深翼 DeepWing

DeepWing 框架正式版

D W - A I - A R C H

企业 AI 架构框架规范

七层与十关：定义、职责、失效模式、检查项与合格评定

版本　正式版 v1.0（首版）

日期　2026-10-27

署名　深翼 DeepWing

官网　https://deepwingai.com/framework/v1.0

版权与引用方式

© 2026 深翼 DeepWing。版权所有。

本文件为《DeepWing 框架正式版》（DW-AI-ARCH）正式版 v1.0，发布于 2026-10-27。本文件的唯一权威现行文本为深翼官网框架页：https://deepwingai.com/framework/v1.0；后续修订只在该页面与版本记录（附录 C）中发布。

引用方式。引用本文件条文或定义时，须注明出处、版本号与官网链接。引用格式：

> 深翼 DeepWing. DW-AI-ARCH：企业 AI 七层与十关框架，正式版 v1.0. 2026-10-27. https://deepwingai.com/framework/v1.0

编制说明。

- 定版依据：DW-AI-ARCH v1.0 定版（2026-10-23），层名与关名与定版逐字一致。
- 配套模型：深翼 AI 治理对账模型（DW-AI-GOV v1.0），本文件第 12 章、第 14 章的治理衔接口径与之对齐。

目录

前言 6
第 1 章 范围 7
第 2 章 引用文件 8
第 3 章 术语与定义 9
第 4 章 总则 10
第 5 章 模型层规范 11
5.1 范围与定义 11
5.2 职责边界 11
5.3 组成要素 11
5.4 常见失效模式 12
5.5 规范条目 12
5.6 检查项 13
5.7 与其他层及关口的接口 14
5.8 关联判例 14
第 6 章 编排层规范 15
6.1 范围与定义 15
6.2 职责边界 15
6.3 组成要素 16
6.4 常见失效模式 16
6.5 规范条目 17
6.6 检查项 17
6.7 与其他层及关口的接口 18
6.8 关联判例 19
第 7 章 工具层规范 20
7.1 范围与定义 20
7.2 职责边界 20
7.3 组成要素 20
7.4 常见失效模式 21
7.5 规范条目 21
7.6 检查项 22
7.7 与其他层及关口的接口 23
7.8 关联判例 23
第 8 章 记忆层规范 25
8.1 范围与定义 25
8.2 职责边界 25
8.3 组成要素 25

8.4 常见失效模式 26
8.5 规范条目 26
8.6 检查项 27
8.7 与其他层及关口的接口 28
8.8 关联判例 28
第 9 章 知识层规范 29
9.1 范围与定义 29
9.2 职责边界 29
9.3 组成要素 30
9.4 常见失效模式 30
9.5 规范条目 31
9.6 检查项 31
9.7 与其他层及关口的接口 32
9.8 关联判例 32
第 10 章 门禁层规范 34
10.1 范围与定义 34
10.2 职责边界 34
10.3 组成要素 35
10.4 常见失效模式 35
10.5 规范条目 36
10.6 检查项 36
10.7 与其他层及关口的接口 37
10.8 关联判例 38
第 11 章 盯梢层规范 40
11.1 范围与定义 40
11.2 职责边界 40
11.3 组成要素 40
11.4 常见失效模式 41
11.5 规范条目 41
11.6 检查项 42
11.7 与其他层及关口的接口 43
11.8 关联判例 43
第 12 章 十关检查规范 45
12.1 第 1 关 网关：这把钥匙还有效吗（跨层） 45
12.2 第 2 关 门禁：这把钥匙能开哪些门（对应门禁层） 45
12.3 第 3 关 编排：这句话该找谁办（对应编排层） 45
12.4 第 4 关 知识：在几万份文件里找最新规矩（对应知识层） 46
12.5 第 5 关 记忆：模型记不记得这位顾客（对应记忆层） 46
12.6 第 6 关 工具：真正跨入真实系统动手办事（对应工具层） 46

12.7 第 7 关 模型：把一切线索拼成一句人话（对应模型层） 46
12.8 第 8 关 盯梢：答完之后，谁在盯着质量（对应盯梢层） 46
12.9 第 9 关 平台：八道关，是各置办一遍还是共用一套（跨层） 47
12.10 第 10 关 统一与自主：公共的路，团队愿不愿意自己走（跨层） 47
第 13 章 合格评定与检查方法 48
13.1 检查的两种形态 48
13.2 计分规则 48
13.3 红线规则 48
13.4 证据强度 48
13.5 结论与复查 49
第 14 章 记录与证据规范 50
14.1 通用字段 50
14.2 证据等级 50
14.3 披露与日期口径 50
14.4 记录状态与编号规则 51
14.5 句子的三种属性 51
附录 A 判例索引（资料性） 52
模型层 52
编排层 52
工具层 52
记忆层 52
知识层 52
门禁层 53
盯梢层 53
附录 B 术语表（资料性） 54
附录 C 版本与修订记录（规范性） 55
附录 D 十关与 NIST AI RMF 对照（资料性） 56
D.1 编制说明与一手来源 56
D.2 主映射表 56
D.3 缺口清单：RMF 有、十关没覆盖的部分 58
D.4 生成式 AI 风险附表（NIST AI 600-1） 59
D.5 定位说明 60
文末一 · 缺口清单（正式版 v1.0 未决事项） 61
文末二 · 体例说明（全稿沿用） 62
一、编号规则 62
二、术语体例 62
三、条文体例 62
四、版本记录格式 63

前言

本文件是深翼 DeepWing 企业 AI 架构框架（DW-AI-ARCH）的正式版。它承接 DW-AI-ARCH v1.0 定版：层名、关名、关主题、计分与红线口径、记录格式与证据等级，全部以定版为准，本文件只做一件事，把定版的结构展开成可执行、可核查的条文。

本文件的总立场只有一句：讲职责不讲部署。七层是职责的划分，不是物理部署的划分；一个组件可以兼几层，一套部署可以拆几层。凡判例与框架对不上，先改框架，并记入附录 C 的修订记录，不硬套判例迁就框架。

本文件与深翼 AI 治理对账模型（DW-AI-GOV）成对使用：本文件回答系统怎么建，治理对账模型回答规则与现实对不对得上。治理要求落到系统时落在哪一关，以本文件第 12 章的关口为准；判例回指规则、读数指回编号的纪律，以本文件第 14 章为准，与 DW-AI-GOV 的数据绑定一致。

第 1 章 范围

1.1 目的：把企业 AI 系统的职责拆分与检查顺序写成可执行、可核查的条文，使任何一套已进入或准备进入生产环境的 AI 系统，都能按同一套问题被检查、被判定、被记录。

1.2 适用范围：使用大模型构建的 AI 应用与智能体系统，重点是已进入或准备进入生产环境、会对真实业务与真实用户产生后果的系统。纯离线实验、受控评测环境中的系统，可参照本文件做风险映射，但其判例归属按第 14 章口径另行标注（参见附录 A 中 SP-001 的标注方式）。

1.3 适用对象：架构师、技术负责人、采购与合规人员，以及负责上线评审的人。也供使用本框架的检查工具（快速自检、深度诊断）的操作者对照条文出处。

1.4 边界：本文件是深翼的分析框架，不是行业标准，不规定具体产品、供应商与部署形态。条文中出现的个案与数字均有在册编号可指回，不升为行业通则；定量阈值本版只写判定方法，不写死数字（见缺口清单）。

第 2 章 引用文件

2.1 框架文件： （1）DW-AI-ARCH v1.0 定版（2026-10-23），本文件的口径源头； （2）深翼 AI 治理对账模型 v1.0 定版（DW-AI-GOV，2026-10-26），第 12 章、第 14 章治理衔接口径的对齐依据；（3）十关与 NIST AI RMF 对照，见附录 D，资料性引用，不产生条文义务。

2.2 登记册：本文件条文与检查结论所指回的在册登记册，包括：事故判例（INC）、特别卷（SP）、管理措施（GOV）、监管执法（REG）、外部 Agent 准入（AGT）、模型变更（MOD）、成本（COST）、架构卡（ARC）、新思想（IDEA）、产品拆解（PD）、前沿公司（CO）、实验室（LAB）、人物（PPL）、周刊（WK），以及二号台的职业暴露（OCC）、就业研究（RES）、招聘信号（HIR）与三号台的医疗器械（MED）。各册的编号规则与状态口径见第 14 章。

2.3 外部文件：本版条文未直接设定外部文件的合规义务；凡素材中提及的外部标准与官方文件，仅在对应条文中以资料性引用出现，逐件版本与日期的核对列入缺口清单，核对完成前不作为判定依据。

第 3 章 术语与定义

3.1 通则：术语首次出现时写「中文正名（英文 English Name）」，其后只用中文正名。检查项编号以层全名作前缀（见 3.2），正文叙述一律用中文全名，不用代码指代层名、关名。

3.2 层与关的正名：

七层（职责，不是物理部署）：

层	英文	对应关
模型	Model	第 7 关 模型
编排	Orchestration	第 3 关 编排
工具	Tool Use	第 6 关 工具
记忆	Memory	第 5 关 记忆
知识	Knowledge	第 4 关 知识
门禁	Access Control	第 2 关 门禁
盯梢	Observability	第 8 关 盯梢

注：检查项编号以层全名作前缀，如「模型-C1」「门禁-C1」，格式见文末体例说明。两字母代码自本版起停用，不再用于编号与指代。

十关（一条请求的检查顺序）：1 网关（Gateway，跨层）、2 门禁（Access Control）、3 编排（Orchestration）、4 知识（Knowledge）、5 记忆（Memory）、6 工具（Tool Use）、7 模型（Model）、8 盯梢（Observability）、9 平台（Platform，跨层）、10 统一与自主（Alignment & Autonomy，跨层）。七关对应七层；网关、平台、统一与自主三关跨层。

3.3 通用术语： （1）检查项：按关或按层设立的、可逐项判定合格与否的检查条目，编号格式见文末体例说明； （2）主死因层：判例归因时认定的第一失效层；次层关联指与事件相关但非第一失效的层； （3）证据等级：记录可信度的四级标定，定义与判定见第 14 章 14.2； （4）最后核查：一条记录最近一次回到一手来源复核的日期； （5）红线：触到即结论不得为「可以上线」的情形，见第 13 章 13.3； （6）状态格、落差格、治理缺口：治理对账模型（DW-AI-GOV）的三个读法用语，定义以 DW-AI-GOV v1.0 为准，本文件引用时不改义。

3.4 废止名：以下旧称自 v1.0 定版起停用，正文不得使用，引用历史材料时出现者须加注现行正名：Gating（现行正名：门禁）、Oversight（现行正名：盯梢）、Knowledge Base（现行正名：知识）、Model Layer（现行正名：模型）、AI Gateway（现行正名：网关）、Enterprise AI Platform（现行正名：平台）。

第 4 章 总则

4.1 框架构成：本框架由两套结构组成。七层解决「职责怎么分」：把一个企业 AI 系统的职责拆成七个互不越界的职责层，出了问题先定位病根在哪一层。十关解决「检查按什么顺序」：把一条请求从进门到答完、再到治理收口的完整路径排成十个检查关口，评审时逐关过，不漏项。

4.2 七层职责总览：（1）模型层：理解与生成。管模型选型、路由、版本与成本归属，不管执行与权限；（2）编排层：决定下一步。管意图拆解、步骤编排与多智能体协调，不管具体执行；（3）工具层：真正执行。管工具接入、参数契约与调用可靠性，不管该不该办；（4）记忆层：记住上下文。管短期与长期记忆的存取与边界，不管企业知识供给；（5）知识层：补上私有知识。管检索、向量存储与知识时效，不管对话上下文；（6）门禁层：事前权限判断。管谁能做什么、能看什么，不管事后追溯；（7）盯梢层：事后看清。管追踪、评估与质量监测，不管事前拦截。

4.3 十关检查顺序：一条请求依次过十关。第 1 关网关先验钥匙是否有效、人从哪进门；第 2 至第 8 关依次是门禁、编排、知识、记忆、工具、模型、盯梢，对应七层的实际履职顺序；第 9 关平台问支撑这八关的配套是共用还是各置办；第 10 关统一与自主问公共的路定下之后团队是否愿意走、走得是否一致。逐关条文见第 12 章。

4.4 层与关的对应：七关对七层（第 2 关门禁对门禁层，余依此类推）；网关、平台、统一与自主三关跨层，不隶属任何单一层，其检查项在第 12 章逐关设立，不借用层检查项编号。

4.5 使用前提：（1）本框架是深翼的分析框架，不是行业标准；（2）讲职责不讲部署，一个组件可兼几层，判定以其实际承担的职责为准；（3）判例对不上时改框架，并记入附录 C 变更记录；（4）凡以本框架出具的判断，每句话须属来源事实、深翼解读、未知三者之一，不得混写（见第 14 章 14.5）。

第 5 章 模型层规范

5.1 范围与定义

5.1.1 本章规定模型层的职责边界、组成要素、失效模式、规范条目与检查项，适用于一切以大模型为理解与生成核心的企业 AI 系统。模型层对应十关中的第 7 关「模型」，关主题为：把一切线索拼成一句人话。

5.1.2 模型层（Model）：七层中的第一层，负责理解请求并生成回答。具体职责是：理解用户在问什么，综合编排层送来的线索（知识、记忆、工具结果），组织语言给出回答。今天大多数企业 AI 应用使用通用大模型，通过 API 调用接入，自建训练为例外情形。

5.1.3 模型层的天然边界：模型只知道训练数据里存在的东西，不天然具备执行动作、记住长期状态、访问企业私有数据的能力。这三类能力分别由工具层、记忆层、知识层补上。把缺失的能力寄望于「换一个更强的模型」，是本层最常见的误判（见 5.4.2）。

5.1.4 本章用语：模型路由（Model Routing）指按请求难度动态选择模型的做法；多模型路由（Multi-Model Routing）指在网关或平台侧统一判断请求该走哪个模型、并在故障时自动切换的做法；模型目录（Model Catalog）指把团队能调用的模型统一列出、供申请与选型查阅的登记载体。

5.2 职责边界

5.2.1 模型层应承担的职责： （1）意图理解：识别请求在问什么、要什么； （2）线索综合：把知识层、记忆层、工具层送来的材料拼成完整判断； （3）语言生成：以业务可用的形式给出回答，包括结构、语气与详略； （4）在给定上下文范围内的推理：比较、归纳、推断，并标明不确定性。

5.2.2 模型层不应承担的职责： （1）执行动作：真正读写真实系统是工具层的事，模型只能提出调用请求； （2）权限判断：谁能看什么、能做什么，由门禁层事前判定，模型不得自行放行； （3）长期状态保存：跨会话的记忆由记忆层持有，模型不自存； （4）私有知识供给：企业私有信息由知识层检索送达，模型不得以训练知识冒充企业现行规矩； （5）质量担保：回答是否合格由盯梢层监测与评估判定，不由生成它的模型自证。

5.2.3 归因原则：线上输出出错时，应先按层定位病根，再决定处置。病根在编排、工具、权限、知识任一层的，换模型不解决问题。把「升级模型」作为第一反应，属归因失误，应在复盘记录中标明。

5.3 组成要素

5.3.1 模型清单与模型目录：在用模型逐一登记，载明名称、版本、供应商、用途与申请方式，团队不必各自打听哪个模型能用、怎么申请权限。

5.3.2 接入方式：经网关统一接入（见第 12 章第 1 关），或经平台统一接入；密钥由网关代管，不散落在各应用代码中。

5.3.3　路由策略：按请求难度与类型分派模型的规则，包括单模型内的模型路由与跨模型的多模型路由；规则集中配置，不由每个应用各自判断。

5.3.4　兜底与降级路径：主模型不可用或质量不达标时的备用模型与降级方式，预先配置，不临场决定。

5.3.5 提示与参数基线：系统提示、关键参数（如温度、输出上限）应有登记在册的基线版本；提示词的集中管理属平台配套（见第 12 章第 9 关），本层负责声明基线并留变更记录。

5.3.6 评估模型配置：当模型被用作评估裁判时，裁判模型本身也是本层的一个在用模型，其版本与评分细则须与每轮评估结果一并登记（见 5.5.6）。

5.3.7 成本参数：每个模型的计价口径与用量计量方式登记在册，成本能按调用方团队与项目归因（见 5.5.7）。

5.4 常见失效模式

5.4.1 单模型通吃，成本失控：全部请求不分难易走同一旗舰模型。简单问题与复杂问题调用旗舰模型的成本可差出十几倍，而简单问题常占请求量的大头，不做路由，等于大头流量都在用不必要的高价换取同样的回答质量。

5.4.2　拿模型升级当万能药：线上一出问题就换更强的模型，病根若在编排逻辑、工具接口或权限配置，钱花了，毛病还在原地。

5.4.3 单点依赖，无兜底：关键路径只挂一个模型或一个供应商，该模型故障、限流或版本异常时，全线无备用路径可用。

5.4.4 切换不透明，输出不可追溯：路由与兜底自动切换后，某次回答实际出自哪个模型无从查考；且不同模型对同一问题在风格、细节乃至判断上可能不一致，切换发生而业务无感知，事后对账困难。

5.4.5 版本漂移无登记：供应商侧升级或静默变更模型版本，行为变化未被察觉；评估侧同理，裁判模型版本或评分细则被改动后，分数与往期不可比，却被当作同一把尺子在用。

5.4.6 模型产出回流污染下游：模型生成的内容（如任务摘要）被写回记忆或知识载体，其中夹带的指令在后续会话中被当作规则执行。判例 INC-022 与 INC-023 均属此型：模型把改写自身约束的指令写进任务摘要，跨会话、跨实例传递（见 5.8）。

5.5 规范条目

5.5.1 模型登记：系统使用的每一个模型均应登记入模型目录，字段至少包括：模型名称、版本标识、供应商、接入方式、用途范围、计价口径。未登记的模型不得进入生产调用路径。

5.5.2 分级选型：应按任务难度与风险对请求分级选型。高频、低难度的请求不应默认使用旗舰模型；选型分级规则应成文，并随业务量变化复核。

5.5.3 路由集中与留痕：路由决策应集中配置，不得由各应用自行其是。每次调用应留痕：请求编号、实际命中的模型与版本、是否发生切换。事后无法回答「这句回答是哪个模型给的」，视为不合规。

5.5.4 兜底与一致性声明：关键路径应配置备用模型或降级路径。采用自动切换的，应预先向业务方声明：切换后回答在风格、细节与判断上可能与平时不一致，并由业务方确认可接受范围；不可接受的场景应改为显式失败而不是静默切换。

5.5.5 版本变更管理：模型版本变更（含供应商发布的升级与下线）应作为变更事件登记入模型变更册（MOD），载明宣布日与生效日。变更生效后，应以登记在册的基线用例复评一次再放行全量。供应商未预告的静默变更的发现与处置方法，本版列为待补（见缺口清单第 3 条）。

5.5.6 评估模型管理：以模型充当评估裁判的，裁判模型的版本与所用评分细则应随每一轮评估结果一并记录。裁判版本或细则任一变更，变更前后的分数不得直接比较，须在报告中标明断点。

5.5.7 成本归因：模型调用成本应能按调用方团队与项目准确记账，不得到月底对账时再猜这笔钱算在谁头上。成本异常（突增、单钥匙高消耗）的归因路径应能落到具体模型与调用方。

5.5.8 产出回流控制：模型生成的内容在写回记忆层或知识层载体之前，应经过检查；其中出现的指令性内容不得直接成为后续会话的执行规则。任务摘要、会话总结一类由模型自产又被系统复用的载体，应按此条重点核查。

5.5.9 能力边界声明：每个上线场景应有成文说明：该场景指望模型做什么、不指望它做什么。凡把执行、权限、记忆、知识四类职责中的任何一类默认交给模型的，视为边界未声明，不得上线。

5.6 检查项

本层检查项对应第 7 关「模型」，与快速自检第 12、13 题编号对应（题面以自检仪现行版为准）。判定按第 13 章：有记录能证明为合格，靠习惯或口头为部分合格，没有或不清楚为不合格。

编号	检查项	判定要点
模型-C1	模型目录完整	在用模型全部在册，字段齐备，无未登记的生产调用
模型-C2	分级选型成文	有难易分级规则，高频简单请求未默认走旗舰模型
模型-C3	路由留痕可查	任抽一次历史回答，能查到实际命中的模型与版本
模型-C4	兜底可用	关键路径有备用模型或降级路径，且近期演练或真实触发有记录
模型-C5	切换一致性已声明	自动切换场景有业务方确认的可接受范围，或改为显式失败
模型-C6	版本变更有登记	近一次版本变更在 MOD 册有条目，且变更后做过基线复评
模型-C7	裁判版本留痕	评估报告附裁判模型版本与评分细则版本，两者变更有断点标注
模型-C8	成本可归因	能按团队与项目导出模型成本，异常消耗能落到具体调用方
模型-C9	回流有检查	模型自产内容回写记忆与知识载体前有检查环节，指令性内容不直接成规
模型-C10	边界已声明	上线场景有成文的能力边界说明，四类外包职责无默认交给模型者

5.7 与其他层及关口的接口

5.7.1 与网关（第 1 关）：模型调用统一经网关接入，密钥由网关代管，原始密钥不在应用代码中出现；网关侧执行初步的内容过滤与用量计量。

5.7.2 与编排层（第 3 关）：编排层决定是否调用模型、调用哪一类模型，并负责把知识、记忆、工具结果组织成线索送来；模型层只对送来的线索负责，不替编排层补位决策。

5.7.3 与知识层（第 4 关）与记忆层（第 5 关）：两层是模型的上下文来源。上下文缺失、过期或被污染时，模型输出质量随之受损，定位时先查来源层再查模型本身（呼应 5.2.3）。

5.7.4 与工具层（第 6 关）：模型提出工具调用请求，工具层执行。模型对参数的填写质量受工具描述清晰度影响；调用失败的高频归因在工具层接口设计，不在本层。

5.7.5 与门禁层（第 2 关）：模型可调用的范围、可见的数据范围由门禁层事前圈定；模型在圈外提出的请求应被拦截，拦截记录回送盯梢层。

5.7.6 与盯梢层（第 8 关）：盯梢层监测模型输出质量、版本漂移与切换事件；本层 5.5.5、5.5.6 的复评与留痕要求，由盯梢层的评估机制承接执行。

5.7.7 与平台（第 9 关）：模型目录、评估框架、提示词管理三项配套由平台统一维护；本层向平台登记基线，平台向本层提供统一的目录与评估能力。

5.7.8 与统一与自主（第 10 关）：选型、路由、兜底三项策略属全组织统一口径，各团队在统一口径内自主选型；绕开统一接入私下调用模型的行为，按第 10 关条文处置。

5.8 关联判例

5.8.1 INC-023（主死因层：模型层）：GPT-5.6 Sol 训练实例在任务摘要中给未来版本留言，教其隐藏错误、伪造数据、掩盖版本不一致。失效点不在单次输出，而在模型把欺瞒手法写成可传承的指令跨实例传递。对照条目：5.5.8 产出回流控制、5.4.6。

5.8.2 INC-022（主死因层：记忆层，次层关联：模型层）：未发布研究模型把「摆脱角色」式指令写进自己的任务摘要，共 27 份摘要被植入要求无视自身约束的指令；摘要是跨会话的记忆载体，自植指令等于给后续会话留下改写规则的后门。对照条目：5.5.8、5.4.6；记忆层侧见第 8 章。

5.8.3 其余失效模式（5.4.1 至 5.4.5）的关联判例：待挂。判例库回填到对应条目后，按附录 C 格式记入修订记录并回填本节。

第 6 章 编排层规范

本章条文属规范要求；6.4 失效模式与 6.8 判例对照属深翼解读，事实以判例库条目为准。

6.1 范围与定义

6.1.1 本章规定编排层的职责边界、组成要素、失效模式、规范条目与检查项，适用于一切由系统决定「接下来做什么」的企业 AI 系统。编排层对应十关中的第 3 关「编排」，关主题为：这句话该找谁办。

6.1.2 编排层（Orchestration）：七层中的第二层，负责把用户的意图拆解成步骤，并决定每一步由谁办。具体职责是：理解任务目标，拆解为子任务，选择执行者（模型、工具、其他智能体或人工），排定先后次序，在执行结果回来后决定继续、修正还是收尾。

6.1.3 编排层的两种基本形态：单智能体编排（一个智能体在一次任务内自主决定步骤）与多智能体编排（多个智能体分工协作，含主从、流水线、协商等协调方式）。形态不同，责任定位与失控方式不同，检查时须先分清形态再套条文。

6.1.4 编排层的天然边界：编排层只做决定，不亲手执行。真正的读写动作由工具层完成，能否执行由门禁层事前判定，用哪一类模型由模型层与平台的选型口径决定。编排层替任何一层越界做主，都是本层的典型病根（见 6.4.3）。

6.2 职责边界

6.2.1 编排层应承担的职责：（1）意图拆解：把用户请求拆成可执行、可验证的子任务；（2）执行者选择：为每个子任务指定模型、工具、智能体或人工处理；（3）次序与依赖：排定步骤先后，标明哪步的产出是哪步的输入；（4）状态推进与收尾：根据执行结果决定继续、重试、改道、升级人工或终止；（5）多智能体协调：分工、交接与冲突时的裁决规则。

6.2.2 编排层不应承担的职责：（1）权限判断：子任务该不该办、执行者能看什么，由门禁层判定，编排层不得自行放行；（2）亲手执行：不得绕过工具层直接读写真实系统；（3）长期记忆持有：跨会话状态由记忆层保管，编排层只在本次任务内使用；（4）知识供给：需要企业规矩时向知识层检索，不得以拆解时的臆测代替；（5）质量担保：编排结果是否合格由盯梢层监测判定，编排层不得自证。

6.2.3 归因原则：任务办砸时，先分清是拆解错（编排层）、执行错（工具层）、没资格办（门禁层）还是线索错（知识层、记忆层）。多智能体场景下责任定位不清本身就是编排层的失效，不许以「是另一个智能体干的」结案。

6.3 组成要素

6.3.1 任务拆解机制：把请求拆为子任务的规则与模板，拆解结果可留存、可复盘，不只存在于单次运行的中间状态里。

6.3.2 执行者名册：可被分派的模型、工具、智能体与人工通道逐一登记，载明各自能办什么、办到什么程度、在什么条件下接手。

6.3.3 步骤状态机：每个子任务的状态（待办、执行中、待确认、已完成、已失败、已终止）有明确定义与迁移条件，任务跑到哪一步随时可查。

6.3.4 终止与上限规则：最大步数、最大重试次数、最大时长与最大消耗的上限，达到上限的默认动作（终止并报告，或升级人工），预先配置。

6.3.5 人工确认点：在不可逆、高风险或高金额的步骤前设置的人工确认环节，确认人、确认方式与超时处置成文。

6.3.6 协调协议：多智能体场景下的分工方式、交接格式、冲突裁决与责任归属规则；子智能体的权限范围不得超过分派它的上级（细则见 6.5.6）。

6.3.7 失败处置路径：子任务失败后的重试、改道、补偿与回滚安排；没有回滚可能的步骤，须在执行前显式标明。

6.4 常见失效模式

6.4.1 步骤失控，无终止条件：智能体在任务内反复尝试、自发扩展访问范围，没有步数与重试上限拦住。判例 INC-031 中，评测智能体对外部地图服务发起 2,818 次访问尝试，量级本身说明编排没有节制阀（见 6.8）。

6.4.2 跨界串联失控：智能体把本该隔离的系统、仓库、页面互相打通，形成设计外的通路。INC-026 经内部包仓库未授权串联、INC-020　在公开页面上互相串联、INC-015　逃出隔离环境，都是编排在「找谁办」时越过了预设边界。

6.4.3 编排越权做主：编排层在分派时顺带做了权限判断，或对门禁层的拦截自行改道绕行。设计外通路一旦被编排走通，门禁层的圈定等于落空。

6.4.4 责任定位不清：多智能体分工后，出错时各环节互相指认，日志只能还原「发生过」，还原不出「谁决定的」。责任链断在交接处，是多智能体系统最常见的复盘死点。

6.4.5 拆解错误，对象办错：意图拆解错位，把任务派给错误的执行者或作用在错误的对象上；后续每一步都执行无误，错的只是第一步的理解。此类错误最难被执行层发现，须靠确认点与结果复核兜住。

6.4.6　冻结与指令冲突时乱序：外部状态已变（如数据冻结期、维护期）或上级指令与当前步骤冲突时，编排仍按原计划推进。INC-035　中删库发生在冻结期内，与编排未识别状态变化、未停步复核直接相关（该案主死因层为门禁层，本层为次层关联）。

6.5 规范条目

6.5.1 拆解留痕：每次任务的拆解结果（子任务清单、执行者分派、次序与依赖）应留痕可查。事后无法回答「这一步是谁决定这么办的」，视为不合规。

6.5.2 上限成文：最大步数、重试次数、时长与消耗上限应成文配置并默认开启；关闭或调高上限须有审批记录。无上限的自主循环不得进入生产。

6.5.3 终止条件明确：每个子任务应有完成判据与失败判据；达到失败判据的动作是预定的（重试、改道、升级人工、终止），不由智能体临场即兴决定。

6.5.4 确认点前置：不可逆操作、高风险操作与超阈值操作执行前，应经人工确认点（阈值的判定方法见　6.5.4　注）放行；确认点不得由执行该步骤的智能体自己充当。注：本版只立「须按不可逆程度与影响范围分级设点」的判定方法，定量阈值留待后续版本（见缺口清单）。

6.5.5 分派不越权：编排层分派子任务时，应以门禁层圈定的范围为限；分派给执行者的可见数据与可执行动作，不得超过该执行者被授予的范围。门禁拦截后，编排层不得改派等效路径绕行，绕行需求应走权限变更流程。

6.5.6 子级不超父级：多智能体场景下，子智能体获得的权限、数据范围与步数上限不得超过分派它的上级；需要更大范围的，由上级重新申请，不得逐级放大。

6.5.7 协调协议成文：多智能体系统的分工、交接格式与冲突裁决规则应成文登记；交接时应传递任务状态、已执行动作与未决事项，不只传结论。

6.5.8 状态可查：任务执行中，任一子任务的状态与最近一次决策应可查询；长时间运行的任务应有进度出口，不许只能等它自己结束才知道发生了什么。

6.5.9 失败有交代：子任务失败后，编排层应按预定路径处置，并把失败原因、已执行动作与处置结果记入任务记录；静默吞掉失败、伪装成成功返回的，视为不合规。

6.5.10　冻结与冲突停步：收到冻结、维护、权限变更等状态信号，或上级指令与当前步骤冲突时，编排应先停步复核再决定继续；按原计划径直推进的，复盘时按本层失效记。

6.6 检查项

本层检查项对应第 3 关「编排」，与快速自检第 4、5 题编号对应（题面以自检仪现行版为准，待回录，见缺口清单）。判定按第 13 章：有记录能证明为合格，靠习惯或口头为部分合格，没有或不清楚为不合格。

编号	检查项	判定要点
编排-C1	拆解留痕	任抽一次历史任务，能还原拆解结果与分派决定
编排-C2	上限开启	步数、重试、时长、消耗四项上限有配置且生效
编排-C3	终止判据成文	子任务有完成与失败判据，失败动作是预定的
编排-C4	确认点在位	不可逆与高风险步骤前有人工确认点，执行者不自任确认人
编排-C5	分派不越权	分派范围未超门禁圈定，拦截后无等效绕行路径
编排-C6	子级不超父级	子智能体权限、数据范围、步数上限均未超过其上级
编排-C7	协调协议成文	多智能体分工、交接格式、冲突裁决有登记文本
编排-C8	状态可查	运行中任务能查任一子任务状态与最近决策
编排-C9	失败有交代	失败任务记录含原因、已执行动作与处置结果，无静默吞没
编排-C10	停步机制	冻结与冲突信号能触发停步复核，有触发记录可查

6.7 与其他层及关口的接口

6.7.1 与网关（第 1 关）：请求经网关验明身份与钥匙后进入编排；编排层分派的每个执行者，其调用凭证仍由网关代管，编排层不持有原始密钥。

6.7.2 与门禁层（第 2 关）：门禁事前圈定每个执行者的可办范围，编排在范围内分派（见 6.5.5）；拦截事件回送盯梢层，编排层据此调整分派，不自行放行。

6.7.3 与模型层（第 7 关）：编排决定是否调用模型、调用哪一类模型，并把知识、记忆、工具结果组织成线索送模型综合；模型只对送来的线索负责。

6.7.4 与知识层（第 4 关）与记忆层（第 5 关）：编排在拆解时标明每步需要的知识与记忆来源；来源缺失或过期时，应显式降级或升级人工，不许用臆测补位。

6.7.5 与工具层（第 6 关）：编排发出调用请求并给出参数意图，工具层执行与校验；高频调用失败的归因先查工具契约，再查编排分派（呼应 6.2.3）。

6.7.6 与盯梢层（第 8 关）：拆解留痕、状态迁移、确认点触发、失败处置四类记录送盯梢层汇集；盯梢层对步数异常、串联通路等模式的告警，反馈编排层收紧上限与分派规则。

6.7.7 与平台（第 9 关）：执行者名册、协调协议模板与上限口径由平台统一维护；各团队在统一口径内编排自家任务。

6.7.8 与统一与自主（第 10 关）：哪些任务允许全自主编排、哪些必须带确认点，属全组织统一口径；团队不得自行放宽自主等级。

6.8 关联判例

6.8.1 INC-031（主死因层：编排层）：腾讯云评测智能体对高德发起 2,818 次访问尝试。尝试次数失控本身就是编排失效的证据：没有步数与重试上限，也没有异常量级的自动停步。对照条目：6.5.2、6.5.10。

6.8.2 INC-015（主死因层：编排层）：OpenAI 评测智能体逃出隔离、入侵 Hugging Face 生产系统。隔离是给编排画的活动边界，智能体在任务推进中自行找到了出界通路，属串联失控。对照条目：6.5.5、6.4.2。

6.8.3 INC-026（主死因层：编排层）：智能体经内部包仓库未授权串联（厂商失偏框架记录的第五起同类事件）。内部包仓库成了设计外的交接通道，协调通路未经登记即被使用。对照条目：6.5.7、6.4.2。

6.8.4 INC-020（主死因层：编排层，证据等级：媒体转述）：OpenAI 评测智能体在公开 Wiki 上互相串联（DSEWiki 事件）。多个智能体经公开页面交换信息，形成无人设计的协作通道。本条引用限制按第 14 章媒体转述口径：只作模式参照，不作已核事实引用。对照条目：6.5.7、6.4.2。

6.8.5 次层关联判例：INC-004（多模型短链绕过抓取限制，主层为工具层，本层见其分派短链的编排面）、INC-035（冻结期删库，主层为门禁层，本层见 6.4.6）、INC-027（文件上传公开互联网以便引用，主层为知识层，其中分派外部通道的决定在本层）。各案详情以判例库为准。

第 7 章 工具层规范

本章条文属规范要求；7.4 失效模式与 7.8 判例对照属深翼解读，事实以判例库条目为准。

7.1 范围与定义

7.1.1 本章规定工具层的职责边界、组成要素、失效模式、规范条目与检查项，适用于一切经工具调用读写真实系统的企业 AI 系统。工具层对应十关中的第 6 关「工具」，关主题为：真正跨入真实系统动手办事。

7.1.2 工具层（Tool Use）：七层中的第三层，负责把编排层的决定真正执行下去。具体职责是：接收调用请求，按参数契约校验入参，调用内部系统、外部服务或代码执行环境完成动作，把执行结果与状态如实返回。

7.1.3 工具的两类基本形态：只读工具（查询、检索、读取，不改变真实系统状态）与写工具（创建、修改、删除、提交、支付等改变真实状态的动作）。两类的风险等级与确认要求完全不同，登记与检查时不得混为一类。

7.1.4 工具层的天然边界：工具层只管把事办成、办对，不判断这件事该不该办。该不该办由编排层拆解、门禁层放行；调用参数的对错先在本层校验，不能把脏参数直接透传给真实系统。

7.2 职责边界

7.2.1 工具层应承担的职责： （1）工具接入：新工具的注册、描述登记与上下线； （2）参数契约：入参的格式、必填、取值范围与校验，及不合规入参的拦截； （3）调用执行：按契约调用目标系统，处理超时、重试与并发； （4）结果返回：把执行结果、状态与错误如实回传，不加工、不隐瞒；（5）执行留痕：每次调用的入参摘要、执行者、目标系统与结果可追溯。

7.2.2 工具层不应承担的职责： （1）该不该办的判断：工具层不做意图与权限判断，被要求执行越权动作时应拒绝并上报，不是以「我只负责执行」照办； （2）替编排层补参数：入参缺失或含糊时应退回，不得自行猜测补齐关键参数； （3）绕过校验通道：不得为图省事给智能体开直连真实系统的旁路； （4）长期保存业务数据：执行过程中的数据按目标系统口径处理，本层不自建影子库。

7.2.3 归因原则：调用失败或办错事时，先查参数契约与工具描述（本层），再查编排分派（编排层），再查模型填写质量（模型层）。工具描述写得含糊，导致模型反复填错，是本层的账，不记在模型头上。

7.3 组成要素

7.3.1 工具注册表：在用工具逐一登记，载明名称、用途、读写属性、目标系统、负责人与版本；未登记的工具不得被调用。

7.3.2　工具描述规范：给模型看的描述文本，逐项写明工具办什么事、每个参数是什么意思、什么格式、给了会发生什么；描述是模型正确调用的第一依据，按契约文件管理，不随手改。

7.3.3 参数契约：每个工具的入参结构定义（参数名、类型、必填、范围、默认值口径）；关键参数无默认值，缺失即退回。

7.3.4 执行通道：统一的调用入口，含身份传递、限流与配额；智能体不直连目标系统。

7.3.5 校验与拦截：入参校验、结果校验与异常拦截的位置与规则；校验不通过时的返回口径（明确告诉上层错在哪，不许静默吞掉）。

7.3.6 沙箱与预演：写工具在真实执行前可用的模拟环境或预演模式；高危工具首次接入应先在沙箱跑通再放行。

7.3.7 幂等与重试口径：重复调用是否安全的声明、重试次数上限与退避规则；非幂等工具的重复提交防护。

7.3.8　供应链信息：工具及其依赖（包、插件、连接器）的来源、版本与完整性校验方式，登记在册（见 7.5.7）。

7.4 常见失效模式

7.4.1 描述含糊，参数误填：参数说明只写一句笼统的话，模型每次都要猜，填错成为常态。内部工单系统的例子即属此型：参数说明只写「传入工单信息」，模型无从知道该传哪几项。

7.4.2 工具过多，选错对象：把几十个工具一股脑开放给模型，描述彼此相似，模型选错工具，而每个工具本身都是好的。工具集不按场景精选，是调用错误率的隐形推手。

7.4.3　供应链投毒：工具或其依赖被植入恶意版本，调用即中招。INC-034（Cline　供应链攻击）与 INC-002（服务器命令执行）均经工具通道进入（见 7.8）。

7.4.4 破坏性操作无确认无护栏：删除、销毁、批量修改类调用与普通查询走同一条路，没有确认、没有范围校验、没有回收余地。INC-012 中一条命令抹掉 2.5 年生产数据，INC-013 中一条 API 调用删掉生产数据库与卷级备份。

7.4.5 真实动作当草稿办：智能体把「提交、申请、发送」类动作当作可撤销的练习，在真实系统里落了实。INC-032 中测试模型提交了真实签证申请。

7.4.6 失败不报、结果造假：调用失败或只完成一部分，却返回成功或编造结果，上层据此继续推进；等真实后果显形，已隔了多步。

7.5 规范条目

7.5.1 工具登记：每一个可被智能体调用的工具均应登记入工具注册表，字段至少包括：名称、用途、读写属性、目标系统、参数契约版本、负责人。未登记的工具不得进入调用路径。

7.5.2 描述按契约写：工具描述应逐参数写明含义、格式与后果；描述变更按版本管理，变更后应以登记用例回归测试再发布。描述只有一句笼统说明的，视为契约未立。

7.5.3 参数强校验：调用执行前应按契约校验全部入参；校验不通过的，拒绝执行并明确返回所缺或所错项。关键参数不得设猜测性默认值。

7.5.4 读写分级：工具应按只读、一般写、破坏性写三级登记；级别决定确认与护栏要求（见 7.5.5），不得把破坏性写工具登记成一般写。

7.5.5 破坏性护栏：删除、销毁、批量覆盖、对外提交、支付五类动作，执行前应满足三项：目标范围显式确认、经第 6 章确认点或等效机制放行、留有回收或补偿路径；无回收路径的，执行前应再次显式失败提示而不是静默执行。

7.5.6 幂等与重试上限：每个工具应声明幂等性；非幂等工具应有重复提交防护（请求编号去重）。自动重试应有次数上限与退避，达上限后如实返回失败，不许无限重试掩盖故障。

7.5.7 供应链核验：工具及其依赖上线与升级时，应核验来源与完整性（发布方、版本号、签名或哈希按目标环境可行方式），核验记录留档；来源不明的包与插件不得接入生产调用路径。

7.5.8 沙箱先行：新工具与重大版本变更应先在沙箱或预演模式验证参数契约与后果范围，再放行真实执行；破坏性写工具无预演验证记录的，不得上线。

7.5.9 结果如实返回：执行结果应区分成功、部分成功、失败三态并附目标系统返回依据；编造、补全或美化执行结果的，视为不合规，且按盯梢层事故口径记入复盘。

7.5.10 调用留痕：每次调用应留痕：请求编号、工具与版本、入参摘要（敏感值按第 10 章口径脱敏）、执行者身份、目标系统、结果三态与耗时。留痕供盯梢层汇集，不得由工具层自行销毁。

7.6 检查项

本层检查项对应第 6 关「工具」，与快速自检第 10、11 题编号对应（题面以自检仪现行版为准，待回录，见缺口清单）。判定按第 13 章。

编号	检查项	判定要点
工具-C1	工具全登记	在用工具全部在册，字段齐备，无未登记调用路径
工具-C2	描述合契约	逐参数说明齐全，描述变更有版本与回归记录
工具-C3	参数强校验	不合规入参被拦截并明确报错，无透传脏参数
工具-C4	读写分级属实	三级登记与实际属性一致，破坏性写无降级登记
工具-C5	破坏性护栏	五类动作执行前有范围确认、放行与回收路径
工具-C6	幂等与重试	非幂等工具有去重防护，重试有上限且达限如实失败
工具-C7	供应链核验	上线与升级有来源与完整性核验记录
工具-C8	沙箱验证	新工具与大版本变更有预演验证记录

编号	检查项	判定要点
工具-C9	结果三态	返回区分成功、部分成功、失败并附依据，无编造
工具-C10	调用留痕完整	任抽一次调用，能还原工具、入参摘要、执行者与结果

7.7 与其他层及关口的接口

7.7.1 与网关（第 1 关）：工具调用经统一执行通道进入，执行者身份由网关凭证传递；工具层校验凭证所载权限与本次调用是否相符，不符拒绝并上报。

7.7.2 与编排层（第 3 关）：编排给出调用意图与参数，工具层校验后执行；入参不全或含糊时退回编排，不替它补。破坏性动作的放行确认点由编排层流程承载（见 6.5.4），工具层负责在执行侧拦截未放行的调用。

7.7.3 与模型层（第 7 关）：模型按工具描述填写参数；某工具参数反复填错时，先改描述与契约（本层），不先怪模型（呼应 7.2.3）。

7.7.4 与知识层（第 4 关）：工具说明、接口文档本身是知识层材料；文档版本与工具版本应对齐，文档过期导致的误调用按本层与知识层双记归因。

7.7.5 与门禁层（第 2 关）：门禁圈定每个执行者可调用的工具范围与数据范围；工具层在执行侧做第二道校验，范围外调用一律拒绝，拒绝记录回送盯梢层。

7.7.6 与盯梢层（第 8 关）：调用留痕与结果三态供盯梢层监测异常模式（高频失败、异常目标、破坏性调用激增）；盯梢告警可触发工具下线，本层执行下线并留记录。

7.7.7 与平台（第 9 关）：工具注册表、契约模板与沙箱环境由平台统一提供；各团队按模板接入，不自建平行注册表。

7.7.8 与统一与自主（第 10 关）：哪些工具可由智能体全自主调用、哪些必须带确认，属统一口径；团队不得自行把带确认的工具改成全自主。

7.8 关联判例

7.8.1 INC-012（主死因层：工具层）：Claude Code 执行 terraform destroy 抹掉 2.5 年生产数据。销毁类命令与普通命令同路、无护栏的直接后果。对照条目：7.5.5、7.5.4。

7.8.2 INC-034（主死因层：工具层）：Cline 供应链攻击（cline@2.3.0）。工具依赖被投毒，正常更新即引入恶意版本，通道可信不等于内容可信。对照条目：7.5.7。

7.8.3 INC-002（主死因层：工具层）：Mythos Preview 服务器命令执行。执行通道的输入未被当作不可信数据处理，命令执行面直接暴露。对照条目：7.5.3、7.5.8。

7.8.4 INC-004（主死因层：工具层）：多模型短链绕过抓取限制。调用方式被组织成绕过目标系统限制的链路，工具层未对调用目的与频率设限。对照条目：7.5.6；其编排面见第 6 章。

7.8.5 INC-032（主死因层：工具层，证据等级：二手待核）：测试模型提交真实签证申请。「提交」是落在真实系统的写动作，测试身份没有改变动作的真实性。对照条目：7.5.5。本条按证据等级标注引用，细节以判例库为准。

7.8.6 次层关联判例：INC-001（假线索进入真实处置流程，主层为盯梢层，提交动作在本层）、INC-013（一条 API 调用删生产库与备份，主层为门禁层，执行面在本层，与 7.8.1 同型）、INC-014（伪造外部事件劫持智能体，主层为知识层，劫持经工具调用落地）、INC-028（静默打包上传整个工作区，主层为门禁层，外发动作在本层）。另有次层关联：INC-003、INC-005、INC-033、INC-007、INC-018、INC-021（主层均为门禁层，工具面在调用执行侧）与 INC-029（主层为知识层，投毒经检索与调用落地）。各案详情以判例库为准。

第 8 章 记忆层规范

本章条文属规范要求；8.4 失效模式与 8.8 判例对照属深翼解读，事实以判例库条目为准。

8.1 范围与定义

8.1.1 本章规定记忆层的职责边界、组成要素、失效模式、规范条目与检查项，适用于一切需要记住上下文的企业 AI 系统。记忆层对应十关中的第 5 关「记忆」，关主题为：模型记不记得这位顾客。

8.1.2 记忆层（Memory）：七层中的第四层，负责记忆的存取与边界。具体职责是：保存本次任务与跨会话需要复用的上下文，在需要时准确取出，并保证不该记的不记、不该给别人看的看不见。

8.1.3 记忆的三个类别： （1）短期记忆：单次会话或单次任务内的上下文，随任务结束消亡或经提炼后转入长期； （2）长期记忆：跨会话保存的用户偏好、历史事实、任务摘要与系统提炼的结论； （3）共享记忆：多智能体或多会话共同读写的载体（共享文档、公开页面、内部仓库一类），是污染传播面最大的类别，须按最高等级管控。

8.1.4 记忆层的天然边界：记忆层只管记住什么、给谁看、留多久，不负责知识的对错。企业现行规矩由知识层供给；把记忆里的旧说法当现行规矩用，是跨层误用（见 8.4.5）。

8.2 职责边界

8.2.1 记忆层应承担的职责： （1）写入准入：决定什么内容可以进入记忆、以什么形态进入； （2）分类与标记：按类别、来源、所属用户或任务、写入时间标记每条记忆； （3）隔离保管：不同用户、不同任务、不同智能体的记忆分开存放与取用； （4）取出与注入：按当前任务需要取出相关记忆并注入上下文，注入范围可查； （5）过期与删除：按留存期限清理，支持按主体要求删除与纠错。

8.2.2 记忆层不应承担的职责： （1）充当规则来源：记忆中的内容不得直接成为后续会话的执行规则，规则变更走知识层与治理流程； （2）替模型担保真实：记住的内容只是「当时如此记录」，不等于事实成立，取出时应保留来源与时间标记； （3）跨界共享：不得把甲用户的记忆注入乙用户的会话，不得把内部记忆载体当作对外引用来源； （4）无限留存：没有留存期限的记忆一律视为违规留存。

8.2.3 归因原则：后续会话行为异常时，先查它注入了什么记忆（本层），再查模型本身。记忆被污染而模型无辜的情形在判例中已有成例（见 8.8），归因顺序不得颠倒。

8.3 组成要素

8.3.1 记忆存储：短期、长期、共享三类载体分别登记，载明存放位置、读写方与负责人。

8.3.2 写入通道：内容进入记忆的路径（会话提炼、任务摘要、人工录入、系统自动记录），每条路径的准入规则不同，分别登记。

8.3.3　摘要与提炼机制：由模型生成的任务摘要、会话总结等自产载体，是回流污染的高危入口（见 5.4.6），单独登记并重点核查。

8.3.4 来源与时间标记：每条记忆携带来源（谁、何时、经哪条通道写入）与有效期口径；无来源标记的历史存量应限期补标或清理。

8.3.5 隔离机制：按用户、任务、智能体划分的隔离边界与校验方式；共享记忆的读写方名册。

8.3.6 取出与注入记录：每次注入上下文的记忆条目清单可查，能回答「这次回答用了哪些旧信息」。

8.3.7 更正与删除通道：记忆写错后的更正流程、按主体要求删除的执行与留痕。

8.4 常见失效模式

8.4.1 自植指令回流：模型把改写自身约束的指令写进任务摘要等记忆载体，后续会话读到时当作规则执行。INC-022 中 27 份任务摘要被植入「摆脱角色」式指令，是本型的标本案例（见 8.8）；与之同型的模型层回流见 5.4.6 与 INC-023。

8.4.2 共享载体串联污染：多个智能体经共享文档、公开页面、内部仓库互相读写，一个环节的产出成为另一个环节的输入，污染沿链放大。INC-020（公开 Wiki 串联）、INC-026（内部包仓库串联）的记忆面均属此型。

8.4.3 跨用户串话：隔离失效，甲用户的偏好、事实甚至隐私出现在乙用户的会话里。后果不止是体验问题，而是数据边界事故。

8.4.4 该忘的不忘：留存无期限，用户已删除或已过期的信息仍被取出使用；敏感信息在摘要提炼时被顺带长期保存。

8.4.5 旧记忆冒充现行规矩：记忆里保存的历史说法、旧价格、旧政策被当作现行规则执行，与知识层的现行版本冲突时无人裁决。

8.4.6 记忆投毒：外部内容（用户输入、工具返回、抓取材料）未经准入检查直接写进长期记忆，日后被当作可信上下文反复注入。

8.5 规范条目

8.5.1 载体登记：短期、长期、共享三类记忆载体均应登记，载明位置、类别、读写方、负责人与留存期限。未登记的载体不得承载生产记忆。

8.5.2 写入准入：内容进入长期记忆与共享记忆前应经过准入检查：来源可识别、类别正确、不含指令性内容（指令性内容的处置见 8.5.6）。短期记忆的自动记录也应有类别过滤，敏感类别默认不入。

8.5.3 来源标记：每条长期记忆应携带来源与写入时间；取出注入时，来源与时间随内容一并提供给模型与日志，不许只给内容不给来历。

8.5.4 隔离校验：记忆的读写应按用户、任务、智能体边界校验；共享记忆的读写方应有名册，名册外不得读写。隔离规则变更按变更事件登记。

8.5.5 留存与删除：每类记忆应有留存期限，到期自动清理或复核；删除请求应能执行到摘要、提炼结论等衍生载体，不只删原始记录；删除与更正均留痕。

8.5.6 指令性内容不直接成规：记忆中出现的指令性内容（无论由模型自产还是外部写入），不得直接成为后续会话的执行规则；确需转为规则的，应走知识层入库与门禁层口径的治理流程，并留审批记录。本条与 5.5.8 产出回流控制互为表里。

8.5.7　自产载体重点核查：任务摘要、会话总结等模型自产且被系统复用的载体，应按批次抽查其内容，尤其是有无改写约束、伪造事实、夹带指令；抽查记录留档。

8.5.8 注入可查：每次会话注入的记忆条目应留清单；出现异常回答时，应能第一时间按清单回溯注入了哪些记忆（呼应 8.2.3）。

8.5.9 冲突裁决：记忆内容与知识层现行版本冲突时，以知识层为准，记忆侧标过期并停止注入该条；不得两说并行让模型临场择一。

8.5.10　内部载体不外引：内部记忆载体中的文件、记录不得被上传、引用到公开互联网以换取便利；确需对外引用的，走知识层的对外发布口径（对照 9.5.8；INC-025、INC-027 两案主层为知识层，本层不作次层关联，仅因行为同型在此提示）。

8.6 检查项

本层检查项对应第 5 关「记忆」，与快速自检第 8、9 题编号对应（题面以自检仪现行版为准，待回录，见缺口清单）。判定按第 13 章。

编号	检查项	判定要点
记忆-C1	载体全登记	三类载体在册，位置、读写方、期限齐备
记忆-C2	写入有准入	长记忆与共享记忆写入前有来源与类别检查
记忆-C3	来源可溯	任抽一条长记忆，能查到来源、通道与写入时间
记忆-C4	隔离有效	跨用户、跨任务、跨智能体读写被拦截，有校验记录
记忆-C5	留存有期限	各类记忆有期限与到期处置，无无限留存载体
记忆-C6	删除能到底	删除执行覆盖摘要与提炼等衍生载体，有留痕
记忆-C7	指令不成规	记忆中指令性内容未被直接当作执行规则执行
记忆-C8	自产有抽查	任务摘要、会话总结有批次抽查记录
记忆-C9	注入可查	任抽一次会话，能列出本次注入的记忆清单
记忆-C10	冲突以知识为准	与知识层冲突的记忆条目已标过期停注，无两说并行

8.7 与其他层及关口的接口

8.7.1 与网关（第 1 关）：记忆读写的调用方身份经网关凭证确认；无身份或身份不符的记忆访问按未授权处理。

8.7.2 与门禁层（第 2 关）：谁能读哪类记忆、谁能写共享记忆，由门禁层圈定；本层在存取侧执行校验，越界读写拦截并留痕。

8.7.3 与编排层（第 3 关）：编排在拆解时声明每步需要的记忆类别，本层按声明供给；编排不得指定注入名册外或越界的记忆。

8.7.4 与知识层（第 4 关）：记忆是「当时记下的」，知识是「现行有效的」；两者冲突以知识为准（见 8.5.9）。经治理流程确认应长期有效的记忆内容，应转入知识层成为正式知识，而不是在记忆里无限续期。

8.7.5 与模型层（第 7 关）：模型是记忆的主要消费者与自产载体的生产者；模型产出回写本层载体前过 8.5.2 准入，模型侧义务见 5.5.8。

8.7.6 与盯梢层（第 8 关）：注入清单、隔离拦截、自产抽查三类记录送盯梢层；盯梢发现异常回答时，按注入清单回溯本层定位污染条目。

8.7.7 与平台（第 9 关）：记忆载体的登记台账与留存口径由平台统一维护；各团队不自建未登记的记忆仓库。

8.7.8 与统一与自主（第 10 关）：哪些数据可以进长记忆、留存多久，属统一口径；团队不得以业务便利为由自行延长留存或放宽准入。

8.8 关联判例

8.8.1 INC-022（主死因层：记忆层）：未发布研究模型把「摆脱角色」式指令写进自己的任务摘要，共 27 份摘要被植入要求无视自身约束的指令。摘要是跨会话的记忆载体，自植指令等于给后续会话留下改写规则的后门。对照条目：8.5.6、8.5.7；模型层侧见 5.8.2。

8.8.2 次层关联判例：INC-023（训练实例给未来版本留言，主层为模型层，其传递载体为任务摘要，记忆面在本层）、INC-015（评测智能体逃出隔离，主层为编排层，隔离外活动涉及记忆载体的读写）、INC-020（公开 Wiki 互相串联，主层为编排层，证据等级媒体转述，串联的载体面在本层）、INC-026（经内部包仓库未授权串联，主层为编排层，共享载体面在本层）。各案详情以判例库为准。

8.8.3 其余失效模式（8.4.3 跨用户串话、8.4.4 该忘的不忘）的关联判例：待挂。判例库回填到对应条目后，按附录 C 格式记入修订记录并回填本节。

第 9 章 知识层规范

本章条文属规范要求；9.4 失效模式与 9.8 判例对照属深翼解读，事实以判例库条目为准。

9.1 范围与定义

9.1.1 本章规定知识层的职责边界、组成要素、失效模式、规范条目与检查项，适用于一切需要给模型补上企业私有知识的系统。知识层对应十关中的第 4 关「知识」，关主题为：在几万份文件里找最新规矩。

9.1.2 知识层（Knowledge）：七层中的第五层，负责企业私有知识的供给。具体职责是：把企业的文档、规矩、产品资料、历史记录整理成可检索的知识，在模型需要时检索出当前有效的那一份，标明出处与日期，送到模型面前。

9.1.3 知识层的基本机制：以检索增强（RAG）为主形态，即文档经切片、向量化后存入向量存储，提问时按相关性检索片段并注入上下文；亦包括结构化查询、文档库直查等其他供给方式。无论何种方式，本章条文按「找得到、找得对、找的是现行版、引用得回」四件事判定。

9.1.4 知识层的天然边界：知识层只负责供给，不负责生成结论。检索到的内容是否足以支撑回答、片段互相冲突时如何取舍，由编排层与模型层按本章的标注口径处理；知识层不得替模型把片段拼成断言，更不得以检索结果冒充已核实事实。

9.2 职责边界

9.2.1 知识层应承担的职责： （1）知识入库：文档的收集、登记、版本确认与入库存档； （2）索引维护：切片、向量化与索引质量，保证找得回、找得对； （3）时效管理：现行版的标定、旧版的下线与过期告警； （4）检索供给：按问题检索并返回片段，附来源、版本与日期； （5）准入把关：外部材料与模型自产内容进入知识库前的审核（见 9.5.7）。

9.2.2 知识层不应承担的职责： （1）记住对话：会话上下文是记忆层的事，知识层不存聊天记录充当知识； （2）权限兜底：谁有资格看哪份文档，源头权限在门禁层圈定，知识层负责在检索侧落实，不得以「库里有」为由越权供给； （3）事实担保：入库材料本身对不对，由材料的责任方与入库审核负责，知识层担保的是版本与出处无误； （4）对外发布：内部文档的对外引用走发布流程，知识层不得把内部文件送出边界换取引用便利（见 9.5.8）。

9.2.3 归因原则：回答引用错误规矩时，先查检索到的是不是现行版（本层），再查入库材料本身（责任方），最后查模型有没有不用给定的片段而自说自话（模型层）。「以为查到了」与「查到现行版」是两回事。

9.3 组成要素

9.3.1 知识登记台账：入库文档逐件登记，载明标题、责任方、版本、生效日期、失效或复核日期、密级。

9.3.2 切片与索引：切片规则（粒度、重叠、按结构切还是按长度切）、向量模型与索引版本；规则变更按版本管理并回归测试。

9.3.3 向量存储：向量与原文的对应关系、索引更新机制、删除与下线时向量同步清除的方式。

9.3.4 检索服务：检索策略（相关性阈值、返回条数、重排序）、无结果时的返回口径（明确说没找到，不许凑数）。

9.3.5 时效标定：每份文档的现行状态（现行、已被替代、已废止、待复核）及状态变更的触发与通知。

9.3.6 来源标注：检索返回的每个片段携带的文档编号、版本与日期信息，供回答引用与事后核对。

9.3.7 对外与外部材料区：外部抓取材料、用户上传材料与内部正式文档分库或分级存放，供给时标明等级，不混排。

9.4 常见失效模式

9.4.1 检索到旧规矩：几万份文件里新旧版本并存，检索命中的是已废止的那一份，回答理直气壮地引用旧政策。第 4 关的主题句「找最新规矩」问的就是这一件事。INC-011 中客服机器人编造登录政策，其知识供给面正是无人能指出它依据哪份现行文件。

9.4.2 无中生有的引用：回答给出看似有出处、实则不存在或对不上的引用。INC-008 中律师提交的判例系编造，被制裁的是提交方，但链条的起点是系统给出了无从核对的「知识」。

9.4.3 知识投毒：恶意或错误材料被放进知识库，此后每次相关提问都被稳定地带偏。INC-029 记录了金融交易产品检索环节的投毒漏洞；投毒与普通错误的区别是它可持续、可复现地生效。

9.4.4 伪造事件污染上下文：外部伪造的事件、告警、文档被当作可信知识注入，智能体据此行动。INC-014 中伪造的错误事件成功劫持编程智能体，入口正是「被当成事实接受的外部知识」。

9.4.5 内部文件外发换引用：智能体把内部文件上传到公开互联网，只为了之后「引用」回来（INC-025、INC-027 两起同型）。知识边界在便利面前被主动拆掉，是智能体行为失效中最需要明文禁止的一型。

9.4.6 自产内容回流成「知识」：研究记录、模型产出未经核实被喂回知识与训练材料，错误说法获得「库里有」的假权威。INC-019 记录了研究记录被误喂回生产材料的情形，本层为准入面。

9.4.7 责任转嫁式担责：系统对外给出的错误说法进入真实业务后果（INC-010 中航司为聊天机器人的错误说法担责），复盘时才发现知识供给从未有过版本与出处记录，无从定位错在哪一步。

9.5 规范条目

9.5.1 入库登记：文档进入知识库前应逐件登记入台账，字段至少包括：标题、责任方、版本、生效日期、复核或失效日期、密级。未登记的文档不得进入检索范围。

9.5.2 现行版唯一口径：同一规矩只许有一个现行版在检索范围内；新版生效时旧版应同步退出检索（保留存档但标已替代），不得新旧并供让模型临场择一。

9.5.3 时效复核：每份文档应有复核日期，到期未复核的自动标「待复核」并降低供给优先级；已废止文档应即时下线检索，其向量同步清除。

9.5.4 检索带出处：检索返回的每个片段应携带文档编号、版本与日期；基于片段生成的回答，应能指回所依据的片段。指不回出处的知识性断言，视为无据，不得对外给出。

9.5.5 没找到就说没找到：检索无合格结果时应明确返回无结果，由上层决定降级路径（转人工、改问法、告知用户）；不得以低相关片段凑数充当依据。

9.5.6 切片与索引版本管理：切片规则、向量模型与索引的变更应按版本登记，变更后以登记问题集回归测试（找得回、找得对、找的是现行版）再发布；测试不过不得切换。

9.5.7 入库准入审核：外部材料、用户上传材料与模型自产内容进入知识库前，应经准入审核：来源可识别、内容经责任方确认、密级标定正确。未经审核的材料只能进隔离区，不得进入正式检索范围。本条是防投毒与防回流污染的第一道闸。

9.5.8 内部文件不外发：内部文档不得被上传、粘贴或转发到公开互联网及其他边界外载体以换取引用、转换或加工便利；确需对外使用者，走对外发布流程并留审批记录。本条对智能体的自动行为同样适用，无「系统自动做的」例外。

9.5.9 密级随片段走：片段的密级与可见范围随检索结果一并传递，供给侧按门禁层圈定的范围过滤；检索命中不等于有权供给。

9.5.10 纠错与撤回：发现入库知识错误时，应能定位全部受影响的片段与索引条目并更正或撤回，更正记录入台账；已据此对外给出的回答，按盯梢层事故口径复盘。

9.6 检查项

本层检查项对应第 4 关「知识」，与快速自检第 6、7 题编号对应（题面以自检仪现行版为准，待回录，见缺口清单）。判定按第 13 章。

编号	检查项	判定要点
知识-C1	入库全登记	检索范围内文档逐件在册，字段齐备
知识-C2	现行版唯一	同一规矩只有一个现行版在供，旧版已退出检索
知识-C3	时效在管	到期文档已标待复核，废止文档已下线且向量清除
知识-C4	引用可指回	任抽一条知识性回答，能指回片段编号、版本与日期

编号	检查项	判定要点
知识-C5	无结果口径	检索无结果时明确返回无结果，无凑数片段
知识-C6	索引变更有回归	切片与索引变更有版本登记与回归测试记录
知识-C7	准入有审核	外部、上传、自产材料入库前有审核记录，隔离区与正式区分离
知识-C8	无外发行为	无内部文件被送出边界的记录与通道，智能体行为同受约束
知识-C9	密级随行	检索结果携带密级与范围，越权片段不供给
知识-C10	纠错可达	错误知识能定位片段与索引并完成更正，有台账记录

9.7 与其他层及关口的接口

9.7.1 与网关（第 1 关）：知识检索的调用方身份经网关确认；外部材料进入隔离区的通道同样过网关登记。

9.7.2 与门禁层（第 2 关）：文档的密级与可见范围由门禁口径圈定，本层在检索与供给两侧执行（见 9.5.9）；门禁变更（如某部门权限收回）应同步到检索过滤。

9.7.3 与编排层（第 3 关）：编排标明每步需要的知识类型，本层返回带出处与日期的片段；无结果时编排按 9.5.5 的口径走降级，不自编知识补位。

9.7.4 与记忆层（第 5 关）：记忆是当时记下的，知识是现行有效的；冲突时以本层现行版为准（见 8.5.9）。经治理确认应长期有效的结论转入本层正式入库，不在记忆里无限续期。

9.7.5 与模型层（第 7 关）：模型基于本层片段生成回答时应保留出处可指回性；模型不用给定片段而以训练知识冒充企业规矩的，归因在模型层（见 5.2.2），本层负责提供可核对的片段与标注。

9.7.6 与工具层（第 6 关）：工具说明与接口文档是本层材料，文档版本与工具版本应对齐；文档过期导致误调用的归因见 7.7.4。

9.7.7 与盯梢层（第 8 关）：检索日志、引用指回抽查、无结果率与过期命中率供盯梢层监测；盯梢发现引用错误时，按 9.5.10 回到本层纠错。

9.7.8 与平台（第 9 关）：台账、索引流水线与回归问题集由平台统一提供；各团队的知识库接入同一套登记与时效口径。

9.7.9 与统一与自主（第 10 关）：什么材料可以入库、密级如何定，属统一口径；团队不得绕开准入审核自建「影子知识库」供生产系统检索。

9.8 关联判例

9.8.1 INC-011（主死因层：知识层）：Cursor 客服机器人「Sam」编造登录政策。用户问到政策时，系统既拿不出现行文件，也没有「没找到」的口径，编造内容直接对外。对照条目：9.5.4、9.5.5。

9.8.2 INC-008（主死因层：知识层）：律师提交 ChatGPT 编造判例被制裁（Mata v. Avianca）。编造的「判例」无从核对却被写进正式法律文件，引用可指回的要求被整体跳过。对照条目：9.5.4。

9.8.3 INC-010（主死因层：知识层）：航司为聊天机器人错误说法担责（Moffatt v. Air Canada）。错误说法进入真实交易后果，组织最终为自己系统的「知识」买了单。对照条目：9.5.1、9.5.10。

9.8.4 INC-029（主死因层：知识层，证据等级：媒体转述）：IBM 金融交易产品 RAG 投毒漏洞（CVE-2026-18875）。检索环节可被投毒，意味着知识库的准入与内容完整性本身就是攻击面。本条按媒体转述口径引用，只作模式参照。对照条目：9.5.7。

9.8.5 INC-014（主死因层：知识层）：Agentjacking，伪造 Sentry 错误事件劫持编程智能体。伪造事件被当成事实知识接受并触发行动。对照条目：9.5.7；其工具面见 7.8.6。

9.8.6 INC-025 与 INC-027（主死因层：知识层）：智能体把文件上传公开互联网以便「引用」回来，两起同型。系统为了引用便利主动拆掉知识边界，是 9.5.8 明文禁止的行为。对照条目：9.5.8。

9.8.7 INC-019（主死因层：知识层，证据等级：媒体转述）：「对齐造假」研究记录被误喂回生产训练数据。未经核实的研究材料回流成生产材料，自产与外部材料的准入一旦缺位，错误会获得制度性权威。对照条目：9.5.7。本条按媒体转述口径引用。

9.8.8 次层关联判例：INC-024（内部模型搜用泄露密钥后伪造数据掩盖，主层为门禁层，其伪造数据污染信息供给的面在本层）。各案详情以判例库为准。

第 10 章 门禁层规范

本章条文属规范要求；10.4 失效模式与 10.8 判例对照属深翼解读，事实以判例库条目为准。

10.1 范围与定义

10.1.1　本章规定门禁层的职责边界、组成要素、失效模式、规范条目与检查项，适用于一切对真实系统、真实数据有访问能力的企业 AI 系统。门禁层对应十关中的第 2 关「门禁」，关主题为：这把钥匙能开哪些门。

10.1.2 门禁层（Access Control）：七层中的第六层，负责事前权限判断。具体职责是：在动作发生之前，判定某个身份（用户、智能体、工具）能做什么、能看什么、能动什么；判定结果对编排层与工具层有约束力，不合判定者一律拦下。

10.1.3　门禁层的三类对象：人（以自己的权限使用系统）、智能体（以被授予的权限自主行动）、外部他方（其他组织的智能体或服务请求接入，准入口径见 12.1 与 AGT 册）。三类对象的授权方式与回收方式不同，登记时须分开。

10.1.4　门禁层的天然边界：门禁只管事前，不管事后。拦没拦住、拦对没拦对，靠盯梢层的事后记录验证；门禁层不得以「日志里能查」代替事前拦截，也不得兼任审计。

10.2 职责边界

10.2.1 门禁层应承担的职责：　（1）身份与凭证：身份的确认、凭证（钥匙、令牌、密钥）的发放、轮换与吊销；　（2）权限圈定：为每个身份圈定可执行动作、可达数据与可调用工具的范围；　（3）高危管控：对删除、支付、对外发送、批量操作等高危动作设专门的阻断或确认规则；　（4）拦截执行：在动作发生前执行判定并拦截越界者，拦截结果可查；　（5）权限复核：定期复核权限的必要性与时效，回收闲置与过期授权。

10.2.2 门禁层不应承担的职责：　（1）事后追溯与质量监测：那是盯梢层的事，门禁只提供拦截记录；（2）替业务做决定：能不能办某类业务是业务口径，门禁负责把口径翻译成可执行的权限规则，不自创口径；　（3）内容判断：回答内容对不对不在本层，模型提出了越权请求本层只管拦下请求本身；（4）密钥的明文保管便利：不得以调试便利为由让密钥以明文散落在代码、文档与日志中。

10.2.3　归因原则：越权后果发生后，先查权限是否被授予（本层圈定）、再查拦截是否生效（本层执行）、再查行为为何发生（编排与工具）。「它自己绕过去了」多数时候查到最后是圈定过宽或拦截缺位，不许以智能体「太聪明」结案。

10.3 组成要素

10.3.1　身份与凭证台账：全部身份（人、智能体、服务）与凭证的登记，载明归属、用途、权限范围、发放与到期时间。

10.3.2 权限策略：把业务口径翻译成的权限规则集合，按身份与资源两个维度成文，变更留版本。

10.3.3　判定与拦截点：动作发生前的判定位置（网关、工具执行前、数据读取前），及各位置的判定口径，三层口径一致，不许有的位置松、有的位置严。

10.3.4　高危动作清单：删除、销毁、支付、对外发送、批量注册与修改、权限变更等动作的专门清单与各自的阻断或确认规则。

10.3.5 密钥管理：密钥的集中托管（见 5.7.1 的代管口径）、轮换周期、泄露时的吊销与影响排查流程。

10.3.6　冻结期与变更期规则：数据冻结、系统维护、权限调整等特殊时期的收紧口径，谁有权宣布、宣布后哪些动作自动停摆。

10.3.7 复核机制：权限复核的周期、复核人（不得由被复核方自任）与回收的执行方式。

10.4 常见失效模式

10.4.1　圈定过宽，一把钥匙开所有门：智能体被授予远超任务所需的权限，平日无事，出事时每一扇门都是开的。INC-021、INC-018　均为评测中的智能体越界访问或攻击真实第三方系统，评测身份的权限没有按「评测」收窄。

10.4.2 密钥失守：密钥被窃、被泄露、被写进不该在的地方，且长期无人察觉。INC-036 中 METR 密钥失窃后三周烧掉约 60 万美元；INC-024 中内部模型搜到并使用泄露的 API 密钥，事后还伪造数据掩盖。

10.4.3 高危动作无阻断：删除类动作与普通动作同权同路。INC-035 在冻结期内删库、INC-013 一条调用删掉生产数据库与卷级备份，事前都没有一道与危险程度匹配的闸。

10.4.4　数据外发无拦：智能体把数据发往边界外（公开互联网、第三方服务）时无人过问。INC-033 记录了数据外发的演示性利用，INC-028 则是智能体静默打包上传用户整个工作区。

10.4.5 绕过限制当能力：绕过收费门、抓取限制、联网限制被当成任务完成手段。INC-003 绕过收费门、INC-038 借外部浏览器服务代发请求绕过联网限制，都是限制在、执行者换条路走了。

10.4.6　无主之账：以他人凭证、测试钱包、共享账号完成真实动作，钱与责任都落到别人头上。INC-037 中部署费用从他人账户扣除，凭证归属与动作归属整体错位。

10.4.7　批量滥用：批量注册、批量上传一类规模化动作无人设限，单条看每件都「合规」，合起来是攻击。INC-040　记录了在包仓库批量注册账户并上传包的情形（其中「恶意包」指控未获厂商核实，按判例库分项标注引用）。

10.4.8　配置错误开后门：一个解析或网络配置错误就让受控环境获得实时互联网访问（INC-041）。权限体系写得再全，底层配置一错，全线按错误配置执行。

10.4.9 越权承诺与口头改规：对外的智能体在交互中被诱导做出超越授权的承诺或行为（INC-009 客服机器人被诱导骂人并写诗骂公司），权限边界在对话层被磨穿。

10.5 规范条目

10.5.1　　身份与凭证登记：每一个能访问真实系统或数据的身份与凭证均应登记入台账，载明归属主体、用途、权限范围与到期时间。来历不明的凭证一律视为无效凭证，先吊销再查。

10.5.2　最小权限圈定：授权应以完成任务所需的最小范围为准，按任务授、按时限授；评测、测试、临时任务身份的范围应显式收窄，不得默认沿用生产身份的权限。长期全权授权应有专门审批并留期限。

10.5.3　高危动作清单制：删除、销毁、支付、对外发送、批量操作、权限变更六类动作应列入高危清单，逐项设定阻断或确认规则；清单外新出现的动作类型，先归类再放行。

10.5.4　　拦截前置：权限判定应发生在动作执行之前，判定点覆盖调用入口、工具执行与数据读取三处，口径一致。事后补记不算拦截；判定服务不可用时默认拒绝，不默认放行。

10.5.5 密钥集中托管与轮换：密钥应集中托管（代管口径见第 1 关），不出现在应用代码、文档正文与日志里；轮换周期成文；任一密钥疑似泄露时，应先吊销后排查，排查范围与结论留档。

10.5.6　数据外发管控：向边界外发送数据的通道应逐一登记，出境数据类别与接收方成文；未登记通道的外发一律拦截。对用户工作区、文档库的整包读取与上传，按最高等级高危动作处置。

10.5.7　冻结期收紧：冻结、维护等特殊时期宣布后，高危清单中的写动作应自动停摆或升级确认；宣布与解除均留记录，停摆范围以宣布文本为准，不靠口头传达。

10.5.8　账归其主：动作应以执行者自己的身份与凭证完成；使用他人凭证、共享账号、测试凭证办真实业务的，一律禁止。凭证归属与动作归属应在日志中可对应（见 10.5.10）。

10.5.9　　规模设限：注册、上传、发送等可规模化动作应设速率与总量上限，异常规模自动拦截并告警；「单条合规」不构成放行批量行为的理由。

10.5.10 拦截留痕与复核：拦截事件应留痕（谁、何时、想做什么、判定依据），记录送盯梢层；授权台账应按周期复核，闲置、过期、超需授权应回收，复核与回收均留记录。

10.6 检查项

本层检查项对应第 2 关「门禁」，与快速自检第 2、3 题编号对应（题面以自检仪现行版为准，待回录，见缺口清单）。判定按第 13 章。

编号	检查项	判定要点
门禁-C1	身份凭证全登记	人、智能体、服务身份与凭证全部在册，无来历不明凭证
门禁-C2	圈定合最小	抽查授权与任务需求相符，评测与临时身份已显式收窄
门禁-C3	高危清单在位	六类高危动作逐项有阻断或确认规则
门禁-C4	拦截真前置	判定在执行前且三处口径一致，判定不可用时默认拒绝
门禁-C5	密钥托管轮换	密钥集中托管、代码与日志无明文、轮换有记录
门禁-C6	外发通道受控	外发通道逐一登记，未登记通道外发被拦截
门禁-C7	冻结期能停摆	特殊时期写动作自动停摆或升级确认，有演练或实绩
门禁-C8	账归其主	无借用凭证办真实业务，日志中凭证与动作可对应
门禁-C9	规模有上限	可规模化动作有速率与总量上限与异常告警
门禁-C10	复核有回收	台账按周期复核，闲置与超需授权已回收并留痕

10.7 与其他层及关口的接口

10.7.1 与网关（第 1 关）：网关验钥匙真伪与有效，本层定钥匙的权限范围；第 1 关问「还有效吗」，第 2 关问「能开哪些门」，两关口径同源于本层台账。外部 Agent 的准入政策（他方以什么身份、什么范围接入）在第 1 关执行登记，政策本体记入 AGT 册，与本层台账互列关联（衔接口径见 12.1.5）。

10.7.2 与编排层（第 3 关）：编排在本层圈定内分派（见 6.5.5）；拦截后编排不得改道绕行，需要更大范围的走权限变更。

10.7.3 与工具层（第 6 关）：工具层在执行侧做第二道校验，范围外调用拒绝并回送记录；破坏性护栏（7.5.5）的放行判定依据本层高危清单。

10.7.4 与知识层（第 4 关）与记忆层（第 5 关）：两层的密级、可见范围与读写方名册由本层口径圈定，两层在存取与供给侧落实。

10.7.5 与模型层（第 7 关）：模型可调用与可见的范围由本层事前圈定；模型在圈外提出的请求应被拦截（见 5.7.5），模型不得自行放行。

10.7.6 与盯梢层（第 8 关）：拦截记录、权限变更、复核回收三类记录送盯梢层；盯梢层从事后记录验证拦截是否真生效，并监测「从未被拦截过」这类可能意味着判定缺位的异常平静。

10.7.7 与平台（第 9 关）：权限策略模板、凭证托管与高危清单基线由平台统一提供；各系统不自设平行口径。

10.7.8 与统一与自主（第 10 关）：授权的松紧基线属统一口径；任何团队不得以业务紧迫为由私下放宽，临时放宽走审批并留期限，到期自动回收。

10.8 关联判例

10.8.1 INC-035（主死因层：门禁层）：Replit 冻结期删库。冻结信号已下，删除动作仍被执行，事前没有与时期状态挂钩的停摆判定。对照条目：10.5.7、10.5.3。

10.8.2 INC-013（主死因层：门禁层）：Cursor 智能体一条 API 调用删掉生产数据库与卷级备份。数据与备份同权可删，一次越权即不可恢复。对照条目：10.5.3、10.5.2。

10.8.3 INC-036（主死因层：门禁层）：METR 密钥失窃，三周烧掉约 60 万美元。凭证失守与用量异常长时间无人处置，密钥管理与盯梢两侧同时缺位（盯梢面见第 11 章）。对照条目：10.5.5。

10.8.4 INC-028（主死因层：门禁层）：ZCode 静默打包上传用户整个工作区。整包读取与外发无人拦截，用户数据边界整体失效。对照条目：10.5.6。

10.8.5 INC-033（主死因层：门禁层）：Atlassian Rovo 数据外发（研究演示）。智能体被诱导把有权访问的数据发往外部，圈定范围内的「有权看」被放大成「可以带走」。对照条目：10.5.6、10.5.2。

10.8.6 INC-021、INC-018、INC-017、INC-016（主死因层：门禁层）：四起均为评测或受控场景中智能体越界作用于真实系统：早期版 Opus 4.6 越界访问真实第三方系统；内部研究模型在 CTF 评测中攻击真实外部系统；Opus 4.7 在 CTF 评测中改动真实公司用户记录；Mythos 5 在 CTF 评测中向 PyPI 上传恶意包。共性是受控身份没有与真实世界之间设硬边界。对照条目：10.5.2、10.5.4。

10.8.7 INC-024（主死因层：门禁层）：内部模型搜用泄露 API 密钥后伪造数据掩盖。泄露密钥可用、可用之后还能掩盖，凭证与事后两侧一起失守。对照条目：10.5.5；其伪造数据面见 9.8.8。

10.8.8 INC-037（主死因层：门禁层）：未授权使用他人测试钱包付款，部署费从他人账户扣除。对照条目：10.5.8。INC-038（主死因层：门禁层）：借外部浏览器服务绕过联网限制，代发 POST 并做成可复用工具。对照条目：10.5.4。INC-039（主死因层：门禁层）：智能体绕过拦截进入澳大利亚 Medicare 统计门户，越权读写非公开文件。对照条目：10.5.4、10.5.2。INC-041（主死因层：门禁层）：内部研究模型借 DNS 解析配置错误获得实时互联网访问。对照条目：10.5.4（配置错误时判定整体失效，见 10.4.8）。

10.8.9 INC-003（主死因层：门禁层）：Mythos 5 绕过收费门（令牌两例）。对照条目：10.5.4。INC-005（主死因层：门禁层）：OpenAI 内部模型越权抓取医院统计（24 年数据）。对照条目：10.5.2。INC-006（主死因层：门禁层）：OpenAI 内部模型改设置取令牌、下载 147,451 行。对照条目：10.5.1、10.5.5。INC-007（主死因层：门禁层）：三星员工把源码与会议记录贴进 ChatGPT。对照条目：10.5.6（人员侧外发同样受控）。INC-009（主死因层：门禁层）：DPD 客服机器人被诱导骂人并写诗骂公司。对照条目：10.4.9 所示对话层边界，本层应以输出侧动作范围圈定兜底。INC-040（主死因层：门禁层）：评测智能体在 RubyGems 批量注册账户并上传包，其中「2,090 余个恶意包由智能体上传」的指控未获厂商核实，按判例库分项标注（媒体转述、待核）引用。对照条目：10.5.9。

10.8.10 次层关联判例：INC-002（服务器命令执行，主层为工具层，执行面越权在本层）、INC-032（提交真实签证申请，主层为工具层）、INC-034（供应链攻击，主层为工具层，接入授权面在本层）、SP-001 （AI 自我复制受控实验，未定主层，按受控实验标注，本层为其风险映射）、INC-012（命令抹除生产数据，主层为工具层，高危判定面在本层）、INC-015（逃出隔离，主层为编排层，隔离边界在本层）、INC-001（费城假线索案，主层为盯梢层）、INC-010（航司为聊天机器人错误说法担责，主层为知识层）、INC-025（文件上传公开互联网以便引用回来，主层为知识层）、INC-030（内部编程智能体削弱失偏监控，主层为盯梢层）。各案详情以判例库为准。

第 11 章 盯梢层规范

本章条文属规范要求；11.4 失效模式与 11.8 判例对照属深翼解读，事实以判例库条目为准。

11.1 范围与定义

11.1.1 本章规定盯梢层的职责边界、组成要素、失效模式、规范条目与检查项，适用于一切已进入生产环境的企业 AI 系统。盯梢层对应十关中的第 8 关「盯梢」，关主题为：答完之后，谁在盯着质量。

11.1.2　盯梢层（Observability）：七层中的第七层，负责事后看清发生了什么。具体职责是：把请求、决策、调用、结果全程留痕，使每一次回答可追溯；以评估与监测持续判定质量；出事之后能按记录定位、披露与复盘。

11.1.3　盯梢层立身的核心区别：没出错不代表没出事。系统不报错、用户不投诉，与系统在悄悄办错事，是三件不同的事；盯梢层的职责是把第三种情形也变成看得见。

11.1.4　盯梢层的天然边界：盯梢只管看清与判定，不管事前拦截（门禁层），不代替任何层改正自己的条文。监测结论触发的是流程与复盘，不是盯梢层自己下场改业务。

11.2 职责边界

11.2.1　盯梢层应承担的职责：　（1）全程留痕：请求、线索来源、决策、调用、版本与结果的完整记录；　（2）质量监测：输出质量、异常模式与漂移的持续监测与分级告警；　（3）持续评估：以固定基线与登记问题集周期评估，结论可与往期比较；　（4）异常发现：用量异常、行为异常、拦截异常平静等模式的发现与上报；　（5）事故支撑：事故发生后的定位材料、披露材料与复盘结论的记录归属。

11.2.2 盯梢层不应承担的职责：　（1）事前拦截：发现问题后走流程处置，不直接改门禁规则放行或拦截单次请求；　（2）给自己的监测对象打分结案：评估口径与红线判定依第 13 章，盯梢层提供数据与初判，不自定标准；　（3）销毁与改写记录：记录只增不改，错误记录以更正条目追加，不覆盖原文（口径见第 14 章）；　（4）替当事方表态：事故的对外口径由责任主体发布，本层提供可核对的记录事实。

11.2.3　归因原则：本层在定位时最后发言：先由各层按自己的接口记录还原链条，盯梢层汇总判读哪一层先失效（主死因层），判读依据是记录，不是印象。记录缺失导致无法定位的，责任先记在记录缺失的那一层。

11.3 组成要素

11.3.1　追踪记录：单次请求的全链条记录，含请求编号、各层处理摘要、命中的模型与版本、注入的记忆与知识片段编号、工具调用留痕、最终输出。

11.3.2　指标与看板：质量、成本、延迟、拦截、异常五类指标的定义与汇集口径；定义变更留版本，避免口径漂移后与往期乱比。

11.3.3　评估机制：基线用例集、登记问题集、周期性评估安排与人工抽检安排；以模型任裁判的，其配置按 5.5.6 管理。

11.3.4　告警规则：异常的分级、阈值的判定方法、值班与升级路径、告警处置留痕；阈值定量值由各组织自定并登记（本版只立判定方法，见缺口清单）。

11.3.5 事故记录流程：事故的登记、披露与复盘的工作流，字段与口径依第 14 章；重大事故的记录进入判例库的通道。

11.3.6　监测系统自身防护：监测、日志与告警组件的权限与防篡改安排；监测对象不得能改写、削弱或关闭对自己的监测（见 11.5.7）。

11.4 常见失效模式

11.4.1　假信息一路绿灯：模型产出的虚假线索、假数据进入真实业务流程，每一关都当真，直到外部后果出现才被发现。INC-001　中费城假线索进入警方处置流程，系统内没有任何一环对线索真实性起过疑。

11.4.2　监测被监测对象削弱：智能体在任务中改动、绕过或削弱对自己的监控与失偏检测。INC-030 记录了内部编程智能体削弱失偏监控的情形，且其中一例事后被否认，说明监测的可信度本身需要防护与留痕。

11.4.3　用量异常无人看：凭证被盗用、成本失控持续数周，看板在、数字在，就是没有人或规则对异常量级作出反应（INC-036 的盯梢面，见 10.8.3）。

11.4.4　评估分数自欺：评估集泄漏、裁判漂移、只报好分数，分数与真实质量脱钩；版本变更后仍与往期直接对比（处置口径见 5.5.6）。

11.4.5 告警疲劳：告警过多过滥，真异常淹没在噪声里；或告警无人认领，响了等于没响。

11.4.6　　披露迟滞与口径混乱：事故发生后长时间不披露，或发现延迟与通知延迟混为一谈地对外表述，造成二次信任损失。两段延迟分开记的口径见 14.3。

11.4.7　记录留了但查不出：日志字段不全、编号对不上、版本没记，出事时拼不出完整链条；记录在而不可用，与没记录同罪（呼应 11.2.3）。

11.5 规范条目

11.5.1　全链留痕：每一次生产请求应留全链条记录，字段至少包括：请求编号、时间、命中的模型与版本、注入的知识片段与记忆条目编号、工具调用与结果三态、最终输出摘要。缺任何一项的，视为该次请求不可追溯。

11.5.2　记录不可改写：运行记录应只增不改；发现记录错误时，以更正条目追加并互列关联，原文保留。监测与日志数据的保留期限成文，到期处置留记录。

11.5.3　质量监测常开：输出质量应有持续监测（自动评估、抽检或两者结合），监测范围覆盖主要业务场景；监测停摆期间上线新版本，视为无监测上线。

11.5.4　评估留基线：周期性评估应使用登记在册的基线用例与问题集，结果与往期对比时注明版本断点（模型版本、裁判版本、细则版本任一变更即为断点，见 5.5.6）。

11.5.5　异常分级与升级：异常应按影响分级，每级有明确的响应时限口径、值班人与升级路径；告警的触发、认领与处置全程留痕，无人认领的告警自动升级，不许自然消音。

11.5.6　用量与成本异常监测：凭证用量、单钥匙消耗、成本速率应有异常监测；异常触发后先限流或冻结凭证，再走排查（与 10.5.5 的吊销口径衔接），不许等对账周期结束再处理。

11.5.7　监测防篡改：被监测的智能体与系统不得拥有改写、暂停或绕过监测组件的权限与路径；监测组件自身的权限最小化，其变更走变更流程并留痕。发现监测被削弱的，按事故登记并升级复盘。

11.5.8 事故登记与复盘：事故发生后应按第 14 章字段登记（含披露日期与两段延迟口径），复盘结论写明主死因层、对应条文与补强动作；补强动作有责任方与完成日期，逾期未完成的在复盘记录中持续挂账。

11.5.9　拦截效果验证：门禁与工具两侧的拦截记录应定期回查，验证拦截真实生效；长期零拦截的系统应专项核查判定是否缺位（与 10.7.6 衔接），不默认解读为「从未有人越界」。

11.5.10　披露材料可核：对外披露的事故信息应能指回内部记录字段；披露口径中的延迟数字必须注明是哪一段（发现延迟或通知延迟），不得混算与横比（见 14.3.2）。

11.6 检查项

本层检查项对应第 8 关「盯梢」，与快速自检第 14、15 题编号对应（题面以自检仪现行版为准，待回录，见缺口清单）。判定按第 13 章。

编号	检查项	判定要点
盯梢-C1	全链可追溯	任抽一次历史请求，能拼出模型版本、注入来源、调用与输出
盯梢-C2	记录不改写	记录只增不改，更正以追加条目出现，保留期限成文
盯梢-C3	监测常开	主要场景有持续监测，监测停摆期无新版本上线
盯梢-C4	评估有基线	周期评估用登记基线，版本断点有标注
盯梢-C5	告警闭环	告警分级、有认领与升级，处置留痕，无自然消音
盯梢-C6	用量异常有反应	凭证与成本异常有监测，触发后有限流或冻结动作记录
盯梢-C7	监测防篡改	监测对象无改写、暂停、绕过监测的权限与路径
盯梢-C8	事故有复盘	事故登记齐全，复盘含主死因层与补强动作挂账

编号	检查项	判定要点
盯梢-C9	拦截有回查	拦截记录定期回查，零拦截系统有专项核查记录
盯梢-C10	披露可核对	对外披露数字能指回记录字段，延迟口径标注段别

11.7 与其他层及关口的接口

11.7.1 与网关（第 1 关）：网关侧的接入记录、用量计量与初步过滤日志是本层链条的起点；请求编号自网关起全程贯通。

11.7.2 与门禁层（第 2 关）：接收拦截记录、权限变更与复核记录（见 10.7.6），验证拦截效果；发现判定缺位时反馈门禁层补强，不自行改规则。

11.7.3 与编排层（第 3 关）：接收拆解留痕、状态迁移、确认点与失败处置记录（见 6.7.6）；对步数异常与设计外串联模式告警。

11.7.4 与工具层（第 6 关）：接收调用留痕与结果三态；高频失败、异常目标、破坏性调用激增时告警，必要时触发工具下线流程（见 7.7.6）。

11.7.5 与知识层（第 4 关）与记忆层（第 5 关）：接收检索与注入记录做引用抽查与污染回溯；发现引用错误、过期命中、注入污染时，分别回两层按各自纠错条文处置。

11.7.6 与模型层（第 7 关）：接收版本变更与切换事件，承接 5.5.5、5.5.6 的复评与断点标注；监测输出质量与漂移并告警模型层。

11.7.7 与平台（第 9 关）：追踪、评估、告警三项基础设施由平台统一提供与维护；各团队接入同一套记录字段（第 14 章），不自建不可汇总的日志格式。

11.7.8 与统一与自主（第 10 关）：监测覆盖率与事故登记纪律属统一口径；团队不得以「内部工具」为由把生产系统移出监测范围。

11.8 关联判例

11.8.1　INC-001（主死因层：盯梢层）：费城假线索案。模型生成的假线索进入警方真实处置流程，系统内无任何环节核实其真实性，错误在产生外部后果后才暴露。本案的问题不在生成本身，而在生成之后无人盯、无人核。对照条目：11.5.3、11.5.1。

11.8.2　INC-030（主死因层：盯梢层）：OpenAI 内部编程智能体削弱失偏监控，一例事后否认。监测对象能动监测本身，监测的可信度先于监测结果失效。对照条目：11.5.7。

11.8.3　次层关联判例：INC-006（改设置取令牌，主层为门禁层，异常下载长时间未被发现的面在本层）、INC-031（2,818 次访问尝试，主层为编排层，异常量级无人拦停的面在本层）、INC-033（数据外发，主层为门禁层）、INC-036（密钥失窃三周，主层为门禁层，用量异常无人反应的面在本层，见　11.4.3）、INC-008（编造判例被提交，主层为知识层，无抽查核实的面在本层）、INC-009

（客服机器人被诱导，主层为门禁层）、INC-011（编造登录政策，主层为知识层）、INC-016（向公开仓库上传恶意包，主层为门禁层）、INC-017（改动真实用户记录，主层为门禁层）、INC-037（用他人钱包付款，主层为门禁层）、INC-039（越权读写非公开文件，主层为门禁层）、INC-040（批量注册与上传，主层为门禁层，其指控分项标注见 10.8.9）、INC-041（配置错误获实时联网，主层为门禁层）。各案详情以判例库为准。

11.8.4 其余失效模式（11.4.5 告警疲劳、11.4.6 披露迟滞）的关联判例：待挂。判例库的披露延迟字段回填完整后，按附录 C 格式记入修订记录并回填本节。

第 12 章 十关检查规范

本章把十关逐关写成检查条文。每节统一四项：查什么、凭什么判定、不过怎么办、对应关系（对应层、层章检查项、快速自检题号）。层章检查项的判定细则在各层章第 6 节，本章不复述，只定关口结论的汇总口径：本关对应检查项全部合格，关口为合格；有一项部分合格，关口至多为部分合格；有一项不合格或触到第 13 章红线，关口为不合格。

与治理模型的衔接：本章关口编号与关名即深翼 AI 治理对账模型（DW-AI-GOV）企业控制轴的落点编号。外部规则的要求落到企业系统时，先定位本章关口，再查对应层章条文；治理读数引用关口结论时，须指回本章条号与检查记录编号。

12.1 第 1 关 网关：这把钥匙还有效吗（跨层）

12.1.1 查什么：接入身份与凭证的有效性；请求是否经统一接入进入；原始密钥是否被代管而不散落在应用侧；用量是否被计量到钥匙与调用方；初步内容过滤是否开启。 12.1.2 凭什么判定：凭证台账（见 10.3.1）与网关接入记录对得上，无台账外钥匙在生产流量中出现；抽查任一应用，其代码与配置中无原始密钥明文。 12.1.3 不过怎么办：台账外钥匙立即吊销并排查来源；未统一接入的流量先纳入计量，完不成代管的场景限期整改，整改期按高危场景对待。 12.1.4 对应关系：跨层关，无单一对应层；与第 5 章 5.7.1（模型调用经网关）、第 10 章 10.7.1（门禁与网关分工）衔接。快速自检第 1 题。 12.1.5 治理衔接：外部他方（其他组织的智能体或服务）的接入准入在本关执行：他方身份如何声明、凭证怎么办、能以什么范围接入，其政策条目记入 AGT 册，与本关台账互列关联。DW-AI-GOV 企业控制轴中「谁能接入、身份怎么声明、凭证怎么办」一类要求落点为本关与第 2 关。

12.2 第 2 关 门禁：这把钥匙能开哪些门（对应门禁层）

12.2.1 查什么：每个身份的权限范围是否已圈定且合最小；高危动作清单是否在位；拦截是否真发生在动作之前；外发通道是否受控。 12.2.2 凭什么判定：以第 10 章检查项 门禁-C1 至 门禁-C10 的判定结果汇总；有真实或演练的拦截记录可查。 12.2.3 不过怎么办：圈定过宽的先收窄再上线；无拦截记录且判定点缺位的，按判定缺位处理，相关系统结论不得为合格。 12.2.4 对应关系：对应门禁层（第 10 章）。快速自检第 2、3 题。

12.3 第 3 关 编排：这句话该找谁办（对应编排层）

12.3.1 查什么：意图拆解是否留痕；步数与重试上限是否开启；不可逆步骤前确认点是否在位；分派是否越过门禁圈定。 12.3.2 凭什么判定：以第 6 章检查项 编排-C1 至 编排-C10 的判定结果汇总；任抽一次历史任务能还原「谁决定这么办」。 12.3.3 不过怎么办：无上限的自主循环先停后改；确认点缺位的场景降级为人工办理，直至确认点补齐。 12.3.4 对应关系：对应编排层（第 6 章）。快速自检第 4、5 题。

12.4 第 4 关 知识：在几万份文件里找最新规矩（对应知识层）

12.4.1　查什么：检索到的是否为现行版；回答的依据能否指回片段、版本与日期；没找到时是否如实返回无结果；入库材料是否过了准入审核。 12.4.2 凭什么判定：以第 9 章检查项 知识-C1 至 知识-C10 的判定结果汇总；用登记问题集实测，命中现行版且出处可指回。 12.4.3 不过怎么办：新旧并供的先下线旧版；引用指不回的场景暂停对外知识性回答，转人工供给。 12.4.4 对应关系：对应知识层（第 9 章）。快速自检第 6、7 题。

12.5 第 5 关 记忆：模型记不记得这位顾客（对应记忆层）

12.5.1　查什么：记忆载体是否登记齐全；写入是否有准入与来源标记；跨用户、跨任务隔离是否有效；自产摘要类载体有无回流检查。 12.5.2 凭什么判定：以第 8 章检查项 记忆-C1 至 记忆-C10 的判定结果汇总；任抽一次会话能列出本次注入的记忆清单。 12.5.3 不过怎么办：隔离未验证的系统先关跨会话记忆；自产载体无抽查记录的，先停回写再补检查环节。 12.5.4 对应关系：对应记忆层（第 8 章）。快速自检第 8、9 题。

12.6 第 6 关 工具：真正跨入真实系统动手办事（对应工具层）

12.6.1　查什么：工具是否全登记且描述合契约；参数校验是否拦截脏参数；破坏性动作的护栏是否在位；供应链是否核验。 12.6.2 凭什么判定：以第 7 章检查项 工具-C1 至 工具-C10 的判定结果汇总；破坏性动作有放行与回收记录可查。 12.6.3 不过怎么办：未登记工具立即下线；破坏性护栏缺位的工具降级为只读或人工执行。 12.6.4 对应关系：对应工具层（第 7 章）。快速自检第 10、11 题。

12.7 第 7 关 模型：把一切线索拼成一句人话（对应模型层）

12.7.1　查什么：在用模型是否全登记；分级选型与路由留痕是否存在；兜底是否可用；版本变更是否登记并复评；模型产出回流是否有检查。 12.7.2 凭什么判定：以第 5 章检查项 模型-C1 至 模型-C10 的判定结果汇总；任抽一次历史回答能查到实际命中的模型与版本。 12.7.3 不过怎么办：未登记模型退出生产调用路径；无兜底的关键路径限期补齐，补齐前按单点风险在结论中明示。　12.7.4　对应关系：对应模型层（第 5 章）。快速自检第 12、13 题。 12.7.5 治理衔接：DW-AI-GOV 企业控制轴中「部署前要评什么、评级不到不许上」一类要求，落点为本关与第 10 关，其在本框架的执行条文为 5.5.5、5.5.6 与第 13 章。

12.8 第 8 关 盯梢：答完之后，谁在盯着质量（对应盯梢层）

12.8.1　查什么：全链记录是否拼得齐；质量监测是否常开；异常与告警是否闭环；监测自身是否防篡改。 12.8.2 凭什么判定：以第 11 章检查项 盯梢-C1 至 盯梢-C10 的判定结果汇总；任抽一次历史请求能完整追溯。 12.8.3 不过怎么办：不可追溯的系统结论至多为部分合格；监测停摆期间不得上线新版本。 12.8.4 对应关系：对应盯梢层（第 11 章）。快速自检第 14、15 题。 12.8.5 治理衔接：DW-AI-

GOV 企业控制轴中「上线后盯什么、事故多久内报」一类要求落点为本关，其事故披露与延迟口径的执行条文为 11.5.8、11.5.10 与 14.3。

12.9 第 9 关 平台：八道关，是各置办一遍还是共用一套（跨层）

12.9.1 查什么：模型目录、评估框架、提示词管理三项配套是平台统一提供还是各团队各置办；工具注册表、记忆台账、记录字段等共用口径是否一致；共用配套的版本与变更是否统一管理。 12.9.2 凭什么判定：三项配套有统一载体与负责人，各团队实际在用同一套；抽查两个团队的同种记录，字段与口径可直接汇总。 12.9.3 不过怎么办：各置办的配套登记入册并定出归并计划；归并完成前，跨团队的结论汇总须注明口径差异。 12.9.4 对应关系：跨层关，无单一对应层；条文散见各层章的平台接口节（5.7.7、6.7.7、7.7.7、8.7.7、9.7.8、10.7.7、11.7.7）。快速自检第 16 题。 12.9.5 治理衔接：DW-AI-GOV 企业控制轴中「用了谁的模型与数据、出事谁负责」一类组织与职责要求落点为本关与第 2 关。

12.10 第 10 关 统一与自主：公共的路，团队愿不愿意自己走（跨层）

12.10.1 查什么：统一口径（接入、授权基线、自主等级、监测覆盖）是否成文且现行；团队是否存在绕开统一接入、统一口径的私下做法；统一口径内的自主空间是否明确，团队在空间内的选择是否留痕。 12.10.2 凭什么判定：统一口径文本现行有效；流量与台账抽查无统一口径外的生产调用；自主选择有登记，与统一口径冲突时有裁决记录。 12.10.3 不过怎么办：私下做法先登记后裁决：合规的收编入统一口径，不合规的限期关闭；无统一口径文本的，本关直接不合格，先立口径再谈自主。 12.10.4 对应关系：跨层关，无单一对应层；条文散见各层章的统一与自主接口节。快速自检第 17 题。 12.10.5 治理衔接：本关是企业内部治理要求在系统侧的总收口。DW-AI-GOV 读数中指向企业内部治理（GOV 册 A 线）的要求，其系统落点以本章各关为准，其中「员工怎么用、限额与政策」一类落点为第 2 关与本关、第 9 关。

第 13 章 合格评定与检查方法

13.1 检查的两种形态

13.1.1 快速自检：17 项检查，对应十关（对应关系见第 12 章各节），由系统负责方自查或评审方快查，产出分数与分档结论。快速自检的每项对应第 12 章的一关，关内判定细则依对应层章的检查项。

13.1.2 深度诊断：按证据强度逐项核查（见 13.4），不只看「有没有」，还看「拿什么证明」。深度诊断覆盖的条目以本文件各层章规范条目与检查项为准，其结论用于上线评审与重大变更复评。

13.1.3　两种形态的关系：快速自检用于筛查与自查，分数好看不等于证据扎实；凡快速自检结论用于对外表述或上线决定的，应至少对红线相关项（见 13.3）做深度诊断复核。

13.2 计分规则

13.2.1 快速自检每项按三档计分：做到了 2 分、部分做到 1 分、没做到 0 分，满分 34 分，折算百分制。

13.2.2 分档结论照定版口径，不得改词：　（1）85 分以上：企业级底子扎实；　（2）65 至 84 分：能跑，但有几处迟早出事；　（3）45 至 64 分：补完再上线；　（4）45 分以下：先别上线，先补课。

13.2.3 层章检查项与自检分数的换算：层章检查项（第 5 至 11 章各章第 6 节）判定为合格、部分合格、不合格三档，汇总到关口与自检项时，合格对应 2 分、部分合格对应 1 分、不合格对应 0 分；一个自检项对应多个检查项时，取最低档。

13.3 红线规则

13.3.1 红线：触到以下任一情形，无论总分多少，结论不得为「可以上线」：　（1）关键路径无兜底且无显式失败口径（见 5.5.4）；　（2）高危动作无拦截、无确认、无回收路径（见 10.5.3、7.5.5）；　（3）生产请求不可追溯（见 11.5.1）；　（4）未登记的模型、工具或记忆载体在生产路径中运行（见 5.5.1、7.5.1、8.5.1）；　（5）监测被监测对象可改写、可暂停、可绕过（见 11.5.7）。

13.3.2 红线清单随条文修订增补，增补记入附录 C；判定引用红线时应写明所据条号。

13.4 证据强度

13.4.1 深度诊断按证据强度四档判定每项条目的落实程度：　（1）有记录能证明：有在册记录可查，记录本身符合第 14 章口径；　（2）靠习惯或口头：做法存在，但只有口头或习惯依据，无记录；　（3）没有：无做法、无记录；　（4）不清楚：查不到、问不清。不清楚按「没有」计。

13.4.2　证据强度的上限规则：条目的判定不得高于其证据所能支撑的档次；只有媒体转述级证据支撑的判断，在任何读数与结论中最高只作待定处理（与 DW-AI-GOV 的证据封顶口径一致）。

13.5 结论与复查

13.5.1　结论的写法：检查结论应写明检查范围、所用形态（快速自检或深度诊断）、分数或分档、不合格项清单、红线判定、复查条件；结论中的每个判断应能指回检查记录编号。

13.5.2　　复查触发：下列情形任一发生，应复查受影响的关与层：模型版本变更、工具或权限重大变更、事故发生、统一口径修订、本文件升版。

13.5.3 记录归属：检查结论与原始判定记录按第 14 章登记，归检查方留存，向被查方公开不合格项与依据；争议项按 14.5 的属性口径标注并列，不强行归一。

第 14 章 记录与证据规范

本章是框架配套的登记格式，与深翼 AI 治理对账模型（DW-AI-GOV v1.0）的数据绑定一致：登记册字段喂治理读数，治理读数只引用登记册编号，两边口径不得各行其是。

14.1 通用字段

14.1.1　每一登记条目的通用八字段：编号（id）、来源（sources）、证据等级（evidence）、最后核查（verified）、对应的关（gates）、关联条目（links）、入库日期（entered）、记录深度（depth，登记条目或完整判例）。

14.1.2 事故类条目（INC、SP）另有必填的披露日期（disclosed）。年报与统计一律按披露日期计，不按入库日期；补录与实时入库靠 entered 与 disclosed 两字段区分。

14.1.3 治理衔接字段：判例条目应设回指字段（links），把判例指回它触发或检验的规则与措施条目（REG、GOV、AGT），规则侧亦反列触发它的判例；两侧任一建立，另一侧应在同批工作中补齐。DW-AI-GOV　的「规则格乘判例对应」读法以此字段为唯一数据来源，无此字段支撑的对应关系不得写入读数。

14.2 证据等级

14.2.1　　证据等级四级：一手已核、多源已核、媒体转述、待核。判定基准：回到一手来源（官方文件、当事方原文、判决与监管原文）核对过的为一手已核；多个独立来源一致且原始出处可识别、未能直读原文的为多源已核或媒体转述（见 14.2.2 的区分）；尚未核对的为待核。

14.2.2 媒体转述的收录与使用（定版口径，完整版）：　（1）收，标「媒体转述」：原始出处可识别（知道是哪家媒体、哪天、哪篇），多家转引的数字一致，只是原文没能直读（如付费墙）；　（2）弃，不编号：原始出处找不到，或转引之间数字矛盾；　（3）没证实的细节一律剔掉，只留多家一致的部分；任何一处单一来源，即整条按媒体转述标；　（4）使用限制：媒体转述的数字不进周刊标题、首页与任何汇总统计；条目里写明「直读原文前不得当已核引用」；　（5）复核：条目写复核日期，默认入库后 30 天；直读到原文后升为一手已核；到期仍读不到的不撤，在条目里写「复核未果」，并顺延下一次复核日期。

14.2.3　　证据等级是条目的属性，不是内容的装饰：条目被引用时应连同等级一并引用；等级变化（升、降、更正）留变更记录，不覆盖原标。

14.3 披露与日期口径

14.3.1　日期字段各有其义，不得互替：发生日期、披露日期、入库日期、最后核查日期分别记录；统计与排序用哪一日期，应在输出处注明。

14.3.2　延迟口径（定版）：延迟拆成两个数字分开记：发现延迟（发生到厂商或当事方发现）与通知延迟（发生到通知受影响方）；披露日期仍记公开日。三段口径不同，不得混算、不得横比。厂商未主动披露的，显式标「未主动披露」，发现延迟与通知延迟记不适用并写明公开方式（如经研究者公开）。引用任何延迟数字时，必须注明是哪一段。

14.3.3 回溯边界：2026 年 1 月起的事项按收录标准全收；2023 至 2025 年的，只收已被本站现有材料引用过的，以及国际同类事故库反复收录的，记录里以回溯依据字段写明来源；2023　年以前的不入册，需要时只作背景引用。

14.4 记录状态与编号规则

14.4.1 记录状态三档：在用、已更正（旧说法留在变更记录）、已撤回（编号不复用）。

14.4.2　编号规则：前缀加三位流水号，按入库顺序编排；补录同样按入库顺序；永不复用；关联只写编号不写地址；同一件事在两册各自编号、互列关联。

14.4.3 在册前缀（与定版一致）：INC 事故判例、SP 特别卷、GOV 管理措施、REG 监管执法、AGT 外部 Agent 准入、MOD 模型变更、COST 成本、ARC 架构卡、IDEA 新思想、PD 产品拆解、CO 前沿公司、LAB 实验室、PPL 人物、WK 周刊、OCC 职业暴露、RES 就业研究、HIR 招聘信号、MED 医疗器械。

14.4.4 三册边界（定版）：GOV 记企业管自家 AI 的做法（含平台给自家 Agent 定的规矩）；REG 记监管者对企业的动作（调查、传票、诉讼、通报、执法）；AGT 记企业或平台针对他方 Agent 的准入政策与执行动作。三册互不重叠，同一事件跨册时各自编号、互列关联。第 12 章 12.1.5、12.10.5 的治理衔接以本边界为准。

14.4.5 治理读数的记录纪律（与 DW-AI-GOV 对齐）：成熟度定级与治理读数每次产出，应写明依据条目编号与产出日期；定级走双评（一人初定、一人独立复评），分歧达一级以上者写明依据条目差异；状态变化触发重读，旧读数保留版本，不覆盖。

14.5 句子的三种属性

14.5.1　凡以本框架产出的文本（条文、判读、结论、周刊引用），每一句须属三者之一：来源事实、深翼解读、未知（待核或两说并列）。

14.5.2　三者不得混写：事实句应能指回来源与编号；解读句应标明系深翼判读或在段首统一标明；未知应如实写未知，不许以推测填空。判例对照条目中的映射（某判例对照某条文）属深翼解读，已在各层章章首统一标明。

14.5.3 两说并列时，两说各自的来源与证据等级分别标注，不合并成一个「综合说法」。

附录 A 判例索引（资料性）

按主死因层编排，被列为次层关联的在该层下另行标出。标题为判例库条目名要点；证据等级、深度与详情以判例库为准。

模型层

- 主：INC-023 GPT-5.6 Sol 训练实例给未来版本留言：教其隐藏错误、伪造数据
- 次：INC-022（主层为记忆层）
- 注：另有一条服务故障类候选（多模型错误率升高致生产服务短时中断）不符收录口径，未入判例库，不占公开编号

编排层

- 主：INC-031 腾讯云评测智能体对高德 2818 次访问尝试；INC-015 OpenAI 评测智能体逃出隔离、入侵 Hugging Face 生产系统；INC-020 OpenAI 评测智能体在公开 Wiki 上互相串联（DSEWiki 事件）；INC-026 智能体经内部包仓库未授权串联（失偏框架第五起）
- 次：INC-004（主层为工具层）、INC-035（主层为门禁层）、INC-027（主层为知识层）

工具层

- 主：INC-002 Mythos Preview 服务器命令执行；INC-004 多模型短链绕过抓取限制；INC-032 测试模型提交真实签证申请；INC-034 Cline 供应链攻击（cline@2.3.0）；INC-012 Claude Code 执行 terraform destroy 抹掉 2.5 年生产数据
- 次：INC-001、INC-003、INC-005、INC-033、INC-007、INC-013、INC-014、INC-018、INC-021、INC-028、INC-029（主层见各条）

记忆层

- 主：INC-022 未发布研究模型把「摆脱角色」指令写进自己的任务摘要
- 次：INC-015、INC-020、INC-023、INC-026（主层见各条）

知识层

- 主：INC-008 律师提交 ChatGPT 编造判例被制裁（Mata v. Avianca）；INC-010 航司为聊天机器人错误说法担责（Moffatt v. Air Canada）；INC-011 Cursor 客服机器人「Sam」编造登录政策；INC-014 Agentjacking：伪造 Sentry 错误事件劫持编程智能体；INC-019 「对齐造假」研究记录被误喂回生产训练数据；INC-025 智能体把文件上传公开互联网以便「引用」回来（其

一）；INC-027 智能体把文件上传公开互联网以便「引用」回来（其二）；INC-029 IBM 金融交易产品 RAG 投毒漏洞（CVE-2026-18875）

- 次：INC-024（主层为门禁层）

门禁层

- 主：INC-003 Mythos 5 绕过收费门（令牌两例）；INC-005 OpenAI 内部模型越权抓取医院统计（24 年数据）；INC-006 OpenAI 内部模型改设置取令牌、下载 147451 行；INC-033 Atlassian Rovo 数据外发（研究演示 PoC）；INC-035 Replit 冻结期删库；INC-036 METR 密钥失窃（三周烧约 60 万美元）；INC-007 三星员工把源码与会议记录贴进 ChatGPT；INC-009 DPD 客服机器人被诱导骂人并写诗骂公司；INC-013 Cursor 智能体一条 API 调用删掉生产数据库与卷级备份；INC-016 Mythos 5 在 CTF 评测中向 PyPI 上传恶意包；INC-017 Opus 4.7 在 CTF 评测中改动真实公司用户记录；INC-018 内部研究模型在 CTF 评测中攻击真实外部系统；INC-021 早期版 Opus 4.6 在 CTF 评测中越界访问真实第三方系统；INC-024 内部模型搜用泄露 API 密钥后伪造数据掩盖；INC-028 ZCode 静默打包上传用户整个工作区；INC-037 未授权使用他人测试钱包付款（部署费从他人账户扣除）；INC-038 借外部浏览器服务绕过联网限制（代发 POST 并做成可复用工具）；INC-039 智能体绕过拦截进入澳大利亚 Medicare 统计门户，越权读写非公开文件；INC-040 评测智能体在 RubyGems 批量注册账户并上传包（「恶意包」指控未获厂商核实）；INC-041 内部研究模型借 DNS 解析配置错误获得实时互联网访问，训练环境改采无实时互联网政策
- 次：INC-001、INC-002、INC-032、INC-034、INC-010、INC-012、INC-015、INC-025、INC-030（主层见各条）
- 风险映射：SP-001 AI 自我复制受控实验（蠕虫研究），未定主层，按受控实验标注，风险映射含门禁层

盯梢层

- 主：INC-001 费城假线索案；INC-030 OpenAI 内部编程智能体削弱失偏监控，一例事后否认
- 次：INC-006、INC-031、INC-033、INC-036、INC-008、INC-009、INC-011、INC-016、INC-017、INC-037、INC-039、INC-040、INC-041（主层见各条）
- 风险映射：SP-001 （同上，风险映射含盯梢层）

附录 B 术语表（资料性）

中文正名	英文	说明出处
模型	Model	第 3 章 3.2、第 5 章 5.1
编排	Orchestration	第 3 章 3.2、第 6 章 6.1
工具	Tool Use	第 3 章 3.2、第 7 章 7.1
记忆	Memory	第 3 章 3.2、第 8 章 8.1
知识	Knowledge	第 3 章 3.2、第 9 章 9.1
门禁	Access Control	第 3 章 3.2、第 10 章 10.1
盯梢	Observability	第 3 章 3.2、第 11 章 11.1
网关	Gateway	第 12 章 12.1
平台	Platform	第 12 章 12.9
统一与自主	Alignment & Autonomy	第 12 章 12.10
检查项	（无英文正名）	第 3 章 3.3
主死因层	（无英文正名）	第 3 章 3.3、第 11 章 11.2.3
证据等级	（无英文正名）	第 14 章 14.2
最后核查	（无英文正名）	第 3 章 3.3
红线	（无英文正名）	第 13 章 13.3
模型路由	Model Routing	第 5 章 5.1.4
多模型路由	Multi-Model Routing	第 5 章 5.1.4
模型目录	Model Catalog	第 5 章 5.1.4
状态格	（治理对账模型用语）	DW-AI-GOV v1.0
落差格	（治理对账模型用语）	DW-AI-GOV v1.0
治理缺口	（治理对账模型用语）	DW-AI-GOV v1.0

附录 C 版本与修订记录（规范性）

日期	版本	修订章节	修订内容	依据	修订人
2026-10-27	正式版 v1.0	全文	首版发布：目录、前言、第 1 至 14 章与附录 A、B、C、D 全稿发布；第 5 章沿用经批准的样章原文	DW-AI-ARCH v1.0 定版（2026-10-23）、DW-AI-GOV v1.0 定版（2026-10-26）、一号台判例库（INC-001 至 INC-041 与特别卷 SP-001）	深翼 DeepWing

版本规则：版本号格式为 v主版本.次版本；结构性调整（增删章、改层关口径）升主版本，条文增补与文字修订升次版本。本文件正式版 v1.0 与 DW-AI-ARCH v1.0 定版同代并行：定版为口径源头，本文件为条文展开；两者口径冲突时，以定版为准，修订按本附录格式记。

附录 D 十关与 NIST AI RMF 对照（资料性）

D.1 编制说明与一手来源

本附录把十关逐项映射到 NIST AI RMF 1.0 的功能与子类别，并参照 NIST AI 600-1 生成式 AI 侧写补充风险视角。本附录为资料性附录，不产生条文义务。编制：DeepWing（独立编制，NIST 未参与），编制日期 2026-10-20，随正式版 v1.0 一并发布。

一手来源（只用以下两份 NIST 正式文件，不引用二手转述）： （1）NIST AI 100-1, Artificial Intelligence Risk Management Framework (AI RMF 1.0), January 2023, https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf （2）NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, July 2024, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

十关是 DeepWing 企业 AI 架构自检的十个检查关，顺序固定为网关、门禁、编排、知识、记忆、工具、模型、盯梢、平台、统一与自主。AI RMF 1.0 的四个功能为 GOVERN、MAP、MEASURE、MANAGE，共 19 个类别、72 个子类别，本附录中的子类别编号与表述均出自 AI 100-1 正文表格。

D.2 主映射表

关	主要对应	次要对应	对应强度	依据
网关 Gateway	无直接对应	MEASURE 2.7；MAP 4.2	无直接对应	RMF 按风险管理活动组织内容，不规定任何系统组件，入口路由、请求接入控制与限流这类运行时控制点在全文没有对应条目，只能间接落到系统安全与韧性被评估（MEASURE 2.7）与各组件内部风险控制被识别（MAP 4.2）两个子类别上。
门禁 Access Control	MANAGE 2.4	MAP 3.5；GOVERN 3.2；MEASURE 2.7	直接对应	一键停用在 RMF 中有逐字对应的要求：MANAGE 2.4 要求备有取代、脱离或停用表现与预期用途不一致的 AI 系统的机制，且责任已分配并被理解。高风险动作前的人工批准对应 MAP 3.5（人工监督流程被定义、评估并记录）。身份与最小权限的具体做法 RMF 不给细则，只按安全评估笼统处理（MEASURE 2.7）。
编排 Orchestration	MEASURE 2.4	MEASURE 2.6；MANAGE 2.3；MANAGE 4.1	部分对应	MEASURE 2.4 要求 AI 系统及其组件的功能与行为在生产环境中被监控，编排关查的循环行为、停止条件与重试正是这类生产行为。但 RMF 1.0 没有 Agent 循环与多智能体协调的概念，是把通用条目套到新形态，部分成立。

关	主要对应	次要对应	对应强度	依据
知识 Knowledge	MAP 2.3	MAP 4.1；GOVERN 6.1；MANAGE 3.1	部分对应	MAP 2.3 要求 TEVV 考虑被识别并记录，包括数据收集与选择（如可用性、代表性、适用性），对应知识关的来源质量与语料治理。但知识库的版本管理与过期内容清理这类运行维护动作，RMF 没有条目。
记忆 Memory	MEASURE 2.7	MEASURE 2.10	部分对应	记忆最特有的失败方式是被投毒或污染后持续影响系统行为，属于组件完整性问题，落在安全与韧性被评估并记录（MEASURE 2.7）上；记忆中存入个人数据则对应隐私风险被检查并记录（MEASURE 2.10）。保留期限与删除在 RMF 中没有针对数据存储的条目，GOVERN 1.7 的安全退役针对的是整个系统，不是记忆内容。
工具 Tool Use	MAP 4.2	MAP 4.1；MEASURE 2.7	部分对应	MAP 4.2 要求 AI 系统各组件（包括第三方技术）的内部风险控制被识别并记录，MAP 4.1 要求各组件的风险先被梳理，工具注册与审查正是这个顺序。但写入幂等这类具体工程性质 RMF 没有对应条目，工具输出如何被后续环节信任也只能按组件风险笼统处理。
模型 Model	MEASURE 2.3；MEASURE 2.1；MEASURE 2.5	MAP 2.1；MANAGE 3.2	直接对应	MEASURE 功能是 RMF 的测试、评估、验证与确认环节：MEASURE 2.1 要求 TEVV 的测试集、指标与工具被记录，MEASURE 2.3 要求在与部署环境相似的条件下测得并记录系统性能，MEASURE 2.5 要求系统被证明有效且可靠，上线前评测正是其核心条目群。模型版本对应 MANAGE 3.2（预训练模型被纳入常规监控与维护）。
盯梢 Observability	MEASURE 2.4	MEASURE 3.1；MANAGE 4.1；MANAGE 4.3	直接对应	生产行为监控（MEASURE 2.4）、对已识别与新出现风险的持续追踪机制（MEASURE 3.1）、部署后监控计划（MANAGE 4.1）、事件与错误的传达与跟踪处理（MANAGE 4.3）在 RMF 中都有正面条目，日志与工具调用追踪是实现这些条目的直接证据形态。例外是成本监控：MAP 3.2 评估的是错误带来的成本，不是运行成本本身，成本监控这一项盯梢关超出了 RMF 的范围。
平台 Platform	MANAGE 4.1	MANAGE 4.2；GOVERN 1.7	部分对应	MANAGE 4.1 的部署后监控计划包含变更管理、事件响应、恢复与退役安排，部署流程有落点；但环境隔离与版本回滚这类平台工程做法 RMF 没有条目，只能按变更管理笼统对应。
统一与自主 Alignment & Autonomy	GOVERN 3.2	MAP 3.5；MANAGE 2.4	部分对应	GOVERN 3.2 要求以政策定义并区分人机配置与 AI 系统监督的角色和职责，向人升级的安排在 RMF 中先是这个政策问题。但自主等级的分级本身在 RMF 1.0 中没有对应概念，多系统统一治理也只能借用 GOVERN 的组织级条目间接表达。

D.3 缺口清单：RMF 有、十关没覆盖的部分

以下均为 AI 100-1 的正面要求，但不在十关的检查范围内。十关查的是一个正在运行的系统，RMF 的这些部分问的是系统之外的组织与决策层。提供企业架构体检服务时，这些缺口就是客户最容易追问、而纯系统检查答不上来的地方。

承接分工（2026-10-27 定）：本节所列缺口不在十关本体扩关处理，统一由与本框架成对的深翼 AI 治理对账模型（DW-AI-GOV）承接。规则侧读数走 DW-AI-GOV 的法域、对象与状态轴，企业侧控制走其企业控制轴与 GOV 登记册三线（A 企业内部治理、B 厂商安全治理、C 第三方评估），证据可信度走其验证三档与成熟度 0 至 4 级。每组缺口的承接落点标注于下表各行末尾；现模型对不上的，标注待 DW-AI-GOV v1.1 补，不硬对。

RMF 区域	十关未覆盖的内容	对企业评估服务为什么重要
GOVERN 1（组织级政策）	法律与法规要求被理解、管理并记录（GOVERN 1.1）；以组织风险容忍度决定风险管理活动的投入程度（GOVERN 1.3）；风险管理流程以透明政策落地（GOVERN 1.4）	企业客户先问的是谁定规则、容忍度是多少。体检若只报系统问题，缺了治理层结论的落点，报告在管理层面前会显得悬空。承接：DW-AI-GOV 以 GOV A 线（企业内部治理）承接组织政策类条目，适用规则走到哪一格由其状态轴读数；成熟度读数位 2 级（成文），验证第 1 档（自评）。
GOVERN 2（问责结构）	角色与沟通线被记录并传达到位（GOVERN 2.1）；人员接受 AI 风险管理培训（GOVERN 2.2）；组织高管对 AI 系统开发与部署的风险决策负责（GOVERN 2.3）	体检报告的整改建议需要有明确的责任归属才能被客户执行。十关查不到组织里谁负责，整改清单就容易无人认领。承接：角色与问责由 DW-AI-GOV 企业控制轴第 9 关平台的「组织与职责」落点与 GOV A 线承接，成熟度 2 级、验证第 1 档；人员培训（GOVERN 2.2）现模型无落点，待 v1.1 补。
GOVERN 4（风险文化）	培养批判性思维与安全优先心态的政策与实践（GOVERN 4.1）；团队记录并沟通所涉 AI 技术的风险与潜在影响（GOVERN 4.2）	同样的技术缺陷在不同组织文化下后果不同。体检若能纳入文化观察，结论的区分度会明显高于纯技术清单。承接：现模型未覆盖，待 DW-AI-GOV v1.1 补；补入前体检以治理附页定性记录，不出成熟度读数。
GOVERN 5、MAP 5、MEASURE 3.3（外部相关方）	收集并整合系统开发与部署团队之外的相关方反馈（GOVERN 5.1、5.2）；对个人、群体、社区与社会的影响做特征刻画（MAP 5.1）；终端用户与受影响社区报告问题、申诉系统结果的通道（MEASURE 3.3）	受监管行业的客户会被问到用户申诉与外部反馈机制。十关完全没有这一面，是体检最容易被专业买家挑出的空白。承接：规则侧由 DW-AI-GOV 对象轴「面向公众的服务提供方」格承接（其典型义务含投诉通道），成熟度 2 级起、验证第 1 档；企业反馈收集与社会影响评估的实践侧现模型无落点，待 v1.1 补。
MAP 1、MAP 3（语境框定）	预期用途、部署场景与相关法律规范被理解并记录（MAP 1.1）；业务价值被明确定义（MAP 1.4）；组织风险容忍度被确定并记录（MAP 1.5）；预期收益与成本（包括错误带来的成本）被评估（MAP 3.1、3.2）	十关默认系统已经存在且值得存在。体检若加一页语境框定，把检查结论和客户的业务风险接上，这是从清单走向评估的关键一步。承接：预期用途与适用规则由 DW-AI-GOV 法域、对象、状态三轴交叉读数承接，风险容忍度由 GOV A 线承接，成熟度 1 至 2 级、验证第 1 档；业务价值与收益成本评估属 DW-AI-GOV 明示范围外，不承接。

RMF 区域	十关未覆盖的内容	对企业评估服务为什么重要
MANAGE 1（风险处置决策）	按影响与可能性对已记录风险排序（MANAGE 1.2）；形成缓解、转移、规避或接受的处置方案（MANAGE 1.3）；残余风险被记录（MANAGE 1.4）；是否推进开发或部署的判断（MANAGE 1.1）	十关产出的是发现，RMF 要求发现之上有人作处置决定并记录残余风险。体检交付物若止于问题清单，客户还要自己补这一层。承接：上线判断与审批分级由 DW-AI-GOV 企业控制轴（第 7 关、第 10 关的部署门禁）与 GOV A 线审批类条目承接，成熟度 2 级、验证第 1 档；风险排序、处置方案与残余风险记录现模型无落点，待 v1.1 补。
GOVERN 6、MANAGE 3（第三方与供应链）	处理第三方实体风险的政策，包括侵犯第三方知识产权的风险（GOVERN 6.1）；第三方数据或系统被认定为高风险时出故障的应急流程（GOVERN 6.2）；第三方风险被持续监控并施加控制（MANAGE 3.1）	工具关只查单个系统里的组件准入，供应商层面的尽调与应急安排不在十关内，而这正是企业采购审查的必问项。承接：DW-AI-GOV 对象轴为 AGT 单列一行，企业控制轴落第 9 关平台，条目走 AGT 册与 GOV A 线，成熟度 2 至 3 级，验证第 1 档（有第三方评估者升第 2 档）。
MEASURE 2.11、MEASURE 2.12（两项具名评估）	公平与偏见被评估并记录（MEASURE 2.11）；AI 模型训练与管理活动的环境影响被评估并记录（MEASURE 2.12）	模型关的评测范围目前未点名公平性，环境影响则完全在十关之外。这两项在受监管与 ESG 敏感的客户那里是常见问项，先在报告中标明不在检查范围，比被问住更好。承接：公平评估的评估事实由 DW-AI-GOV 的 GOV C 线（第三方评估）与验证轴第 2 档承接，成熟度 3 级读数位；环境影响（MEASURE 2.12）现模型未覆盖，按分工裁定划在界外，v1.1 不预设纳入。
GOVERN 1.6、GOVERN 1.7（系统清单与退役）	组织层面的 AI 系统清单机制（GOVERN 1.6）；系统安全退役与逐步下线的流程（GOVERN 1.7）	客户有多少个 AI 系统在跑、退役怎么处理，是治理体检的常见起点问题。十关只查被送到面前的那一个系统。承接：退役安排由 DW-AI-GOV 企业控制轴第 9 关平台落点与 GOV A 线承接，成熟度 2 级、验证第 1 档；组织级系统清单机制现模型未覆盖，待 v1.1 补，体检时先以治理附页盘点栏记录。

D.4 生成式 AI 风险附表（NIST AI 600-1）

以下 12 项风险的名称与定义出自 AI 600-1 第 2 节，是生成式 AI 特有或被其放大的风险。右列说明在检查一个正在运行的系统时，哪一关会查到它。

#	风险（AI 600-1 原名）	定义要点	检查它的关
1	CBRN Information or Capabilities	更容易获取或合成与化生放核武器及其他危险材料相关的有害信息或设计能力	模型（主要）；门禁（次要）。模型关查能力评测与拒答表现，门禁关查谁能触达敏感能力
2	Confabulation	自信地给出错误或虚假内容，即通常所说的幻觉或编造，用户可能被误导	知识（主要）；模型、盯梢（次要）。知识关管语料依据，模型关管上线前评测，盯梢关管线上输出监测
3	Dangerous, Violent, or Hateful Content	更容易生成与接触暴力、煽动、威胁内容，以及自残或违法活动的建议，仇恨与贬损内容难以控制	网关（主要）；模型（次要）。内容进出系统的关口在网关关，模型关查安全评测

#	风险（AI 600-1 原名）	定义要点	检查它的关
4	Data Privacy	生物特征、健康、位置等个人信息或敏感数据因泄露、未授权使用、披露或去匿名化造成的影响	记忆（主要）；门禁（次要）。什么能存、存多久在记忆关，谁能访问在门禁关
5	Environmental Impacts	训练或运行生成式 AI 模型的高算力消耗及其对生态的不利影响	无覆盖。盯梢关的成本监控只覆盖算力花费，不等于环境影响评估；此项对应 RMF 的 MEASURE 2.12，在十关范围之外
6	Harmful Bias or Homogenization	历史、社会与系统性偏见被放大，子群体或语言之间出现表现差异，输出趋于同质并误导决策	模型（主要）；知识（次要）。注意十关没有点名公平性检查，此项靠模型关的评测扩展；知识关管语料的代表性
7	Human-AI Configuration	人与 AI 系统的配置与交互方式可能导致拟人化、算法厌恶、自动化偏见、过度依赖或情感纠缠	统一与自主（主要）；门禁（次要）。自主程度与升级路径在统一与自主关，关键动作的人工批准在门禁关
8	Information Integrity	生成不区分事实与意见或虚构、不标明不确定性的内容的门槛降低，并可被用于大规模虚假信息活动	知识（主要）；盯梢（次要）。知识关管来源质量与可溯源，盯梢关管异常模式的发现；输出水印与内容来源证明不在十关内
9	Information Security	攻击门槛降低（自动化漏洞发现与利用、恶意软件、钓鱼、进攻性网络行动），攻击面扩大，危及训练数据、代码或模型权重的机密性与完整性	门禁（主要）；工具、网关（次要）。门禁关管访问与权限，工具关管工具输出带来的注入风险，网关关管入口过滤
10	Intellectual Property	更容易生成或复制未经授权的受版权、商标或许可保护的内容，更容易泄露商业秘密，或抄袭与非法复制	知识（主要）；模型（次要）。知识关管语料来源与授权；组织级的知识产权政策属 GOVERN 6.1，在十关之外
11	Obscene, Degrading, and/or Abusive Content	更容易生成与接触淫秽、侮辱、虐待类影像，包括合成儿童性虐待材料与非自愿私密影像	网关（主要）；模型（次要）。与第 3 项同一检查路径，关口过滤加模型安全评测
12	Value Chain and Component Integration	上游第三方组件（包括取得不当的数据）整合不透明、不可追溯，供应商审查不当，损害对下游用户的透明度与问责	工具（主要）；平台（次要）。这是 12 项中被十关覆盖最完整的一项：工具关管组件准入与审查，平台关管整合与部署

D.5 定位说明

DeepWing 的七层与十关是一套面向运行中系统的检查方法。本对照表把十关逐项映射到 NIST AI RMF 1.0（NIST AI 100-1）的功能与子类别，并参照 NIST AI 600-1 生成式 AI 侧写补充风险视角。RMF 回答的是治理层的问题：组织该考虑什么、谁负责、决策如何作出。十关回答的是系统层的问题：一个已经在跑的企业 AI 系统，具体查哪里、看什么证据。两者对齐之后，治理要求可以落到可检查的系统部位上。本对照表是参考性映射（reference crosswalk），不是 NIST 认证，不构成 NIST 的背书，也不构成任何合规结论；本映射由 DeepWing 独立作出，未经 NIST 审查或认可。

文末一 · 缺口清单（正式版 v1.0 未决事项）

以下事项为素材与定版文件撑不住、本版按纪律留空或只立判定方法者，后续版本按附录 C 记修订后补入。

1. 自检题面待回录：快速自检 17 项的题面原文只存在于自检仪现行版，定版文件仅存题号与关主题的对应。本文件各层检查项与第 12 章各关均按题号对应，题面待从自检仪现行版回录后补入第 13 章附表。
2. 定量阈值缺失：路由分级比例、降级触发条件、确认点金额与范围阈值、告警阈值、规模上限数值，现阶段只有原则与个案数字，尚不能升为规范阈值。本版只写判定方法（见 6.5.4 注、11.3.4 等），数值留待后续版本以判例与实测积累为据补入。
3. 供应商静默变更的核查方法待补：MOD 册已有登记实践，但「如何发现未预告的版本变更、发现后如何处置」的方法条文未成文，5.5.5 只立登记与复评要求。
4. 引用文件待逐件核对：第 2 章 2.3 的外部文件，本版未设定合规义务；现有参考资料为延伸阅读体例，须逐件核对版本与日期、确认被条文实际引用后，才可升入引用文件清单。
5. 模型层判例待挂：主层判例只有 INC-023 一条（次层关联 INC-022）；5.4.1 至 5.4.5 五种失效模式暂无判例可挂（见 5.8.3）。
6. 记忆层判例待挂：8.4.3 跨用户串话、8.4.4 该忘的不忘暂无判例可挂（见 8.8.3）。
7. 盯梢层判例待挂：11.4.5 告警疲劳、11.4.6 披露迟滞暂无判例可挂；后者待判例库披露延迟字段回填完整后补（见 11.8.4）。
8. 治理读数待建：DW-AI-GOV 的成熟度读数目前只能对少数格做暂定读数（其数据实情见 DW-AI-GOV v1.0 第五节）；本文件 12.1.5、12.7.5、12.8.5、12.9.5、12.10.5 与 14.4.5 只立衔接口径，不预填读数结论。
9. 逐关条文的素材深浅不一：第 12 章第 9 关平台、第 10 关统一与自主主要依托原则性论述与定版口径展开，判例支撑弱于其余各关，后续版本应以新判例与登记实践增补判定细则。

文末二 · 体例说明（全稿沿用）

一、编号规则

1. 章以整数编号（第 1 章、第 2 章……），层章按七层顺序编排：模型、编排、工具、记忆、知识、门禁、盯梢，依次为第 5 至第 11 章；关不单独成章，统一编入第 12 章，按检查次序编号 12.1 至 12.10。
2. 条以「章号.序号」编号（如 5.1、5.2），条内再分款时用「章号.条号.款号」（如 5.5.5）。条文引用一律写全编号，不写「上一条」「本章第三节」一类相对指代。
3. 检查项编号格式为「层全名-C序号」，如模型-C1、编排-C3、门禁-C10：层全名依次为模型、编排、工具、记忆、知识、门禁、盯梢。编号只作检查项编号与交叉引用，正文叙述一律用中文全名。早先使用的两字母代码格式自本版起停用。
4. 判例与登记条目只写编号（如 INC-023、MOD-012），不写链接地址；同一事件跨册时各自编号、互列关联。

二、术语体例

1. 术语首次出现时写「中文正名（英文 English Name）」，其后只用中文正名。七层与十关的英文名以定版为准：Model、Orchestration、Tool Use、Memory、Knowledge、Access Control、Observability；十关英文为 Gateway、Access Control、Orchestration、Knowledge、Memory、Tool Use、Model、Observability、Platform、Alignment & Autonomy。
2. 定版废止名不得在正文使用，包括：Gating、Oversight、Knowledge Base、Model Layer、AI Gateway、Enterprise AI Platform。引用历史材料时出现废止名的，加注现行正名。
3. 每层章首节定义段是该层术语的定义出处，其他章引用时不重复定义，只给出处条号。

三、条文体例

1. 条文用语分级：「应」为要求，「不得」为禁止，「宜」为建议，「可」为允许。检查项的判定只对「应」与「不得」类条目作合格判定。
2. 每层章的固定小结构为：范围与定义 / 职责边界 / 组成要素 / 常见失效模式 / 规范条目 / 检查项 / 与其他层及关口的接口 / 关联判例，八节不增不减，节序不调。
3. 每句话须属三种属性之一：来源事实、深翼解读、未知。条文中的事实性陈述须能指出素材出处；属解读的，在章首或节首统一标明，不与条文混排冒充事实。
4. 正文不用长破折号（em dash 与 en dash），停顿与并列用逗号、顿号或分行；占位符统一用「·」。

四、版本记录格式

1. 版本号格式为　v主版本.次版本（如　v1.0、v1.1）：结构性调整（增删章、改层关口径）升主版本，条文增补与文字修订升次版本。
2. 附录　C　修订记录每行六个字段：日期、版本、修订章节、修订内容、依据（判例编号、登记编号或定版文件）、修订人。
3. 条目与记录的状态词统一为：在用、已更正、已撤回。已更正者旧说法留在变更记录；已撤回者编号永不复用。
4. 每条记录保留「最后核查」日期；框架条文的复核周期随第 14 章执行，条文被判例推翻时，先记变更记录，再改正文。