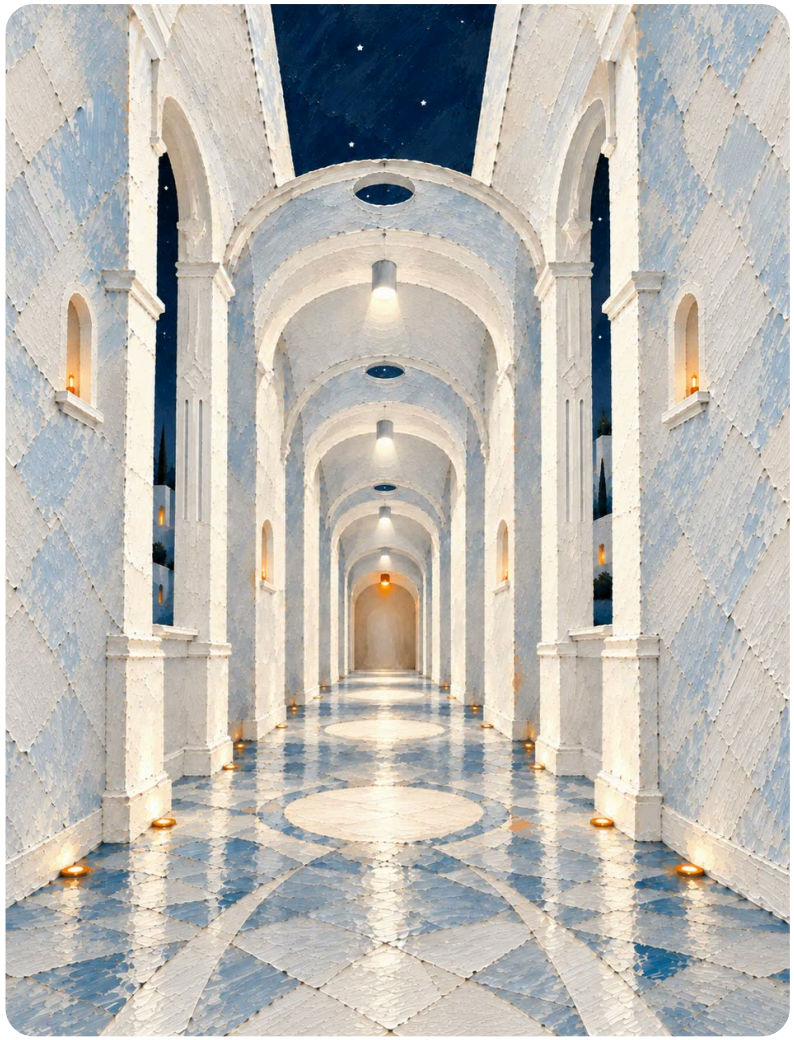
AI网关

AIRouter

智能AI网关解决方案

在账单到来之前掌握AI费用，
在数据发往供应商之前过滤敏感信息。



本手册不罗列功能，
而是从运营AI模型的企业
真正面临的七大难题出发。

一页难题、一页解决方案，
可按顺序依次阅读。



引言

引言	运营AI模型的企业面临的七大难题	p.03
解决方案	AIRouter 是什么，在哪里化解难题	p.04

七大难题与解决方案

难题 01	收到账单才知道AI花了多少钱	p.05
难题 02	每家供应商的密钥 · 控制台 · 合同各自分散	p.07
难题 03	不知道AI费用浪费在哪里	p.08
难题 04	担心个人信息原样发送给AI	p.10
难题 05	出了事故，说不清智能体做了什么	p.12
难题 06	换一个模型就得重新部署	p.14
难题 07	一家供应商停摆，服务也跟着停摆	p.16

应用与导入

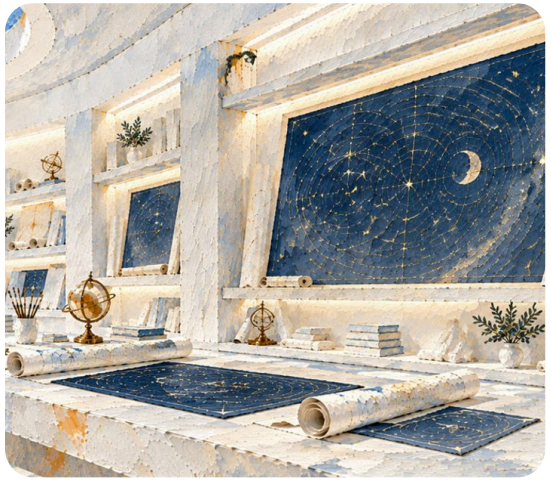
应用示例	导入 AIRouter 后AI运营的变化	p.17
导入步骤	从一项服务起步、逐步推广的导入步骤	p.19
咨询	演示视频 · 咨询导入	p.20

常听到的难题

运营AI模型的企业最先提出的七大难题

一旦将多个AI模型接入服务，
管理部门要求管控，
研发团队要求速度。

我们汇总了双方
反复提到的声音。



点击您最关心的难题，
即可直接跳转查看
AIRouter 的解决方案。

难题 01 “AI花了多少钱，要等账单来了才知道。” 预算管控 →

难题 02 “每家供应商的密钥、控制台、合同都是分开的。” 供应商整合 →

难题 03 “AI费用到底浪费在哪儿，我们根本不知道。” 降低成本 →

难题 04 “就怕客户的个人信息原封不动地发给AI。” 个人信息 →

难题 05 “智能体做了什么，出了事根本说不清。” 智能体管控 →

难题 06 “换一个模型，就得改代码、重新部署。” 模型切换 →

难题 07 “一家供应商出故障，我们的服务也跟着停。” 故障应对 →

统一入口

AIRouter 在统一入口化解这些难题。

智能AI网关解决方案

将分散在各供应商控制台和各服务自研代码中的预算 · 安全 · 故障应对汇集到统一入口，并由AI分析请求记录、给出处置建议。



点击“难题”下方的编号即可跳转到对应难题。

难题 01 · 预算管控

AI花了多少钱，
要等账单来了才知道。

这是财务 · 采购负责人最先提到的问题。



现场常见的情况

用完了才知道金额

AI费用随请求不断累积，
金额却要到下个月的账单上才能确认。

分不清谁用了多少

供应商账单以账号为单位，
各部门 · 项目的用量只能另行核算。

没有办法叫停

即使某个服务突然大量调用，
也无法在超出预算之前拦住。

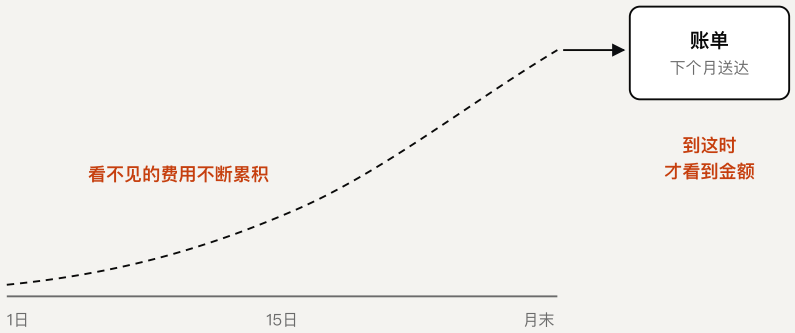
背后的担忧 **“失去对AI预算的控制权”**

查看 AIRouter 的解决方案 →

解决 01 · 不是事后核对，而是事前管控

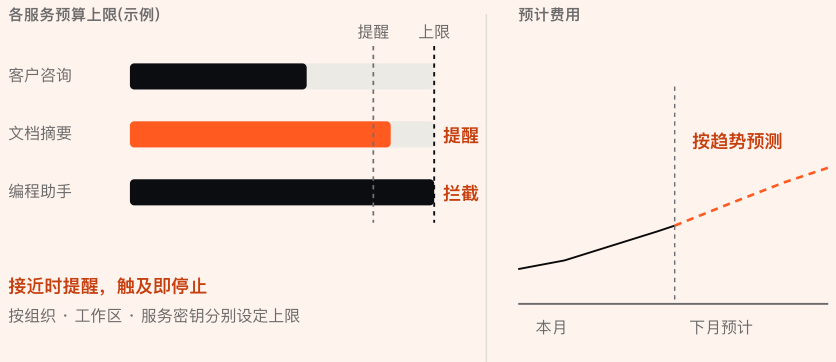
接近设定的上限时提前提醒，触及上限即停止调用。

现状



从使用到收到账单的一个月里，费用始终不可见。

导入 AIRouter 后



使用时即可看到按单价折算的费用，超出上限的调用在发送前即被拒绝。

分级预算上限

按组织 · 工作区 · 服务密钥
分别设定可用金额。

接近时提醒，触及即拦截

接近上限时通知负责人，
触及上限即停止调用。

按团队 · 项目 · 模型查看

多家供应商的用量和按单价折算的费用
集中在一处，可按需分维度查看。

预测下月费用

根据目前的趋势，提前显示
下周 · 下月的预计费用。

您将获得

- ✓ 比账单更早掌握预计费用
- ✓ 在设定的预算范围内运营
- ✓ 依据各部门用量分配预算

难题 02 · 供应商整合

每家供应商的密钥、控制台、合同都是分开的。



每个服务各自对接

每个应用都要逐一对接供应商，
需要管理的地方越来越多。

有几家供应商，就要看几个控制台

用量 · 费用 · 安全设置要在各家控制台分别查看。

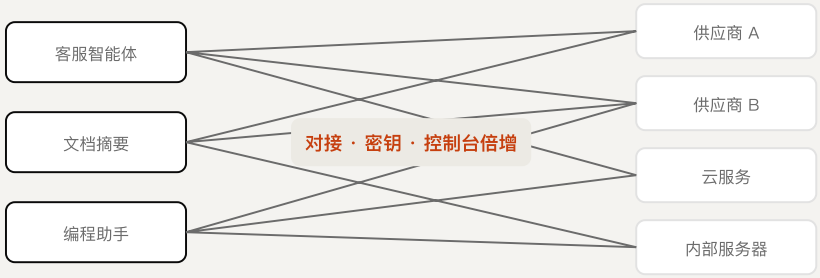
背后的担忧

“供应商越多，管理越难以招架”

解决 02 · 统一入口

一个服务密钥，即可使用多家供应商的模型。

现状



服务和供应商每增加一个，连接就成倍增加。

AIRouter



服务只需连接一处，供应商在入口后方统一管理。

您将获得

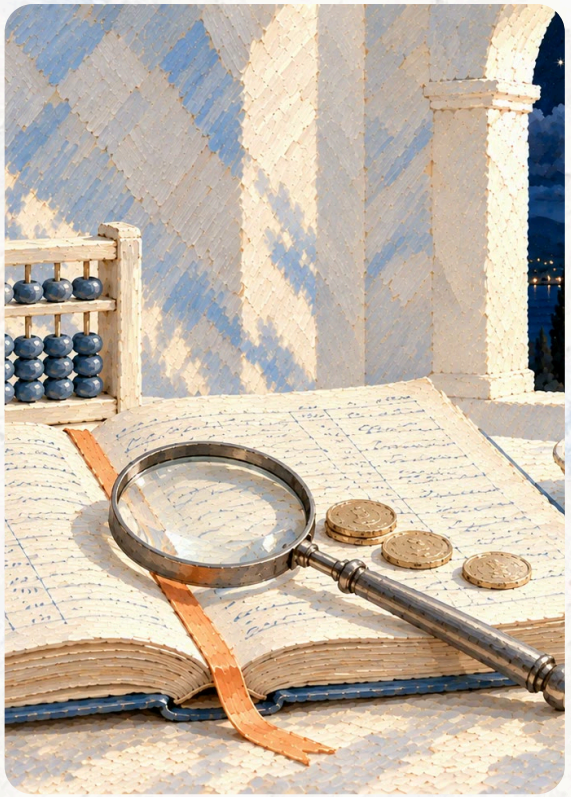
- ✓ 沿用现有供应商合同
- ✓ 各供应商费用一目了然

返回目录

难题 03 · 降低成本

AI费用到底浪费在哪儿，我们根本不知道。

这是运营AI服务的研发 · 财务负责人共同的心声。



现场常见的情况

相同内容每次都重新计算

反复发送的长指令没有使用缓存，
本可享受折扣的费用照单全付。

失败的请求也在花钱

重试 · 失败 · 响应被截断产生的费用
都埋在账单里，根本看不出来。

简单任务也用最贵的模型

简单的摘要和分类
也在使用顶级模型。

背后的担忧 **“明明有可以节省的费用，却找不出来”**

查看 AIRouter 的解决方案 →

解决 03 · 从可节省的费用入手

每5分钟检查一次请求记录，告诉您哪些费用可以省、怎么省。



九条成本规则

找出未使用缓存 · 重试 · 失败 · 响应截断 · 长对话，并计算预计节省金额。

自动提示词缓存

在重复的开头部分设置缓存点，同一对话发往同一供应商。

按条件分配模型

按输入长度 · 是否调用工具 · 是否含图片等您设定的条件，将请求分发到不同模型。

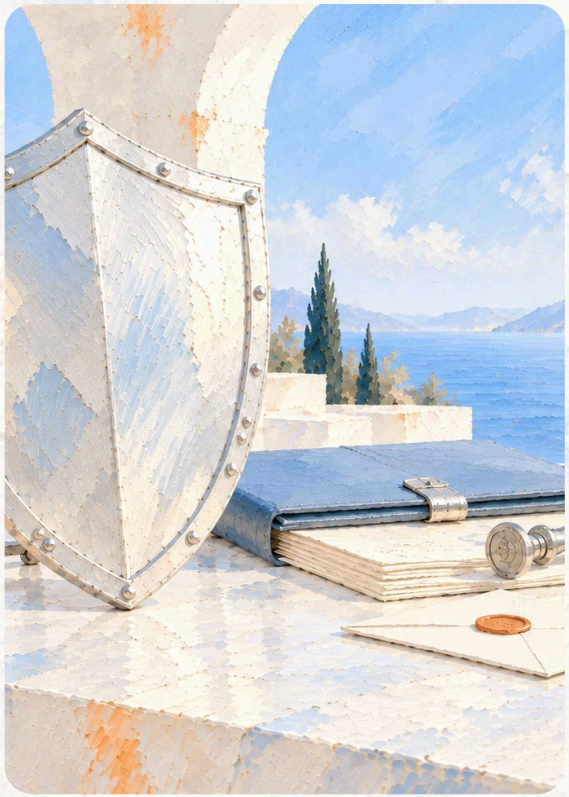
数字靠规则，解释靠AI

金额和次数依据请求记录计算，AI负责梳理原因与处置方法。

您将获得

- ✓ 以预计金额查看可节省的费用
- ✓ 同时获知问题所在与改进方法
- ✓ 处置后还能确认是否已解决

难题 04 · 个人信息 · 机密信息



就怕客户的个人信息原封不动地发给AI。

这是安全负责人在审批AI服务前最先问的问题。

现场常见的情况

中间没有检查环节

各服务直接调用供应商，
没有地方能看到哪些内容被发了出去。

照着文档里的指令行事

智能体读取的文档 · 检索结果中隐藏的指令，
有时会改变智能体的行为。

响应里也会夹带机密

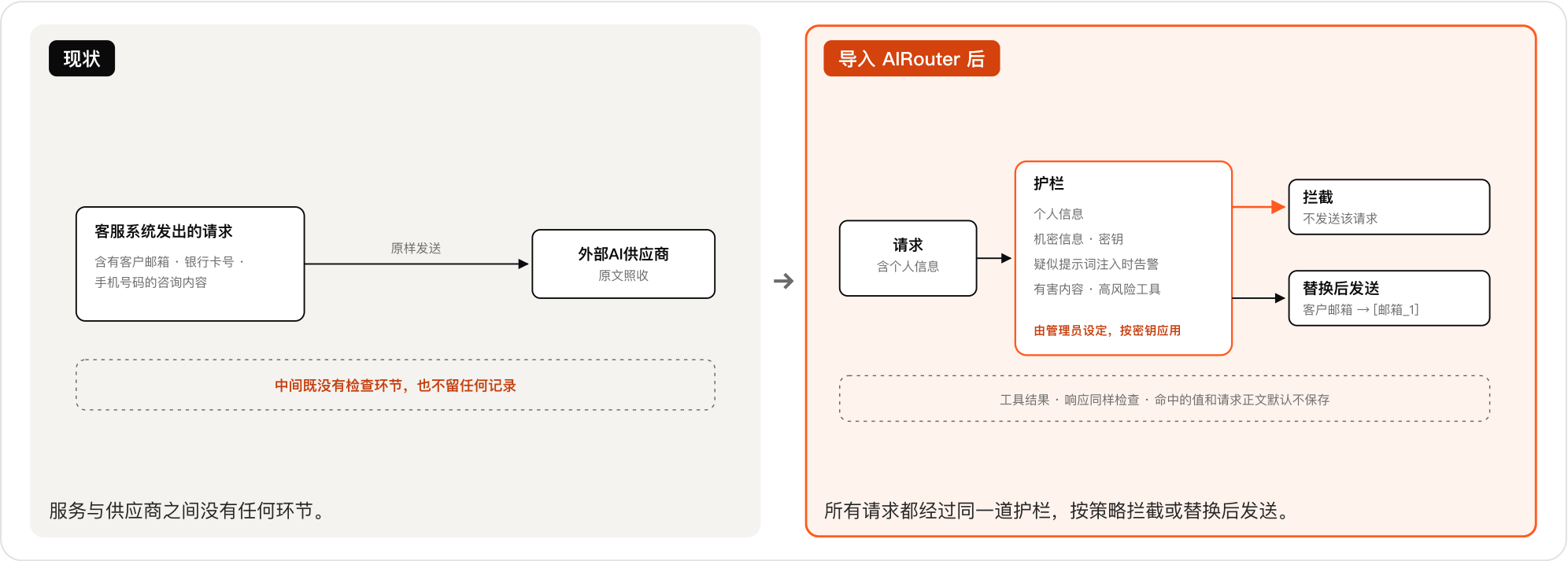
模型的响应中即使混入密钥 · 密码等
机密信息，也无法过滤。

背后的担忧 “一旦发生泄露，责任落到批准引入AI的人身上”

查看 AIRouter 的解决方案 →

解决 04 · 在发往供应商之前

证件号码 · 银行卡号 · 密钥，在发往供应商之前即被过滤。



检查请求 · 工具结果 · 响应

不仅检查发出的请求，
还检查智能体收到的工具结果
和模型的响应。

拦截，或替换后发送

居民登记证 · 外国人登记证 · 护照 · 驾驶证 · 账户号码等，默认替换为类型标签后发送。

开发者无法关闭的策略

护栏由管理员设定并应用到各服务密钥，调用方无法解除。

先“仅记录”，确认后再拦截

将规则设为仅记录，即可先看到若启用拦截将被挡下的请求数。

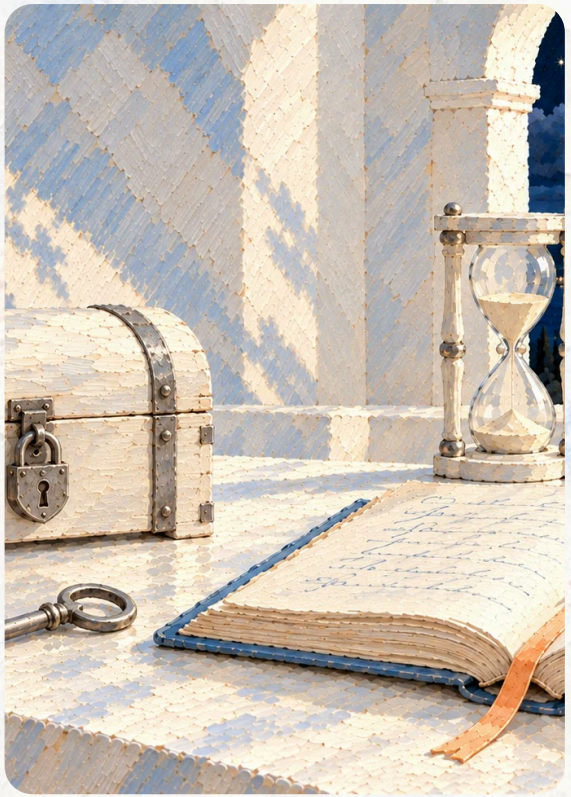
您将获得

- ✓ 敏感信息在外发前被拦下
- ✓ 不中断业务即可测试策略
- ✓ 所有服务适用同一安全标准

难题 05 · 智能体管控 · 审计

智能体到底做了什么，出了事故根本说不清。

这是准备将智能体投入业务的IT · 安全部门提出的问题。



现场常见的情况

一个任务变成几十次调用

智能体的一个任务会引发数十次
模型调用和工具调用，单看一次看不出风险。

高风险工具也照样执行

发送邮件 · 支付 · 删除 · 执行命令等
工具调用，中间没有地方可以确认。

不知道密钥在哪里被使用

来自陌生IP的调用、
长期闲置密钥被重新使用，都无从察觉。

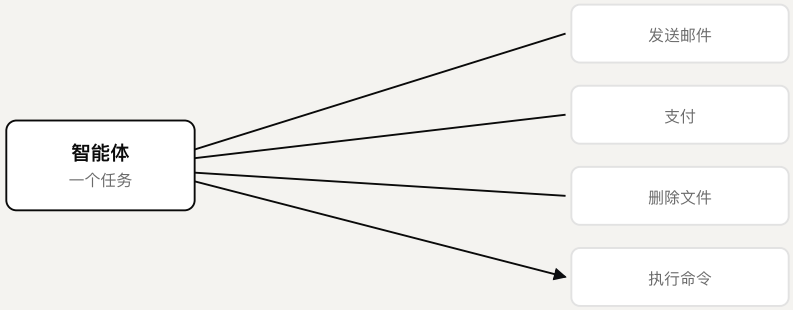
背后的担忧 “事故发生后，没有记录可以追溯经过”

查看 AIRouter 的解决方案 →

解决 05 · 管到智能体的行为

管控高风险工具调用，从记录中提前发现风险征兆并告警。

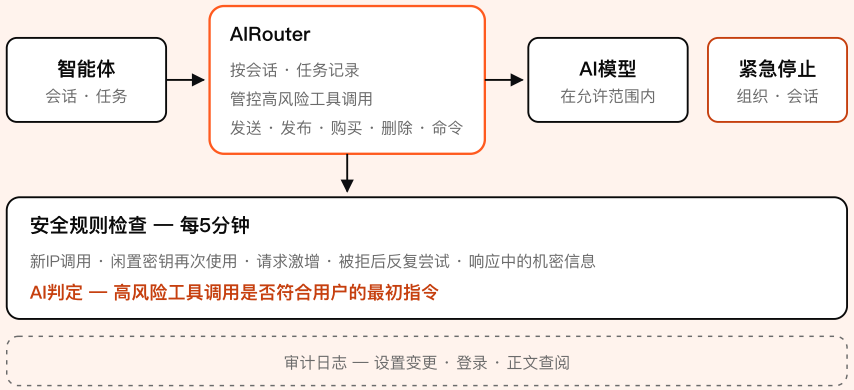
现状



既无确认环节，也没有可追溯经过的记录

智能体与工具之间既无确认，也无记录。

导入 AIRouter 后



所有调用均被记录，风险由规则与AI共同发现并告警。(界面示例)

管控高风险工具调用

通过策略管控发送 · 发布 · 购买 · 删除 · 执行命令等工具调用。

十五条安全规则

每5分钟查找新IP · 闲置密钥 · 请求激增 · 被拒后反复尝试等征兆。

行为是否符合指令

对调用了高风险工具的会话，由AI对照用户的最初指令进行判定。

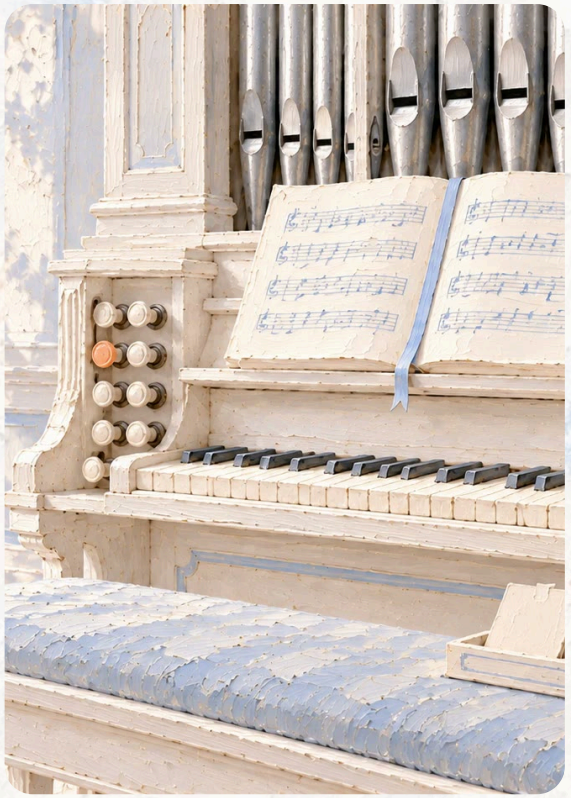
立即停止与审计日志

可按组织 · 工作区 · 会话立即停止，并记录设置变更 · 登录 · 正文查阅。

您将获得

- ✓ 在事故发生前察觉风险征兆
- ✓ 用记录说明事情经过
- ✓ 出现问题可立即停止

难题 06 · 模型切换



换一个模型，就得改代码、重新部署。

这是开发AI服务的团队最常提到的问题。

现场常见的情况

模型名称写死在代码里

每出一个新模型，
就要重复修改 · 验证 · 部署。

回退也需要部署

换上的模型出了问题，
想恢复原样也得重新部署。

各供应商调用方式不同

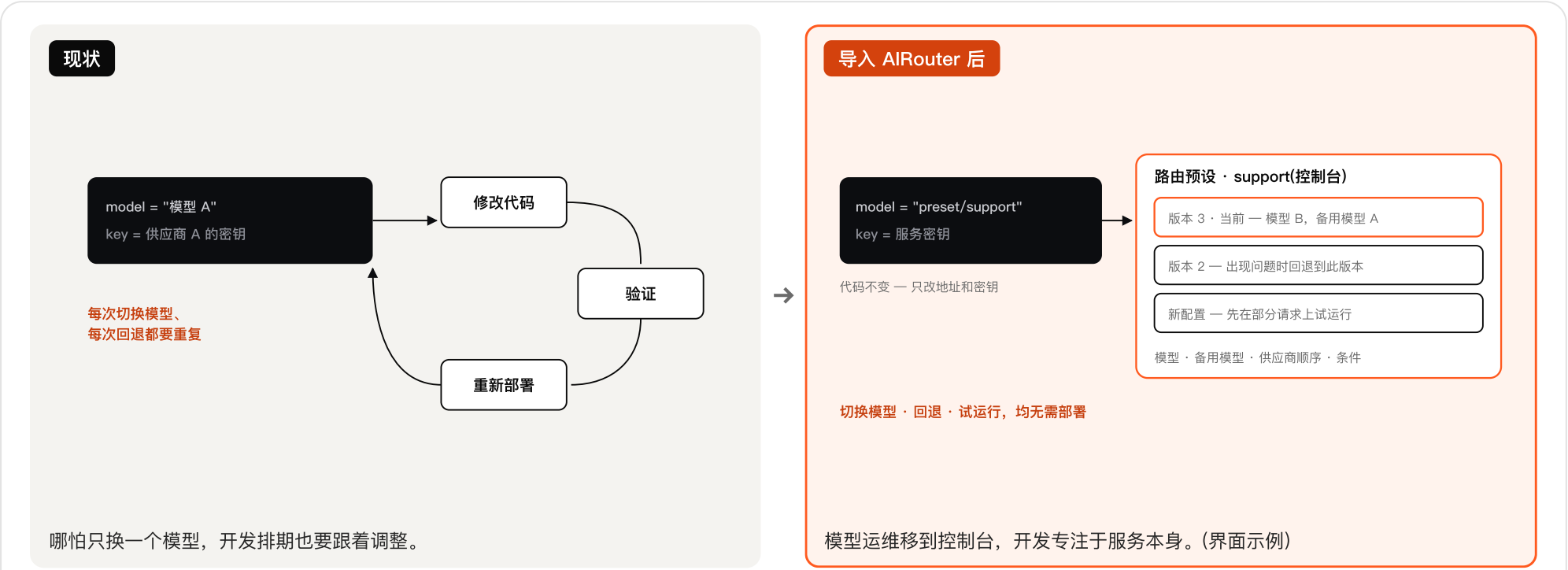
更换供应商时，请求格式和响应处理
都要从头适配。

背后的担忧 “因为怕切换，用不上更好的模型”

查看 AIRouter 的解决方案 →

解决 06 · 代码不变，运维在控制台

代码里只写预设名称，模型切换和回退都在控制台完成。



只改地址和密钥即可接入

兼容两种广泛使用的AI API格式，
现有代码只需更换地址和密钥即可运行。

路由预设

将模型 · 备用模型 · 供应商顺序 · 条件
保存在控制台，代码中只调用名称。

版本管理与回退

按版本管理预设，出现问题时
可立即回退到上一版本。

先用部分请求试运行

新配置只应用于部分请求，
确认费用和响应速度后再扩大范围。

您将获得

- ✓ 切换模型无需部署
- ✓ 出现问题可立即回退
- ✓ 轻松试用更好的模型

难题 07 · 故障应对

一家供应商出了故障，服务也跟着停摆。



超出调用限额，报错直达用户

一旦超出供应商的调用限额，
错误会一直传到用户界面。

每个服务都要自建容灾逻辑

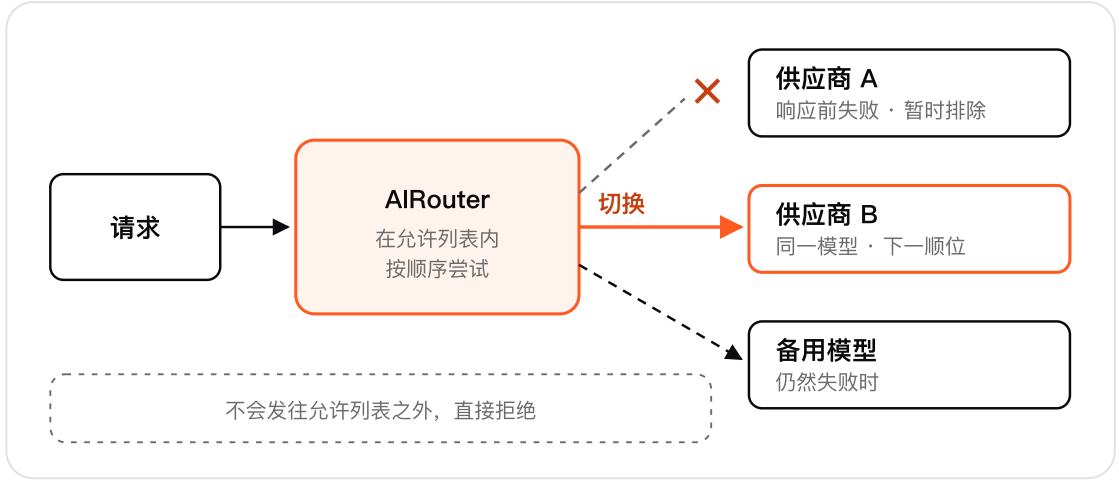
重试和切换到其他供应商，
都要研发团队为每个服务单独开发。

背后的担忧

“AI供应商的故障，变成我们服务的故障”

解决 07 · 在允许范围内切换

供应商出现问题时，切换到已允许的其他供应商。



故障时自动切换

响应前失败时，先切换到
下一家供应商，再到备用模型。

频繁失败时暂时排除

短时间内反复失败的供应商，
暂时不再向其发送请求。

不越出策略范围

不符合允许列表 · 不存储数据 ·
不用于训练等条件的，
不予发送并直接拒绝。

您将获得

- ✓ 无需逐个服务自建容灾
- ✓ 切换也不越出安全策略

返回目录

业务应用示例

导入 AIRouter 后，AI运营将发生这些变化。

智能体开发 · 上线前检查

不必等到上线后才发现费用和安全问题

您将获得 — 在上线前及时修复费用和安全问题。

现在	AIRouter 负责的工作	负责人要做的事
客服智能体会产生多少费用、会调用哪些工具，要等真实客户用过之后才会显现。	追踪开发环境中运行的会话，找出缓存浪费 · 高风险工具调用 · 护栏判定，整理原因与对策。	修复被指出的问题后重新运行，确认问题消失后再上线。

支撑该场景的 AIRouter：解决 03 成本分析；解决 05 工具管控；解决 04 护栏

连接对象：开发工作区；服务密钥；会话 · 任务记录

财务 · 部门预算

部门AI费用不再手工拆分

您将获得 — 依据各部门用量分配预算。

现在	AIRouter 负责的工作	负责人要做的事
从各供应商控制台下载用量，按部门拆分，再用电子表格制作月度报告。	让各工作区在上限内运行；只需在对话中提出要求，即可生成包含表格 · 图表的月度报告。	审阅报告，确认AI助手提出的预算调整建议后执行。

支撑该场景的 AIRouter：解决 01 预算管控；解决 02 统一汇总；解决 03 成本分析

连接对象：工作区；预算上限；AI助手

以上为业务应用示例。适合贵公司服务的范围与效果，将在咨询导入(第20页)时与您共同规划。

业务应用示例(续)

以同样的方式，用于其他业务。

安全 · 受监管行业

策略先测试，再启用拦截

您将获得 — **在不中断业务的前提下**强化安全策略。

现在

担心突然启用拦截规则会影响业务，结果哪条规则都不敢启用。

→

AIRouter 负责的工作

先将规则设为仅记录，并按服务密钥显示若启用拦截将被挡下的请求数。

→

负责人要做的事

确认影响后切换为拦截，并将审计日志用作安全审查材料。

支撑该场景的 AIRouter：解决 04 护栏 · 解决 05 安全规则 · 解决 01 预算管控

连接对象：护栏 · 审计日志 · 告警 Webhook

服务运营 · 模型切换

新模型不会一下子全面上线

您将获得 — **无需部署**即可平稳迁移到更好的模型。

现在

更好的模型出来了，也因全面替换负担太大而一直沿用旧模型。

→

AIRouter 负责的工作

将路由预设的新版本只应用于部分请求，供应商故障时切换到已允许的其他供应商。

→

负责人要做的事

用请求记录比较费用和响应速度后，再决定推广到全部或回退。

支撑该场景的 AIRouter：解决 06 路由预设 · 解决 07 自动切换 · 解决 03 成本分析

连接对象：路由预设 · 请求记录 · 备用模型

以上为业务应用示例。适合贵公司服务的范围与效果，将在咨询导入(第20页)时与您共同规划。

导入步骤

不必一次迁移所有服务。先接入一项服务验证，再逐步推广。

1

共同梳理现状

梳理哪些服务在使用哪些供应商，
密钥和费用分别在哪里管理。

您将获得

分散的AI调用与密钥的现状

2

接入一项服务

在现有代码中只改地址和密钥，
接入一项服务。护栏先以
“仅记录”模式启用。

您将获得

只需改回地址和密钥即可复原，
验证毫无负担

3

启用策略

设定预算上限和路由预设，
已确认影响的护栏规则
率先切换为拦截。

您将获得

在设定的预算 · 安全标准内运营

4

推广到其他服务

按部门 · 项目划分工作区和
服务密钥，在同一策略下
逐步推广。

您将获得

在一处掌握全部AI调用的运营

首次咨询只需准备以下部分资料

正在使用的AI供应商
已签约的AI供应商和
云服务清单

调用AI的服务
智能体 · 应用 ·
内部工具清单

每月AI费用规模
各供应商大致的
使用金额

安全标准
必须禁止外发的信息与
内部安全策略

以 SaaS 轻松起步
获取服务密钥即可直接接入，
现有供应商合同照常使用。

专属环境协商
对安全要求较高的组织，
可协商客户专属云配置。

请求正文默认不保存
仅在组织开启时，
才在设定期限内保存请求正文。

咨询指南

请先告诉我们
贵公司的难题。

只需告诉我们一项调用AI的服务，我们将以该服务为例，
通过演示为您指出费用流失之处 · 外发的敏感信息 ·
故障时的切换路径。资料准备第19页清单中的一部分即可，
我们也会一并介绍导入费用的核算标准。
如想先行了解，欢迎观看演示视频。

咨询导入 →

▶ 观看演示视频



咨询时，请告诉我们您最迫切的一个难题

预算管控 | 供应商整合 | 降低成本 | 个人信息 | 智能体管控 | 模型切换 | 故障应对

邮箱
sales@vicsnc.com

网站
vicsnc.com