

企业大模型调用与智能体应用治理规范（大模型调用篇一）

第一章 总则与治理目标

（一）目的

随着企业智能体（AI Agents）应用的规模化，为彻底解决“接口分散、用量看不见、账单不统一、业务高峰链路中断”等痛点，防范算力成本失控与数据安全风险，特制定本治理规范。

（二）核心原则

坚持“统一接入、集中管控、安全合规、降本增效”的原则。通过引入企业级 AI 策略网关（如 TokenBoss 盒子）作为基础设施，将大模型算力转化为可管控、可审计、可度量的标准化 IT 资产。

第二章 接口与账号治理

（解决“多个大模型接口”与“独立账号”痛点）

（三）统一网关接入

禁止各业务部门、开发团队私自注册、采购第三方大模型 API（如 OpenAI、阿里云、百炼等）。所有大模型调用必须统一收敛至企业的 AI 策略网关，实现底层模型服务的“黑盒化”，保障业务稳定。

（四）协议统一与转换

网关需提供统一的 API 出口（如支持 OpenAI/Claude/Gemini 统一协议自动转换）。业务开发团队只需对接网关标准接口，无需为不同大模型的异构接口重复开发，彻底解决“接口分散”造成的开发壁垒。

（五）模型池化管理

由 IT 架构部负责构建企业级“模型池”，将同质化的大模型服务（如多家供应商提供的同等能力模型）纳入同一个模型池。业务端按模型能力调用，

屏蔽底层供应商差异。

第三章 成本与财务治理

（解决“账单分散”与“成本不可控”痛点）

（六）统一账单与用量透明化

财务部与 IT 部需依托网关的“图表统计”大盘，建立统一的算力账单中心。实时追踪各业务线、各工作组的请求次数与 Token 消耗，杜绝“月底对账痛苦”和“模型盲盒”现象。

（七）精细化额度配额

严禁“无上限”的算力滥用。实施基于工作组的硬性限额策略：

- 1) **部门/工作组配额：** 设定全组每日/每月的 Token 总消耗上限。
- 2) **个人专属配额：** 针对高耗能岗位（如投研分析师、智能体高频使用人员）单独设定限额。
- 3) **并发权重策略：** 针对智能体高并发场景，设定 1-10 级的并发权重，防止单一业务突发流量挤兑整体算力资源。

第四章 研发与调度治理

（解决“分单分散”与“调度困难”痛点）

（八）自动化故障切换与负载均衡

智能体开发必须遵循高可用原则。网关需强制配置“主备模式”，当主用服务商接口异常时（如连接超时、响应报错），网关自动执行故障隔离，并将流量无缝切换至备用服务商，确保业务链路不中断。

（九）智能路由与缓存加速

针对高频重复问题，智能体应优先调用网关的“轻量级智能问答缓存”，

减少上游 API 请求，降低响应延迟与 Token 消耗。同时，依托网关的“定向加速”能力，保障海外模型调用的网络稳定性。

第五章 安全与合规治理（解决“看不见”与“安全风险”）

（十）数据防泄露（DLP）与动态脱敏

所有流经网关的大模型对话必须经过双向安全围栏。

1) **输入检测：** 识别并阻断发送给大模型的身份证号、银行卡号、API 密钥、企业统一信用代码等敏感信息。

2) **输出检测：** 对模型返回内容进行合规性熔断，并开启流式回答脱敏。

（十一）零信任接入与审计

1) **准入控制：** 智能体及内网客户端必须绑定授权的 IP 与设备 MAC 地址，防止非法设备调用。

2) **全量日志审计：** 开启网关的远程日志审计功能（Webhook 异步推送），做到“日常对话提问与回答全文”完全记录。所有调用日志需留存备查，满足金融级合规的“可审计、可追责”要求。

第六章 附则与执行监督

（十二）本规范由企业 IT 治理委员会负责解释与修订。

（十三）任何未通过 AI 策略网关直连外部大模型的“私搭乱建”智能体项目，一经发现，将立即断网并通报批评；造成严重数据泄露或资金浪费的，依法追究责任。

（十四）本规范自发布之日起试行。

落地建议：

这套企业侧“大模型调用与智能体应用治理规范”的“大模型调用”篇一的核心逻辑是“**管住入口、看清流向、守住底线**”。在执行时，建议 IT 部门将 **MLMtoken** 网关的内置的可视化看板转化为内部周报/月报，用数据说话，让业务部门真正意识到“通过网关合规调用，既能保障业务的高可用，又能避免个人和部门背上高昂的成本包袱”。