

企业大模型调用与智能体应用治理规范（试行）

第一章 总则与治理对象

第一条【目的】

随着企业智能体规模化应用，为防范“接口分散、用量不可见、账单不统一、业务高峰链路中断”，以及“智能体越权、任务越界、责任不清、数据泄露、成本失控”等风险，制定本规范。

本规范旨在将大模型算力与智能体任务转化为可管控、可审计、可度量、可追责的标准化 IT 资产与业务能力。

第二条【治理对象与分级】

本规范治理两类对象，适用不同条款：

1. **模型调用级治理**：谁调用、调用哪个模型、消耗多少 Token、是否夹带敏感信息、链路是否可用、账单是否统一。主要适用第二章至第五章。
2. **智能体任务级治理**：智能体代表谁行动、任务边界是什么、权限如何授予与收缩、验收标准是什么、失败如何接管与复盘。主要适用第六章至第九章。

任何智能体上线，应先通过任务级准入，再实施调用级管控。未通过任务级准入的智能体不得进入生产环境；试验类智能体按第五条分级管理。

第三条【核心原则】

坚持“入口集中、权限分级、最小授权、安全合规、降本增效、责任可追”六项原则。

原“集中管控”调整为“入口集中、兼容权限分级”，避免以集中管控替代分级放权。

第四条【治理架构】

设立联合治理机制，由 IT 治理委员会、业务责任人、法务/合规、数据安全四

方组成。

每个上线智能体必须设“智能体责任人”（业务 Owner），负责任务边界、验收标准、扩权申请与事故复盘。

本规范由联合治理机制解释与修订。

第二章 接口与账号治理

第五条【统一网关接入与分级】

1. 生产类智能体的大模型调用必须统一接入企业级 AI 策略网关（以下简称网关），实现底层模型服务黑盒化。禁止生产类项目私自注册、采购第三方大模型 API 或直连外部模型。
2. 试验类智能体（沙箱、脱敏数据、限额、登记用途与退出期限）允许临时直连，但须登记用途、责任人、数据范围、限额与退出期限；到期未转正即断连。不得以“试验”名义长期规避网关。
3. 网关应提供统一 API 出口，支持主流协议自动转换。业务团队只需对接网关标准接口，减少异构接口重复开发。

第六条【协议统一与工具调用审计】

网关除统一模型调用协议外，还须登记并审计“工具调用”：谁调用了哪个内部系统、传递了什么参数、返回了什么结果。

跨智能体委托时，须满足第七章身份遗传与授权递减要求。

工具调用是权限与提示注入风险的重点，不得只审计模型对话而遗漏工具入参与出参。

第七条【模型池与任务模型下限】

IT 架构部构建企业级模型池，将同质化模型纳入同一池，业务端按能力调用。每个任务须登记“可用模型等级”与“降级下限”；智能路由不得低于该下限。因故障切换、降级或成本优化更换模型时，须记录并触发回归测试，防止“模

型能力相近”导致任务结果不可接受。

第三章 成本与财务治理

第八条【统一账单与用量产出透明】

财务部与 IT 部依托网关可视化看板建立统一算力账单中心，实时追踪各业务线、工作组、智能体、任务的请求次数与 Token 消耗。

看板至少增加两类产出指标：

1. 每个成功任务成本；
2. 单位 Token 产出的有效任务数。

归集维度不限于部门，还应按任务/智能体维度归集，避免“Token 便宜即随便跑”的误判。

第九条【任务配额与精细化额度】

严禁无上限算力滥用。配额对象由“部门/个人 Token 上限”调整为“任务配额”：

1. 任务配额：每日可完成任务数、单任务 Token 上限；
2. 部门/工作组配额：全组每日/每月总量上限；
3. 个人专属配额：高耗能岗位单独设定；
4. 并发权重：高并发场景设 1—10 级权重，防止单一业务挤兑整体资源；
5. 临时提额：允许业务高峰申请临时提额，事后复核；禁止月初一次性锁死额度导致月末误伤。

第十条【成本口径】

成本治理不得等同于 Token 配额。企业智能体成本大头通常包括系统集成与改造、权限治理、评测集建设、失败案例排查与回归、知识库维护人力等。

预算应按“总算力成本+集成摊销+评测与运维成本”归集，按成功任务成本考核，避免预算按错误单位分配。

第四章 研发与调度治理

第十一条【故障切换与负载均衡】

智能体开发应遵循高可用原则。网关应配置主备模式：主用服务商接口异常时，自动故障隔离并切换备用服务商。

但不得承诺“无缝切换、业务链路绝对不中断”。不同供应商在工具调用格式、输出风格、拒答策略上存在差异，链路切换成功不等于任务成功。

切换须记录，并对关键任务触发结果校验与回归；无法通过校验的，应降级为人工接管或只读建议。

第十二条【缓存与加速治理】

高频重复问题可优先调用网关轻量级智能问答缓存，但须满足：

1. 缓存有效期（TTL）按任务设定；
2. 命中缓存前校验调用者权限与数据隔离，防止跨用户串数据；
3. 含敏感信息的内容不得入缓存；
4. 命中日志单独统计，命中率与“事后被纠正率”并列展示；
5. 过期答案不得继续作为唯一依据，关键任务须回源或显示知识时效。

海外模型“定向加速”仅解决网络稳定性；涉及数据出境与个人信息处理的，须同时满足第二十九条合规要求。

第五章 安全与合规治理

第十三条【数据防泄露与提示注入检测】

所有流经网关的大模型交互须经过双向安全围栏。检测对象包括五处：提示词、检索片段、工具入参、工具出参、模型输出。

1. 输入侧识别并阻断身份证号、银行卡号、API 密钥、统一社会信用代码等敏感信息，并增加提示注入检测：识别用户输入或检索内容中夹带的越权指令、绕过指令、工具滥用指令。

2. 输出侧进行合规性熔断与流式回答脱敏，防止输出内部制度、客户名单等敏感内容。
3. 命中安全策略须触发“阻断—降级—记录”三段动作，并纳入审计与复盘。

第十四条【准入控制与身份授权分离】

网络准入不等于身份授权。智能体及内网客户端绑定授权 IP 与设备 MAC 地址，仅作为网络准入措施；主体授权按第七章执行。

智能体不得使用共享网关令牌以管理员名义修改配置。

审计日志须能回答“谁授权它这么做”，而不仅是“某个 IP 用了某个令牌”。

第十五条【日志审计与留存】

网关开启远程日志审计与异步推送。“全量留存”对象为结构化元数据：谁、何时、哪个智能体、哪个任务、调用了什么、花了多少、结果如何、是否人工接管、是否命中安全策略。

对话正文按需留证，不全量常备；设留存期限与分级访问；日志须签名、异地留存，防篡改、防抵赖。

涉及个人信息的最小必要处理，由法务与数据合规部门制定专项细则。

第六章 智能体与任务治理

第十六条【任务准入】

任何智能体上线前须提交一页纸任务说明：输入、输出、验收标准、错误代价、数据范围、责任人。

没有验收标准的任务不得上线。上线任务书完备率为硬门槛。

第十七条【任务三态与判据筛选】

每个流程节点显式标注为“人工/AI/人机协作”，并写明交接点判定条件。

任务能否交给智能体，按三项判据筛选：结果能否被机器判定、错误代价是否

可控、是否有足够历史样本。不满足者只能做辅助建议，不给执行权。

第十八条【只读起步与逐级扩权】

新上线智能体默认只读，先输出建议、人工执行；积累真实失败样本后再逐级扩权。

每次扩权须配套增加审计与回滚措施。写权限智能体只读起步比例应作为考核指标。

第十九条【验收指标与人在环路】

上线门槛与考核使用“任务完成率、人工接管率、失败后纠正耗时”，不得使用“回答好不好”等不可验收表述。

关键节点保留人工确认；无人化流程必须定义失败兜底与接管条件。

第二十条【事故分级与复盘】

智能体错误按影响分级：可自愈、需人工介入、造成业务或资金损失。每级对应处置动作、上报路径与复盘要求。

事故复盘须回填回归集，防止同类错误重复发生。

第七章 身份、委托与授权

第二十一条【独立身份与三身份区分】

每个智能体有独立身份，不与个人共享令牌。

一次调用须区分三种身份：发起人、智能体运行时身份、被代表的业务主体。

身份、代表谁、有效期在调用链上必须可追溯。

第二十二条【委托链递减】

A 委托 B、B 委托 C 时，C 获得的权限不得大于 B 获得的权限，越往下越窄；必要时重新取授权。

跨智能体通信须携带可验证身份与委托链；缺失身份遗传的调用应拒绝。

第二十三条【工具动作授权与撤销】

工具调用按动作粒度授权，读写分离；写操作默认需要显式确认。

授权可撤销、可限时；撤销须即时生效，并能回滚已产生的副作用。

高风险动作须记录授权人、授权时间、有效期与撤销记录。

第八章 评测、变更与回归

第二十四条【回归集】

每个上线智能体须有回归集（历史真实案例+期望结果），随提示词、知识库、模型、工具变更自动运行。

回归覆盖率作为考核指标。

第二十五条【灰度与放量】

新版本先在影子模式或小流量运行，达标后放量；未达标自动回退。

第二十六条【变更管理】

换模型、换供应商、改提示词、改知识库、改编排，均属于变更，均须走回归。

故障切换、模型降级、缓存策略调整，同样触发关键任务回归。

第二十七条【知识库版本与时效】

每条知识有版本与有效期；过期知识自动降权或下架。

引用时须显示依据与时间。缓存答案与知识库版本须可追溯。

第九章 供应商、数据与退出

第二十八条【数据不用于训练】

合同须写明企业数据、提示词、输出不被用于供应商模型训练或跨客户优化，并给出可核验技术措施或审计权。

第二十九条【数据出境与地域合规】

调用境外模型涉及个人信息或重要数据出境的，须先通过合规清单。

海外加速不得绕过数据出境合规评估。

第三十条【SLA 与降级条款】

合同应约定可用性、故障通知时限、切换义务、数据可导出格式、退出协助与迁移期。

第三十一条【避免隐式锁定】

日志、评测集、提示词、知识索引必须是企业自有、可迁移资产。

供应商退出时须提供迁移协助，不得以格式或接口锁定阻碍退出。

第十章 附则、执行监督与考核

第三十二条【联合治理与责任人】

本规范由联合治理机制解释与修订，设智能体责任人制度。

纯 IT 治理不得裁定“任务做错算谁的”，须由业务责任人、法务合规、数据安全共同参与。

第三十三条【处罚分级与紧急止血】

违规处置分两级：轻微违规先“撤销授权/令牌降级为只读/限流”并限期整改；确认外泄或资金浪费再断网追责。

断网正在跑生产任务的智能体，可能比违规本身损失更大。故须规定紧急止血授权链：谁有权在几分钟内拉闸、降级、限流，并记录事后复核。

第三十四条【试行与复审】

本规范自发布之日起试行 6 个月，先在 2—3 个低风险业务线试行；期满复审。

规范每半年评估一次是否与模型能力匹配，并设退出机制与复审周期。

本规范由联合治理机制负责解释与修订。

任何未通过网关直连外部大模型的“私搭乱建”智能体项目，一经发现，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。

第三十五条【考核硬指标】

建议纳入考核的硬指标包括，但不限于：

1. 任务完成率；

2. 人工接管率；
3. 失败后纠正耗时；
4. 上线任务书完备率；
5. 回归覆盖率；
6. 写权限智能体只读起步比例；
7. 每个成功任务成本；
8. 单位 Token 产出的有效任务数。

阈值由治理委员会按业务线设定并公示。

第三十六条【实施提示】

本规范的核心是“管住入口、看清流向、守住底线”。

IT 部门应将网关可视化看板转化为内部周报/月报，用数据说明：通过网关合规调用，既能保障业务高可用，也能避免个人和部门承担高昂成本。