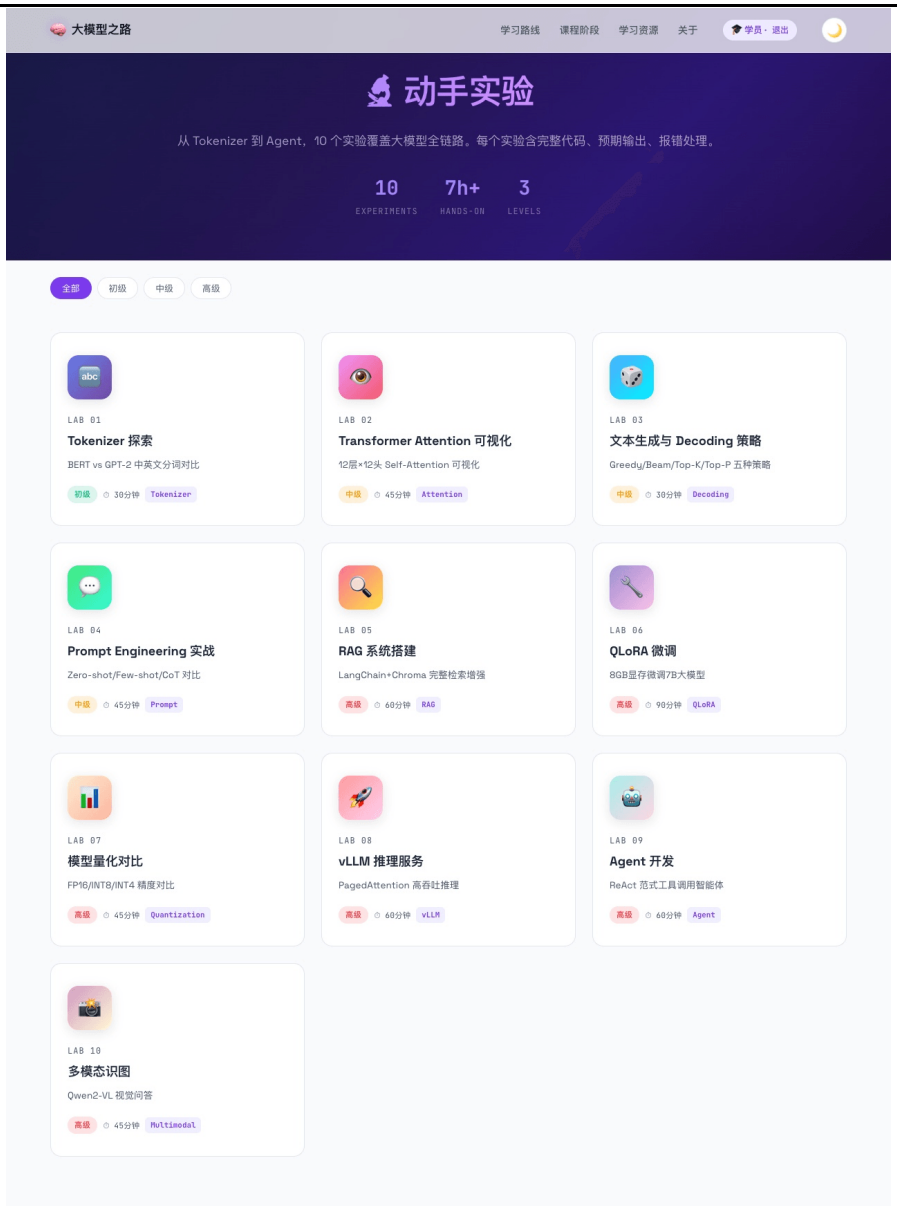
请务必阅读

更多免费行业报告 并购家 www.ipoipo.net

4. 实践体系：以实验与毕业项目补齐从认知到交付的能力闭环

平台的关键训练环节并不止于阅读和理解，而是强调面向真实工程问题的构建能力。除 32 个模块外，学习者还需要通过 10 个动手实验进行反复练习，内容覆盖 PyTorch 张量、Tokenization、Attention 与 KV Cache 最小实现、LoRA 微调、RAG 检索、vLLM 部署、Agent 工具调用等场景。页面顶部设置难度筛选器和技术栈标签，便于学习者根据自身能力缺口选择对应实验。

图3：动手实验 | 10 个实验列表，含难度筛选器与技术栈标签



数据来源：AI 大模型学习平台，东吴证券研究所

在实验体系之上，平台进一步设置 4 个毕业项目，对应 AI 工程链路中的关键层级。

项目 1：从零实现 Transformer。学习者需要实现注意力机制、编码器 / 解码器结构以及基础训练流程。其意义不在于重复已有框架能力，而在于帮助学习者理解模型能力的结构来源。只有真正实现过 Attention 机制，后续讨论 KV Cache、上下文长度和推理成本时，才更容易建立工程层面的理解。

项目 2：从零预训练一个小参数模型。该项目模型规模相对有限，但训练链路较为完整，涵盖数据准备、分词、训练、损失函数、优化器、评估和调参等环节。学习者可以由此理解，模型训练并非不可解释的过程，而是一组可观察、可调试、可权衡的工程步骤。

项目 3：搭建完整 RAG 系统。该项目训练重点不是简单接入向量库，而是围绕检索质量、上下文拼接、召回评估、生成控制和失败处理展开。学习者需要直接面对 RAG 在实际应用中较常见的问题，包括召回不准确、文档版本过旧、上下文过长、答案难以追溯以及模型过度自信等。

项目 4：使用 vLLM 部署开源模型并进行压测。该项目将学习者进一步带入生产部署层面，关注吞吐、延迟、显存、批处理、量化、缓存和成本等指标。对于企业级 AI 应用而言，系统上线后的关键约束往往不只是模型调试，也包括推理效率、资源利用率和成本曲线管理。

上述 4 个项目共同形成一条较清晰的能力链：

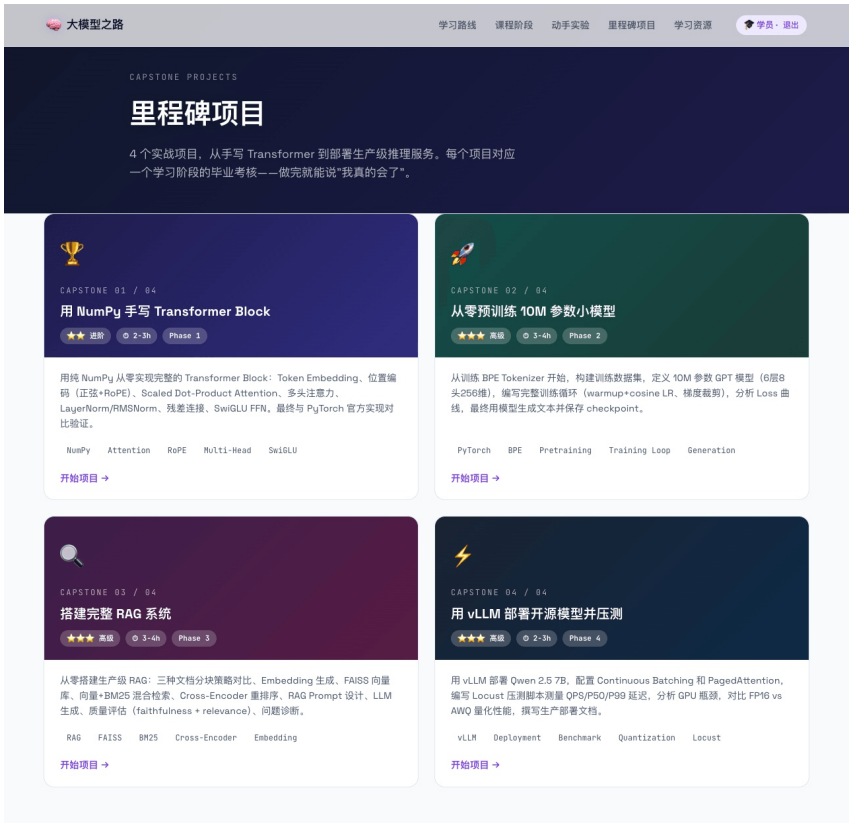
理解架构 → 训练模型 → 构建应用 → 部署优化

完成该路径的学习者未必能够立即承担复杂 AI Infra 系统建设，但有望建立继续深入所需的基础认知框架。其能力边界不再停留于“会使用某个框架”，而是能够理解框架背后对应的系统问题和工程取舍。

从训练目标看，4 个毕业项目主要回答的是“是否完整做过 Transformer、训练、RAG 和部署”等能力链问题。在此基础上，平台还提供 8 个 Capstone 业务场景案例，包括 AI 问答系统、文档摘要、代码生成助手、情感分析和多模态应用等，作为学习者进行项目沉淀和能力展示的出口。前者侧重夯实底层能力，后者侧重形成可展示的业务化作品。

请务必阅读
更多免费行业报告 并购家 www.ipoipo.net

图4：Capstone 项目库 | 8 个业务场景案例



数据来源：AI 大模型学习平台，东吴证券研究所

5. 教学方法：既训练直觉，也训练排错

AI 工程栈仍处于快速迭代阶段。Prompt 设计、推理框架、量化策略、Agent 框架以及模型 API 均在持续变化。对于工程师而言，仅掌握某一时点的实现方式并不足以支撑长期能力迁移，更重要的是形成面对新模型、新框架和新问题时的工程直觉与排查方法。

5.1. 直觉优先：降低理论学习门槛，建立工程化理解框架

数学基础往往是工程师进入大模型领域的第一道门槛。平台并未回避数学内容，但也避免将其处理为偏论文式的复杂推导，而是按照“先建立直觉、再引入公式、最后通过数字案例验证”的顺序展开。

以矩阵乘法为例，平台先将其解释为“批量线性变换的流水线”，再引入 C＝AB 的数学表达和逐元素计算过程；对于梯度下降，则通过可视化图示和具体数值演算，展示参数如何逐步逼近较优解。此类设计并非降低理论深度，而是为学习者进入复杂系统建立更清晰的认知入口。

以模块 05《Transformer 架构》为例，页面首先通过五维对比表说明 Transformer 相

较于 RNN / LSTM 在并行性、长距离依赖、复杂度、内存占用和最大长度等方面的差异；随后进入自注意力机制的数学推导，包括 Q / K / V 矩阵维度、四步计算流程，并以“我爱你”三个词作为示例进行直觉化演算；最后落到缩放点积注意力的完整 PyTorch 实现。对比表、公式推导、直觉理解框和代码实现依次呈现，构成该教学方法的典型样本。

图5：模块阅读页 | 对比表 + 数学推导 + 直觉理解框 + 代码实现的四位一体教学

大模型之路　学习路线　课程阶段　关于　0/32　学员 · 退出

← 返回首页

第一阶段：基础理论

模块 05 — Transformer 架构

进阶 | 预计学习时间：90 分钟

Transformer 是现代大语言模型的“骨架”。LLaMA、Qwen、DeepSeek、ChatGPT 背后都是同一个核心架构：堆叠的 Transformer Block。本模块从直觉到代码，把每个组件讲透——为什么需要注意力、为什么除以 sqrt(d_k)、为什么用 RoPE 而不是正弦编码、为什么 SwiGLU 取代了 ReLU。读完这一节，你能看懂任何一个开源模型的 config.json。

一、为什么 Transformer 取代了 RNN

在 Transformer 之前，处理序列数据的主力是 **RNN（Recurrent Neural Network，循环神经网络）**——一种通过循环连接处理序列数据的网络。每个时间步接收当前输入和上一步的隐藏状态，像人读书一样逐词处理。问题是：读到后面就忘了前面（长距离梯度消失）。为缓解这个问题，人们提出了 **LSTM（Long Short-Term Memory，长短期记忆网络）**——RNN 的改进版，通过门控机制（遗忘门、输入门、输出门）控制信息的保留和遗忘，缓解了梯度消失问题。但仍需逐时间步计算，无法并行。模块 03 讲过这两种结构，这里只做回顾。

维度	RNN/LSTM	Transformer
并行性	❌ 必须逐时间步计算	✅ 所有位置同时计算
长距离依赖	❌ 梯度消失/爆炸	✅ 任意距离直连（注意力）
计算复杂度	$O(n \times d^2)$	$O(n^2 \times d)$ — 长序列更贵
内存	$O(n \times d)$	$O(n^2)$ — 注意力矩阵
最大长度	实际~512	理论上无限（实际128K-1M+）

核心突破： 2017年论文 “Attention is All You Need” 提出的 Transformer 彻底改变了 NLP——用注意力机制替代了循环结构，实现了全并行计算和任意距离的直连。RNN 必须先处理第 1 个词才能处理第 2 个词（串行），而 Transformer 可以同时处理整句话所有词（并行），这让 GPU 的并行算力得以充分发挥，也为大模型 Scaling Laws 奠定了基础。

二、自注意力机制——数学推导

2.1 注意力公式

```
缩放点积注意力：
Attention(Q, K, V) = softmax(Q · K^T / sqrt(d_k)) · V

维度分析：
Q: (seq_len, d_k) — Query 矩阵，“我在找什么”
K: (seq_len, d_k) — Key 矩阵，“我有什么”
V: (seq_len, d_v) — Value 矩阵，“我的内容是什么”

Step 1: scores = Q · K^T   → (seq_len, seq_len) — 每对token的相关性
Step 2: scores = scores / sqrt(d_k)   → 缩放
Step 3: weights = softmax(scores)   → 归一化为概率分布
Step 4: output = weights · V   → (seq_len, d_v) — 加权求和
```

直觉理解： 注意力机制像“图书馆找书”。你有一个问题（Query），每本书有一个标题（Key）和内容（Value）。你拿问题和每本书的标题对比（Q·K^T），越相关的书分数越高。然后用这些分数做加权平均——相关度高的书内容占比大，相关度低的书内容占比小。最终得到一个综合了所有相关书籍信息的答案。

走一遍数字：假设有3个词“我 爱 你”，每个词有4维向量：

- Q = [[1,0,0,0], [0,1,0,0], [0,0,1,0]]（每个词在“找什么”）

数据来源：AI 大模型学习平台，东吴证券研究所

5.2. 渐进式代码构建：从最小实现逐步过渡到完整系统

对于复杂系统而言，如果一开始即提供完整代码，学习者容易停留在“能够运行”的层面，而难以理解其背后的设计逻辑。平台将核心代码实现拆分为多个步骤：先构建最小骨架，再逐步增加输入处理、核心计算、边界条件、性能优化和错误处理。

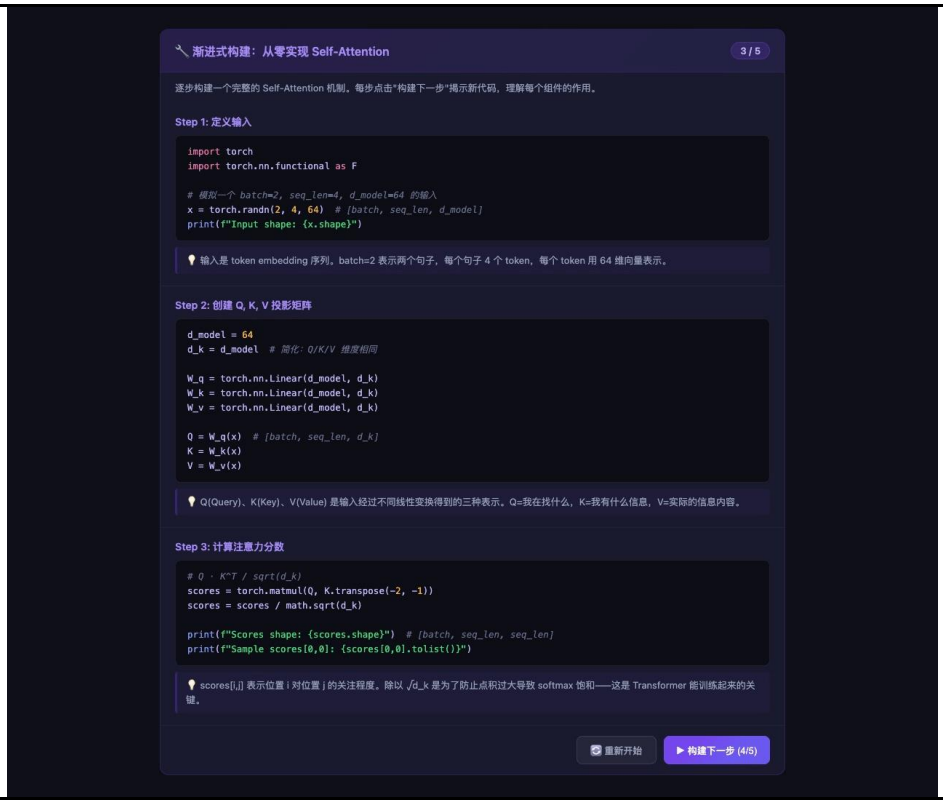
每一步均要求学习者观察预期输出，并理解下一步引入新逻辑的必要性。由此，代码不再只是最终答案，而是成为学习者理解系统构建过程的重要载体。

请务必阅读

更多免费行业报告 并购家 www.ipoipo.net

平台中的代码块也并非简单的复制粘贴示例。以缩放点积注意力函数为例，在语法高亮之外，每一行代码均配有中文注释，用于解释张量形状和计算含义；对于关键步骤，例如 QK 相乘后的缩放处理、softmax 前的掩码操作等，还增加额外的行内说明。较高密度的中文注释主要服务于真实学习场景：工程师阅读的往往是英文代码，但理解和推演过程通常需要中文语境下的解释层，二者之间的衔接对学习效果具有重要意义。

图6：渐进式代码 | 完整实现 + 详尽中文注释



数据来源：AI 大模型学习平台，东吴证券研究所

5.3. 引入调试挑战：从展示正确路径转向训练排错能力

工程能力往往在排错过程中形成。平台在部分实验中有意设置典型错误，例如注意力分数未缩放、梯度未裁剪、KV Cache 更新异常、RAG 切分策略不当导致召回污染等。

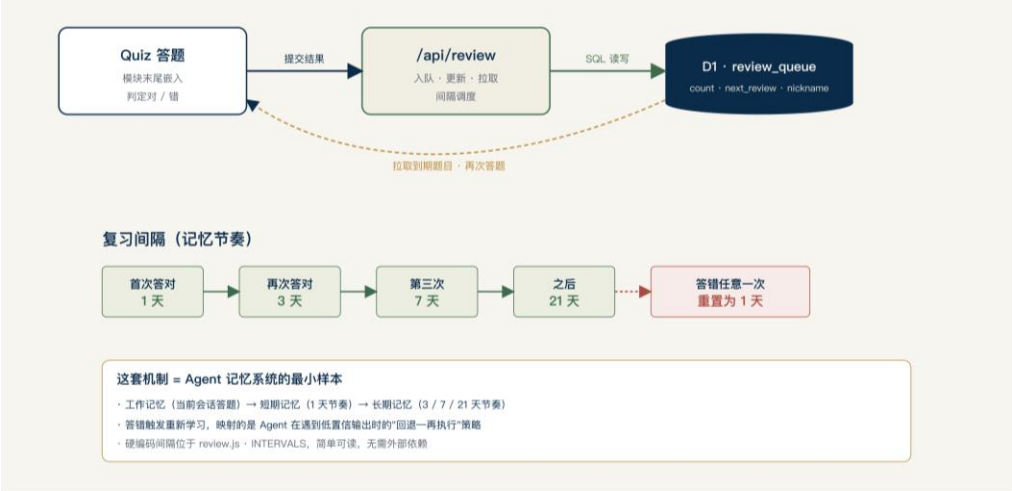
学习者需要完成问题定位、原因解释和代码修复。相比单纯运行标准示例，该过程更接近真实工作中的问题处理方式，也更有助于训练学习者的问题拆解能力、系统边界意识和工程判断力。

5.4. 间隔重复机制：以记忆、状态和触发机制提升长期掌握效果

AI 工程知识密度较高，且存在较强的遗忘风险。平台设计了基于遗忘曲线的复习系统，将关键概念和测验题纳入复习队列，并按照 1 天、3 天、7 天、21 天等节奏触发复习。

从产品功能看，该设计属于学习辅助工具；从系统机制看，其也可以被视为 Agent 记忆系统的简化样本：知识点被记录，学习状态被持续更新，触发策略决定下一次交互时点。学习者在使用该功能的过程中，也能够直观体验 Agent 系统中“记忆—状态—触发”的基本逻辑。

图7：间隔重复系统数据流示意图



数据来源：东吴证券研究所绘制

6. 国产大模型实践：降低从“可调用”到“可落地”的迁移成本

国产模型生态正在从能力追赶逐步进入工程落地深化阶段。对于工程团队而言，选择国产模型并不只是替换 API 地址，更重要的成本来自后续验证环节，包括长上下文场景下的稳定性、Function Calling 的可控性、推理框架支持情况、量化后的质量损失、中文任务适配能力以及规模化调用下的成本曲线等。

平台将国产模型实践嵌入多个关键环节，而非仅在示例代码中替换模型名称，具体包括：

- AI 助教当前以 DeepSeek V4 Flash 作为推理底座，并保留未来按需扩展至其他国产模型 API 的空间；
- 推理部署实验围绕 DeepSeek、GLM 等模型展开，覆盖 vLLM、量化和压测等环节；
- Agent 开发模块关注国产模型 API 在工具调用、多轮对话和错误恢复方面的工程表现；
- 长上下文模块结合 Kimi 等模型能力，讨论其工程利用方式和边界条件；
- 模型评估模块对比不同模型在中文理解、代码生成、复杂推理和成本等维度的表

请务必阅读
更多免费行业报告 并购家 www.ipoipo.net

现；

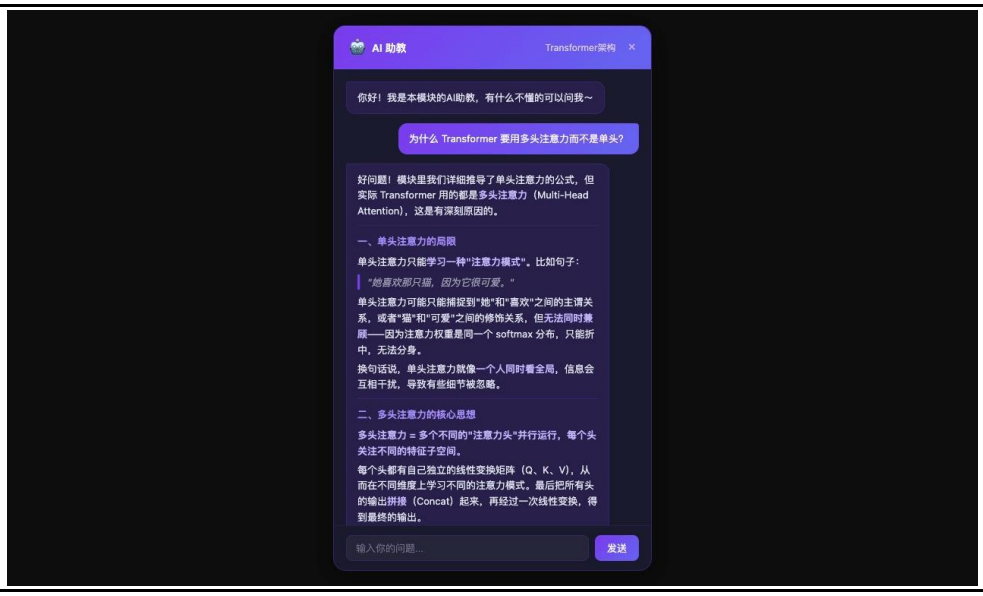
- 量化与优化模块聚焦 W4A8、INT8 等策略在具体模型上的精度—吞吐取舍。

这类内容的价值不在于简单证明某类模型“更优”，而在于帮助工程师降低从模型调用到业务落地之间的迁移成本。生态成熟的重要标志之一，不是每个团队都重复经历相同问题，而是将已有实践中的问题、经验和解决路径沉淀为可复用的工程知识。

我们认为，模型能力决定国产大模型是否具备应用基础，而工程知识体系则影响其能否实现更大范围、较低成本和可持续的产业落地。

在学习者一侧，国产模型实践最直接的产品载体是页面右下角的 AI 助教浮窗。该组件不遮挡主学习内容，每次提问会自动带入当前模块标题和关键概念，并由 DeepSeek V4 Flash 生成解释。回答内容的语气、边界和参考范围均受到当前模块上下文约束，从而降低泛化回答偏离学习主题的概率。

图8：AI 助教 | 浮窗式对话，实时解答与代码调试建议



数据来源：AI 大模型学习平台，东吴证券研究所

从产品形态看，AI 助教表现为页面右下角的悬浮窗；从系统架构看，其背后对应一条较完整的边缘代理链路。该链路包括 chat-widget、/api/chat、TokenHub、deepseek-v4-flash 以及作为 System Prompt 注入的模块元数据，并设置相应的失败降级路径。

图9：AI 助教请求链路示意图

数据来源：东吴证券研究所绘制

7. 平台架构具备 Agent Infra 教学样本属性

平台较具辨识度的设计在于，其并非仅以课程内容呈现 AI 工程知识，而是将学习过程本身组织为一个简化版 Agent 系统。学习者在平台中的一次完整行为链路大致包括：

- 打开某一模块后，系统读取学习状态；
- 学习者向 AI 助教提出问题；
- 边缘函数注入当前模块标题、关键概念和相关上下文；
- 模型生成回答，并以流式方式返回；
- 学习者完成测验后，系统更新知识掌握状态；
- 相关知识点进入复习队列；
- 系统基于间隔重复策略触发后续复习。

从 Agent 系统视角看，上述链路已包含若干核心元素：

- 推理服务：AI 助教调用大模型生成解释；
- 上下文注入：系统根据当前模块动态组织提示信息；
- 状态追踪：知识地图记录学习进度和可访问节点；
- 长期记忆：复习队列保存学习者对知识点的掌握状态；

请务必阅读
更多免费行业报告 并购家 www.ipoipo.net

- 触发策略：间隔重复算法决定后续复习时点，例如 1 / 3 / 7 / 21 天；
- 成本治理：通过边缘函数和按需推理降低固定资源占用。

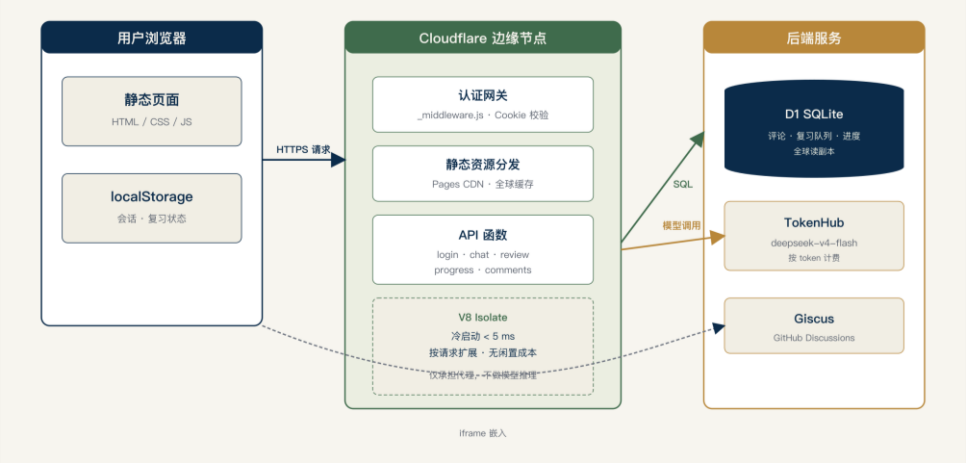
该平台并非完整企业级 Agent 平台，也无需被包装为复杂生产系统。但作为教学样本，其价值在于将推理、上下文、状态、记忆、触发和成本控制等 Agent 系统关键要素，以较低复杂度呈现在同一条产品链路中。学习者并非仅通过架构图理解 Agent Infra，而是在实际使用过程中体验这些模块之间如何协同工作。相较于单纯展示静态架构图，这种“在系统内部学习系统”的方式，更有助于学习者理解 Agent 应用从功能设计到工程实现之间的关系。

本节从 Agent 系统视角对平台功能层进行了拆解。下文将进一步从技术架构角度展开，分析相关能力分别落在哪些层级，以及平台在部署方式、成本控制、扩缩容和访问控制方面作出的工程取舍。

8. 技术架构：边缘优先、低固定成本、弱中心化依赖

平台采用 Cloudflare Pages + Functions + D1 的边缘架构，前端以原生 HTML / CSS / JS 为主，AI 推理通过边缘函数调用模型 API 完成。

图10：系统架构总览



数据来源：东吴证券研究所绘制

该技术选择主要服务于三个目标。

第一，保障轻量化阅读体验。对于包含大量长文、代码示例和交互组件的学习平台而言，并不必然需要复杂 SPA 框架。采用原生静态前端有助于减少 JavaScript Bundle 体积、改善首屏加载体验，并在一定程度上降低前端供应链依赖复杂度。

第二，降低早期固定基础设施成本。在早期访问规模和调用频次相对有限的情况下，

平台可主要利用静态托管、边缘函数和 D1 等能力，降低固定基础设施投入。需要注意的是，这并不意味着规模化之后整体成本可以长期保持接近零。AI 推理仍会随调用量产生费用，未来随着活跃用户数、对话频次和上下文长度提升，仍需通过缓存、模型路由、上下文压缩和调用策略等方式持续进行成本治理。

第三，使架构本身成为教学材料。边缘函数负责请求路由和上下文注入，D1 承担学习状态和复习队列存储，模型 API 负责无状态推理，客户端负责交互与渲染。该分层方式可以清晰呈现许多 Agent 应用中的基础工程问题：状态存放在哪里、上下文在哪里组织、模型如何被调用、异常如何降级、成本如何管理。

当前平台架构可概括为五层：

表2：平台五层技术架构

层级	职责	技术选择
内容分发层	静态内容、代码示例、页面资源分发	Cloudflare Pages / CDN
边缘函数层	请求路由、上下文注入、AI 助教转发	Cloudflare Functions
数据持久层	学习状态、复习队列、评论等数据	Cloudflare D1 / Giscus
AI 推理层	对话辅导、解释生成、模块问答	DeepSeek 等国产模型 API
客户端层	阅读、测验、交互组件、知识地图	原生 HTML/CSS/JS

数据来源：东吴证券研究所整理

该架构的价值不只在于降低成本，更在于将 AI 应用中的若干关键工程问题拆分得相对清晰。对学习者而言，其既是平台运行底座，也是可以被拆解和复用的工程案例。

8.1. 认证与访问控制：以边缘网关完成轻量化权限校验

平台对受保护页面设置登录访问机制。未登录用户访问受保护页面时，将被重定向至登录页。该页面既承担平台首屏访问体验，也作为后续鉴权、Cookie 下发和状态维持的入口。

图11：平台登录页示意图

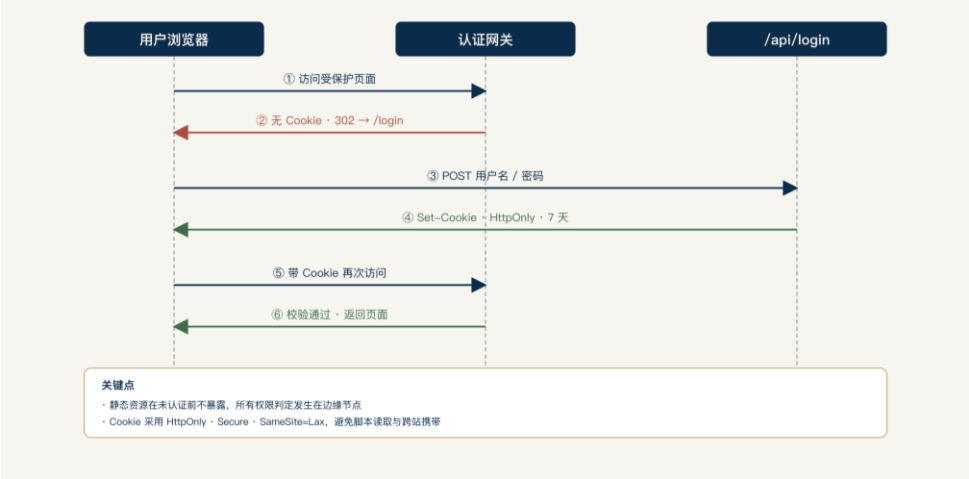
数据来源：AI 大模型学习平台，东吴证券研究所

请务必阅读

更多免费行业报告 并购家 www.ipoipo.net

从工程视角看，认证流程主要包括三部分：白名单放行、Cookie 校验和失败重定向。平台默认将核心页面置于受保护状态，仅对 /login、/api/、/css/、/js/* 等公开路径放行。_middleware.js 在边缘节点完成 Cookie 校验，未认证请求将被重定向至登录页，受保护内容不会直接返回给未授权客户端。

图12：认证流程时序示意图



数据来源：东吴证券研究所绘制

8.2. 数据模型：以 D1 承载学习状态、复习队列和社区互动

D1 中主要承载三张核心表：comments、review_queue 和 user_progress，分别对应社区评论、间隔重复复习队列和学习进度。平台统一使用 Unix 时间戳记录时间字段，并在应用层实现字段限长和速率控制，以降低单条记录膨胀对 D1 读写效率的潜在影响。

图13：D1 数据模型示意图

数据来源：东吴证券研究所绘制

8.3. 计算范式：V8 Isolate 有助于降低轻量 API 的部署与扩缩容复杂度

平台未采用传统 Docker 容器承载 API，而是基于 Cloudflare Workers 的 V8 Isolate 执行边缘函数。相较于传统容器模式，V8 Isolate 通常有助于降低冷启动开销，并简化轻量 API 的扩缩容管理。对于平台当前以 I/O 密集型请求转发为主的业务形态而言，该计算范式具有较好的匹配度。

图14：边缘计算 vs 传统容器 | 冷启动、内存占用、扩缩容与月成本对比

数据来源：东吴证券研究所绘制

此外要注意，V8 Isolate 并不适用于所有计算场景。Cloudflare Workers 在不同套餐下存在 CPU 时间、内存和执行时长等资源限制。平台当前 API 端点主要承担请求代理、上下文注入和结果转发，AI 推理本身在远端模型服务中完成，因此该层资源约束对当前功能影响相对有限。但对于 CPU 密集型任务，例如大规模数据处理、本地模型推理或复杂同步计算，则不适合放在该边缘函数层执行。

9. 竞争定位：差异化切入 AI 工程实践与国产模型适配场景

平台并不试图替代现有主流 AI 学习资源，而是在现有资源体系之外选择差异化切入点。

从当前供给格局看，Coursera、edX 等平台的优势主要体现在课程体系、证书背书和学习约束机制；Hugging Face 的优势在于模型社区、开源工具链和实验环境；技术博客通常具备更新速度快、案例覆盖广、问题颗粒度细等特点；论文与学术课程则更侧重理论深度、方法来源和前沿研究。

“AI 大模型学习之路”选择的是上述资源之间的交叉位置，即**系统化工程实践、国产大模型适配和 Agent Infra 教学样本**。其核心定位并非提供泛化 AI 科普，也不是单纯复刻高校课程或开源社区，而是围绕工程师从模型理解、应用构建到生产部署的能力迁移路径进行组织。

请务必阅读

更多免费行业报告 并购家 www.ipoipo.net

平台的相对优势主要体现在以下几个方面：

- 学习路径较为连续，从数学基础逐步延伸至生产部署；
- 强调从零构建，而非仅停留在框架调用层面；
- 将 Agent 与 Agent Infra 纳入核心训练模块；
- 围绕国产模型沉淀工程适配经验；
- 平台自身可作为低成本、轻量化 Agent Infra 样本被拆解学习。

与此同时，平台也需要正视若干阶段性短板。

第一，完成率可能天然承压。免费、低门槛、浏览器直达等特征有助于提升可及性，但也可能削弱学习者的外部约束。相比之下，Coursera 等平台通过付费机制、作业提交、证书激励和截止日期等方式强化学习压力，而本平台目前尚未形成同等强度的约束机制。

第二，社区生态仍有待建设。AI 工程学习不仅需要内容输入，也需要同伴讨论、代码互评、项目反馈和经验交流。AI 助教可以提升即时答疑效率，但难以完全替代人类社区在复杂问题讨论、代码审查和项目协作中的作用。

第三，内容维护压力较高。模型、推理框架、Agent 工具链和模型 API 均处于高频迭代状态。平台包含 32 个模块、10 个实验、8 个 Capstone 项目和 4 个毕业项目，后续需要持续更新与版本管理，否则过时内容可能削弱学习者信任。

第四，模型 API 与云平台免费额度存在不确定性。当前低固定成本架构在早期阶段具备可行性，但若用户规模、调用频次、模型价格或托管平台政策发生变化，成本治理将成为平台长期运营中必须处理的问题。

10. 发展规划：从学习平台走向工程知识基础设施

平台下一阶段的建设重点不宜仅停留在内容数量扩充，而应围绕**实践深度、学习约束和生态连接**三类能力展开。

近期来看，未来 3–6 个月可重点补齐 Agent 与国产模型工程实验。平台可进一步增加 DeepSeek、GLM、Kimi、混元等模型在不同推理框架和任务场景下的部署对比实验，持续记录吞吐、延迟、显存占用、上下文长度、量化损失和调用成本等指标。同时，应加强 Agent 开发实战模块，覆盖工具调用、多轮规划、错误恢复、状态管理和可观测性等环节。Agent 教育若仅停留在“规划—执行—观察”的概念图层面，难以真正训练生产级工程能力。

中期来看，未来 6–12 个月可建立学习数据驱动的个性化路径。平台可引入学习行为数据分析，包括模块停留时长、测验正确率、复习完成率和项目提交情况等指标，据此识别学习者薄弱环节，并推荐对应复习内容、补充模块或更合适的项目路径。若该机制持续完善，平台有望从静态课程站点进一步演进为自适应学习系统。

长期来看，12 个月以上的建设方向可以聚焦标准化 AI 工程能力评估。当前招聘和培训场景中，“会 RAG”“会 Agent”“熟悉大模型应用”等描述往往较为宽泛，缺乏可验证的能力刻度。平台可围绕模型机制理解、工程实现能力、生产部署能力和 Agent 系统设计能力四类维度建立评估体系。毕业项目、代码评审、压测报告和调试挑战均可成为能力证明的一部分。

11. 风险提示

（1） **内容过时风险：**AI 技术迭代速度较快，尤其是推理框架、Agent 工具链和模型 API。若平台缺乏持续维护机制，深度内容反而可能在一段时间后成为低效甚至误导性的学习材料。

（2） **成本口径风险：**AI 推理、存储、调用频次、模型价格和云平台政策均会影响长期成本结构，后续仍需建立更明确的成本监控和治理机制。

（3） **国产模型生态变化风险：**国产模型的版本、API 格式、定价策略和开源政策均可能持续调整。平台需要避免过度绑定某一模型或某一版本，而应将重点从“适配特定模型版本”升级为“训练学习者适配模型变化的能力”。

请务必阅读
更多免费行业报告 并购家 www.ipoipo.net

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。经授权刊载、转发本报告或者摘要的，应当注明出处为东吴证券研究所，并注明本报告发布人和发布日期，提示使用本报告的风险，且不得对本报告进行有悖原意的引用、删节和修改。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持：预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性：预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持：预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn