

公共云 AI 推理服务稳定性建设研究报告

(2026 年)

中国信息通信研究院云计算与数字化研究所

阿里云计算有限公司

2026年9月

版权声明

本报告版权属于中国信息通信研究院和阿里云计算有限公司，并受法律保护。转载、摘编或利用其他方式使用本报告文字或者观点的，应注明“来源：中国信息通信研究院、阿里云计算有限公司”。违反上述声明者，编者将追究其相关法律责任。

前 言

以大模型为核心的人工智能（Artificial Intelligence, AI）技术正加速规模化应用。公共云承担模型部署、算力调度和应用接入等关键功能。随着推理规模扩大、应用链路延长，稳定性问题已从模型可用性延伸至分布式链路和基础设施，成为产业发展的重要制约因素。

本报告所称“公共云 AI 推理服务稳定性”，是指在多租户共享资源、弹性按需供给和跨可用区、跨地域部署条件下，公共云持续提供可预期、可恢复、可审计推理服务的能力。其对象不仅是模型接口，还包括算力、网络、存储、模型调度、智能体（Agent）工具链和知识服务构成的端到端链路。

围绕上述内涵，本报告将发展背景、风险变化与建设路径贯通分析，提出“纵向分层、横向协同”的治理框架：基础设施层提升算力、网络和存储的连续供给能力；模型服务层加强上线准入、动态调度和故障恢复；应用服务层面向模型即服务（MaaS）、人工智能辅助编程（AICoding）、智能体应用和知识服务强化任务级保障；全栈可观测能力横向贯穿各层，形成监测、分析、处置和复盘闭环。

未来，随着算力架构演进、推理场景拓展和标准体系完善，公共云 AI 推理服务稳定性将由单点保障转向系统治理。本报告旨在为云服务提供者、行业应用方和标准制定者提供参考，推动构建安全可信、高效韧性的公共云 AI 推理服务体系。

目 录

一、稳定性建设背景与内涵	1
（一）推理服务规模化，稳定性成为公共云核心能力	1
（二）服务链路复杂化，单点保障难以支撑全局韧性	2
（三）安全可信要求提升，推动稳定性治理体系升级	3
二、稳定性分层治理框架	4
（一）基础设施筑底，保障资源供给连续可靠	5
（二）模型服务承中，提升推理链路交付韧性	16
（三）应用服务牵引，强化业务场景端到端保障	16
（四）可观测体系贯通，形成监测评估处置闭环	16
三、基础设施稳定底座	17
（一）算力弹性调度，保障推理资源稳定供给	17
（二）网络高可用设计，降低分布式推理链路抖动	17
（三）存储可靠支撑，稳定模型加载与知识检索	17
四、模型服务稳定能力	17
（一）上线准入前置，降低模型变更风险	17
（二）动态调度优化，提升推理服务质量	17
（三）容错恢复闭环，增强推理链路韧性	18
（四）度量评价牵引，支撑稳定性精细治理	18
（五）租户隔离与配额管控	19
五、AI 应用服务稳定实践	20
（一）AI 应用稳定性治理框架与指标体系	20
（二）应用服务稳定性实践，保障服务连续性	27
（三）Harness 稳定性工程实践，构建通用智能体运行底座	31
（四）智能体应用实践，增强业务任务执行韧性	38
六、全栈可观测能力体系	41
（一）全栈可观测的技术地基	41
（二）AI 基础设施层可观测	44
（三）模型服务层可观测	46

（四）应用服务层可观测	49
（五）贯穿三层的横切能力	51
七、发展趋势与建设建议	52
（一）技术智能化：推动稳定性能力自治演进	52
（二）行业分级化：建立差异化服务保障机制	53
（三）标准协同化：完善评测基准与责任边界	53
（四）运营常态化：强化演练复盘与持续改进	54

图 目 录

图 1 公共云 AI 推理服务稳定性分层治理框架示意	4
图 2 稳定性内涵变化	6
图 3 面向 Token 为中心的新型网络架构	11
图 4 云原生分层解耦的网络弹性承载技术示意图	12
图 5 意图感知的自适应高精度网络监控技术示意图	13
图 6 Overlay 网络驱动的租户无感快速容灾技术示意图	14
图 7 Agent 应用新挑战	21
图 8 面向 Agent 应用的稳定性五层参考架构图	22
图 9 AI 应用服务稳定性成熟度模型	26
图 10 通用 Agent Harness 六子系统概览图	33
图 11 Agentloop	34
图 12 上下文与状态治理	35
图 13 错误恢复与降级处置	38
图 14 任务状态与工具健康检查	39
图 15 长任务续跑、记忆隔离与多 Agent 协作	40

表 目 录

表 1 配额管控：从被动限流到分层容量治理.....19

表 2 AI 应用服务稳定性指标体系 24

表 3 关键阈值 25

表 4 成熟度模型 27

表 5 多模型切换机制 28

表 6 监控指标 30

表 7 常见错误分类与处置机制 31

表 8 Harness 核心组件 32

表 9 工具治理分级建议 37

表 10 基础设施层关键指标 45

表 11 推理过程维度关键指标 47

表 12 推理引擎状态维度关键指标 48

一、稳定性建设背景与内涵

（一）推理服务规模化，稳定性成为公共云核心能力

推理服务规模化和链路复杂化，使稳定性成为公共云 AI 服务从“可调用”走向“可依赖”的基础能力。大模型应用正从试点走向生产，推理调用量、并发量和业务依赖度持续提升。据 MarketsandMarkets 报告，2025 年全球 AI 推理平台即服务（AI Inference PaaS）市场规模达 188.4 亿美元，预计 2030 年增至 1052.2 亿美元，年复合增长率（Compound Annual Growth Rate, CAGR）为 41.1%；2025 年国内 AI 推理算力市场规模达 438.5 亿元，2021—2025 年 CAGR 为 66.3%。

企业投入也加速向推理侧倾斜。MaaS 平台中，企业对企业（Business-to-Business, B2B）场景已成为 Token 消耗的重要来源。

企业对 AI 服务的关注点也由模型准确率转向生产环境中的稳定表现。服务等级协议（Service Level Agreement, SLA）、故障恢复、安全合规、可用率和审计能力，正成为企业采购的重要要求。因此，公共云平台需加强资源调度、模型交付、业务保障和安全合规，推动 AI 服务从“可调用”走向“可依赖”。

（二）服务链路复杂化，单点保障难以支撑全局韧性

公共云 AI 推理服务涉及网关、鉴权、模型路由、推理实例、缓存、检索、工具调用和结果处理等多个环节，任一环节异常都可能影响端到端服务质量。与传统云服务相比，AI 推理服务主要具有四类复杂性：一是请求规模差异大，简单问答与长上下文请求共享资源池；二是输出具有非确定性，故障复现和定位困难；三是多轮对话和长任务状态复杂，容易出现上下文溢出或丢失；四是尾延迟明显，第 99 百分位延迟（P99 延迟）通常远高于中位延迟。根据 Ookla 旗下 Dnswat 的统计，2025 年第一季度至 2026 年第一季度，主流 AI 平台的高故障天数由 6 天增至 51 天，表明 AI 服务的稳定性风险正在增加。

因此，公共云 AI 推理服务稳定性建设应以模型服务和 AI 应用为核心，重点提升模型交付、请求调度、异常恢复、任务连续性和结果质量保障能力。算力、网络 and 存储等基础资源作为推理服务运行的关键依赖，应提供必要的资源可用性、容量状态和故障信息，支撑上层开展服务切换、弹性调整、降级和恢复，推动 AI 推理服务从“可调用”走向“可依赖”。

稳定性建设不应停留在服务器、容器或接口网关等单点组件，也不需要底层基础设施进行全面展开，而应围绕 AI 推理服务形成“基础资源支撑—模型服务治理—应用任务保障”的协同体系。其中，基础资源侧重点保障关键依赖可用、状态可感知、故障可隔离；模型与应用侧重点保障服务可调用、任务可完成、异常可恢复、结果可验证。

（三）安全可信要求提升，推动稳定性治理体系升级

随着 AI 推理服务进入政务、金融、医疗和工业等重点行业，稳定性已与数据安全、内容安全、合规审计和责任认定紧密结合。政策层面，《关于深入实施“人工智能+”行动的意见》提出加强模型、数据、基础设施和应用系统安全能力建设；《人工智能生成合成内容标识办法》及强制性国家标准 GB 45438—2025《网络安全技术 人工智能生成合成内容标识方法》要求对人工智能生成合成内容设置显式和隐式标识，并在生成、传播和下载导出等环节落实标识责任。海外合规方面，对于纳入高风险范围的系统，欧盟《人工智能法案》要求其在全生命周期保持适当的准确性、鲁棒性和网络安全水平，并通过风险管理、日志、质量管理及冗余或故障安全措施提高一致运行能力；GDPR 第 32 条要求处理系统持续具备保密性、完整性、可用性和韧性，并在物理或技术事件后及时恢复数据可用性和访问；SOC 2 不是法律法规，而是围绕安全性、可用性、处理完整性、保密性和隐私评价服务组织控制设计及运行有效性的鉴证框架。

安全可信要求主要从三个层面推动治理升级：一是准入层面，模型上线前应完成备案审核、安全评估和红队测试。红队测试是指通过模拟攻击和异常输入检验系统防护能力。二是运行层面，应具备内容安全检测、异常识别、访问审计和故障分级响应能力。三是持续运营层面，应通过故障复盘、应急演练和安全审计完善服务等级、变更管理、监测和应急机制，形成“设计即合规、上线即受控、运行即可观测”的治理闭环。

二、稳定性分层治理框架

公共云 AI 推理服务稳定性治理以模型服务层和应用服务层为核心，以基础资源支撑层为运行依赖，并由全链路可观测能力横向贯通。基础资源侧重点保障资源可获得、关键依赖可感知和基础故障可隔离；模型服务侧重点保障模型可用、调用稳定和容量可调度；应用服务侧重点保障业务任务连续执行、异常恢复和结果质量。



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 1 公共云 AI 推理服务稳定性分层治理框架示意

（一）基础设施筑底，保障资源供给连续可靠

基础资源支撑层主要保障 AI 推理服务所依赖的计算、调用链路和数据访问能力持续可用，并将相关状态和异常影响传递给模型服务层。

基础资源侧应重点关注推理容量可用性、容量水位、实例健康状态和资源分配结果，为模型服务的并发控制、请求排队、弹性调整和异常降级提供依据。

调用链路侧应重点关注推理调用链路的连通性、时延、丢包和跨区域通信状态，为请求路由、超时控制、异常绕行和区域切换提供支撑。

数据访问侧应重点保障模型文件加载、缓存访问和知识数据检索的可用性与时效性，并通过重试、缓存、冗余和降级机制降低存储异常对模型调用和应用任务的影响。

（二）模型服务承中，提升推理链路交付韧性

MaaS 平台面临租户干扰、版本异常、突发流量和区域故障等风险，需要建立覆盖事前预防、事中处置和事后改进的全生命周期治理体系。

一是建立上线准入前置机制。将依赖检查、性能压测和安全评估纳入持续集成与持续交付（Continuous Integration/Continuous Delivery, CI/CD）流水线，并通过大语言模型（Large Language Model, LLM）

回归测试和灰度发布降低变更风险。二是实施动态调度优化。根据请求特征、实例负载和业务优先级实施动态路由，在容量紧张时采用排队、降级和异步处理。三是构建容错恢复闭环。通过异常检测、故障隔离、熔断、回退和流量切换形成容错恢复闭环。四是完善度量评价体系。建立覆盖模型、实例和租户的性能与质量指标，提高服务可预期性。五是强化租户隔离与配额管控。通过每分钟请求数（Requests Per Minute, RPM）、每分钟 Token 数（Tokens Per Minute, TPM）、弹性预留和增速限制，加强租户隔离与配额管控。

（三）应用服务牵引，强化业务场景端到端保障

随着 AI 应用由单轮模型调用向智能体（Agent）演进，稳定性问题已从“接口是否可用”扩展为“业务任务能否完成”。多轮会话、任务编排、工具调用、知识检索和长上下文管理等任一环节异常，都可能导致任务中断、循环失控、权限越界、成本失控或结果不可信。

稳定性内涵之变：从「接口可用」到六个维度

公共云 AI 推理服务正从单轮模型调用加速演进到 Agent 化应用。「模型接口可用但业务任务不可用」已成为规模化落地阶段最突出的稳定性问题。

传统 Web 应用稳定性

围绕请求成功率、接口延迟、服务可用性、资源水位；以「模型 API 是否返回成功」衡量。

Agent 应用稳定性

新增多轮会话状态、动态任务编排、异步工具调用、外部知识依赖、长上下文管理、内容可信治理；任一变量异常都可能导致任务中断、循环失控、权限越界、成本失控或结果不可验证。

AI 应用稳定性的六个衡量维度



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 2 稳定性内涵变化

因此，AI 应用稳定性应覆盖任务可完成、结果可信任、过程可恢复、链路可追溯、成本可控制和安全可治理六个维度。应用服务层应围绕智能体运行时支撑体系（Agent Harness）、智能体应用和检索增强生成（Retrieval-Augmented Generation, RAG）服务，建设状态管理、工具治理、异常恢复、结果校验和持续评估能力。

（四）可观测体系贯通，形成监测评估处置闭环

AI 应用请求可能经过终端、认证、会话管理、任务编排、知识检索、工具调用、模型路由和底层硬件等多个环节。传统以基础设施指标和日志为主的监控方式，难以解释故障位置、原因及业务影响。

因此，应构建以模型服务层和应用服务层为核心，并关联基础资源支撑层的全栈可观测体系，以 OpenTelemetry 为数据基础，以端到端链路追踪为骨架，统一监测性能与可用性、质量与安全以及资源成本。

应用服务层重点观测 RAG 链路、Agent 及模型上下文协议（Model Context Protocol, MCP）工具调用、语义质量和成本；模型服务层重点监测模型调用、vLLM 和 SGLang 等推理引擎；基础设施层重点监测 GPU、显存、远程直接内存访问（Remote Direct Memory Access, RDMA）网络、存储和 Kubernetes。横向建设统一的指标、日志和链路追踪（Metrics、Logging、Trace）存储，以及 AI 网关、自动化评估和智能告警能力。

体系建设应遵循开放标准、低侵入低开销、语义统一、数据安全和业务价值导向五项原则，形成从异常发现、关联分析到处置改进的治理闭环。

研究归纳：公共云场景的核心不是单项技术叠加，而是把多租户隔离、跨故障域部署、SLA 约束和共享责任映射到分层能力，并由统一可观测数据驱动跨层协同。

三、基础设施稳定底座

（一）算力弹性调度，保障推理资源稳定供给

1. 挑战与背景

基础设施稳定性的核心在于通过算力弹性、网络高可用和存储韧性，保障异构资源持续稳定供给。大模型推理具有流量波动明显、并发需求高和服务时延敏感等特点，不同模型在资源需求、启动时间和运行方式方面也存在差异。当算力容量不足、服务实例异常或资源交付不及时，可能造成请求排队、响应时延升高和模型服务不可用。

因此，算力资源保障应以推理服务连续运行为目标，重点关注资源供给的确定性、资源使用效率、容量调整速度和异常恢复能力，并将资源状态与模型、版本、租户及服务区域建立关联，为上层模型服务的调度、限流、降级和恢复提供依据。

2. 资源确定性保障

对于具有明确服务等级要求或流量相对稳定的模型服务，可通过资源预留方式保障基础推理容量；对于流量波动明显或保障要求较低

的服务，可采用按需供给方式补充弹性容量。

资源预留与弹性供给应根据模型服务的重要程度、历史流量、容量水位和恢复目标进行配置。当可用容量不足时，应优先保障核心模型和关键租户，并通过请求排队、并发限制、备用模型切换等措施控制影响范围。

3.资源池化与弹性复用

通过软件定义资源池，可打破物理设备与固定规格的绑定，减少资源碎片。代表性公共云容器平台（如阿里云 ACK、Amazon EKS、Microsoft Azure AKS 和 Google Kubernetes Engine）均可通过节点池、标签/污点及跨可用区调度策略组织弹性资源，具体能力边界因平台而异。GPU 资源可通过虚拟化切分和多容器共享实现细粒度复用，并通过显存隔离降低租户干扰。Binpack 策略优先集中调度以减少碎片，Spread 策略分散调度以提高可用性，从而将 GPU 由“整卡独占”转变为按需切分和共享。

4.智能弹性伸缩

弹性伸缩系统应根据请求量、排队长度、并发情况、响应时延和容量水位等指标，动态调整模型服务容量。弹性调整还应考虑模型加载和服务启动时间，合理设置触发条件，避免容量调整滞后或频繁变化造成服务波动。

5.故障自愈闭环

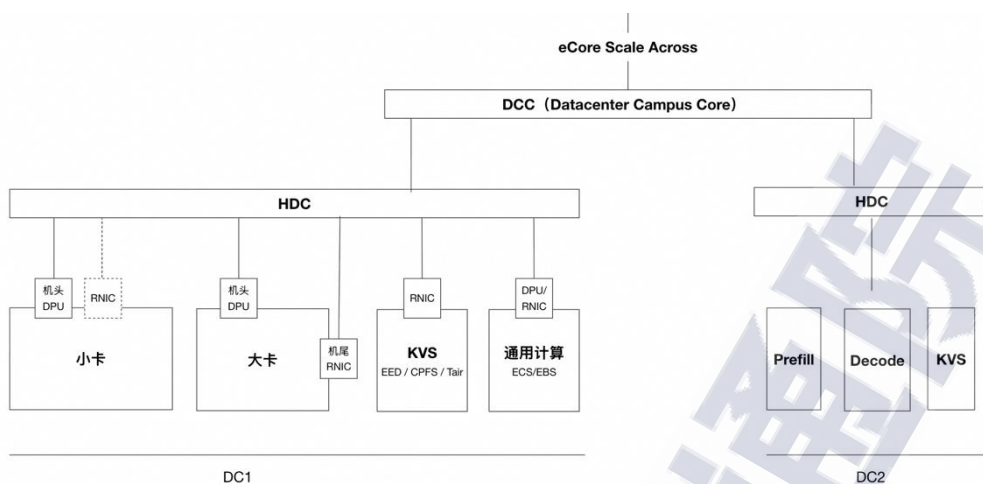
大规模异构算力集群需建立“检测—诊断—分级处置”的故障自愈机制。虚拟机异常时，系统可自动创建新实例替换，并根据宿主机状态执行重启或迁移。GPU 诊断应覆盖显存双比特错误、Xid 异常、设备丢卡、驱动、固件、供电和网卡等问题。软件故障可通过重启或重装驱动恢复；硬件故障则通过运维事件通知，在用户确认后完成资源更换。

综上，算力弹性调度通过资源预定、池化复用、快速伸缩和故障自愈，将推理容量由静态分配转向动态保障，为模型服务连续运行提供必要支撑。

（二）网络高可用设计，降低分布式推理链路抖动

在 Token 原生架构中，预填充（Prefill）、解码（Decode）、键值缓存（KV Cache）、混合专家模型（Mixture of Experts, MoE）和跨集群协同逐步网络化，网络已成为 Token 生产链路的重要组成部分。行业通用方案通常从流量分平面承载、缓存感知路由、QoS 调度和跨地域容灾四个方面治理。

1. 面向 Token 为中心的新型网络架构



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 3 面向 Token 为中心的新型网络架构

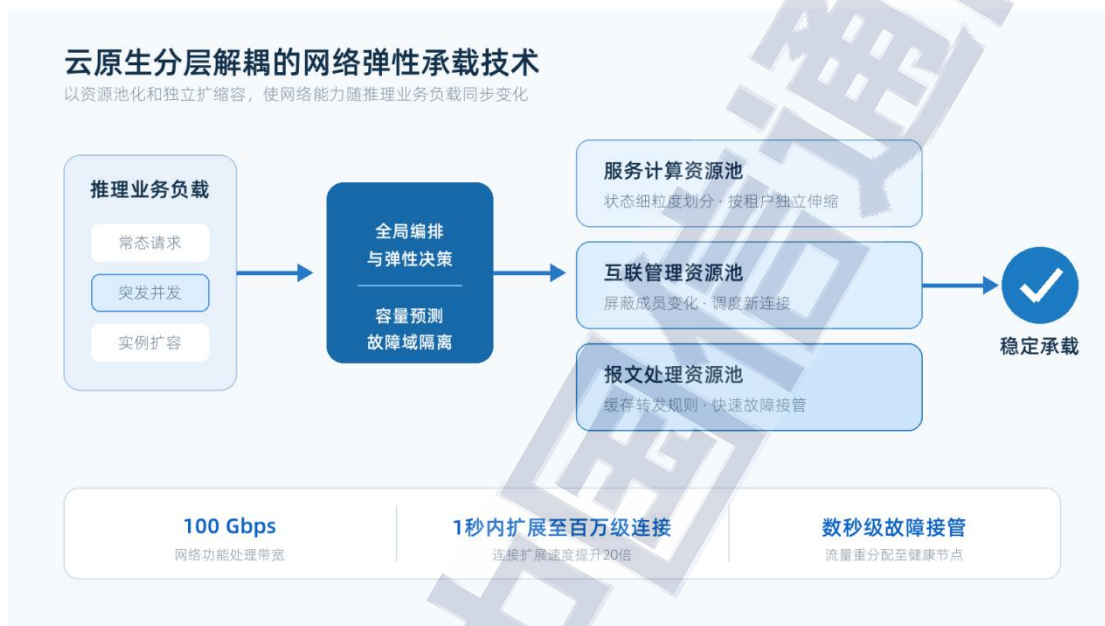
该架构主要包括五项能力：一是按 Token 生产链路特征实施多平面、多路径差异化承载；二是面向预填充与 PD 分离和 KV Cache，通过近存储访问、缓存感知路由和零拷贝卸载降低状态迁移压力；三是通过组播等方式优化智能体场景的缓存流量；四是结合推理引擎实施流量整形和服务质量（Quality of Service, QoS）调度；五是依托 TPN（Token Performance Network，基于 Token 性能的网络架构）与 eCore 实现跨集群缓存访问和推理请求调度。

2. 高性能极致弹性容器网络

容器网络应支持模型服务实例的快速创建、调整和迁移，并保证服务容量变化期间的连接连续性。相关连接、访问策略和服务路由应能够与模型服务实例同步调整，缩短服务上线和容量调整所需时间。对于长连接、流式输出和智能体多轮调用，还应支持异常重连和超时控制。

3. 云原生分层解耦的网络弹性承载技术

该技术将网络功能拆分为服务计算、互联管理和报文处理资源池，各资源池可根据连接数、带宽和负载独立扩缩，降低状态同步对扩容效率和业务连续性的影响。



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 4 云原生分层解耦的网络弹性承载技术示意图

相关实践显示，主流云厂商的单实例或单 VIP 能力已普遍达到数十 Gbps、数百万至千万级并发连接以及数十万至百万级每秒新建连接，阿里云使用该技术可实现 100 Gbps 处理带宽，并在 1 秒内扩展至百万级连接；实例故障后，流量可在数秒内切换至健康节点。实际应用中，应通过弹性阈值、故障域隔离和实例限速控制异常流量影响。

4. 意图感知的自适应高精度网络监控技术

该技术结合虚拟拓扑、路由配置、资源变更和活跃路径动态生成探测任务，通过端到端、逐跳和边界探测发现丢包、时延和节点异常，并区分虚拟网络、物理网络、用户侧及运营商侧问题。



图 5 意图感知的自适应高精度网络监控技术示意图

相关评估显示，在探测任务减少 78% 的情况下，可覆盖超过 90% 的租户流量，虚拟拓扑通常在 6 秒内完成更新。实际部署中应实施分层采样，控制探测开销。

5.Overlay 网络驱动的租户无感快速容灾技术

该技术通过调整隧道报文外层源端口，改变等价多路径路由的哈希结果，将流量引导至已验证的可用路径，在底层路由完全收敛前完成快速绕障。相关实践显示，按活跃 IP 探测可降低 60%~70% 的成本，路径状态数据可压缩约 25 倍，18 个月内累计网络中断时间降低

92.71%。该技术依赖冗余路径，不能替代基础容量和跨故障域容灾；多路径同时受损时，仍需进行可用区或地域级切流。



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 6 Overlay 网络驱动的租户无感快速容灾技术示意图

总体来看，上述技术分别从弹性供给、主动监控和快速恢复三个方面提升网络稳定性，并可将网络指标与请求成功率、首字时延、输出中断率和任务完成率关联。

（三）存储可靠支撑，稳定模型加载与知识检索

随着大模型和生成式 AI 发展，公共云对象存储逐步承担模型加载、智能体数据管理和知识检索等任务，普遍面临高并发、海量命名空间、向量检索存算融合和高频变更等挑战。阿里云对象存储服务（Object Storage Service, OSS）作为联合研究案例，可用于说明相关稳定性机制的工程实现。

从行业共性机制看，可将基础对象存取的数据平面与弹性调度、元数据管理和向量索引计算的智能平面解耦，并通过存算隔离降低智能计算对主存储链路的影响。阿里云 OSS “双平面稳定性架构”是该机制的一项实践，其数据可靠性目标为 99.999999999%（12 个 9），AWS S3 的“跨可用区多副本冗余架构”存储的数据可靠性是 99.999999999%（11 个 9）；该指标属于特定服务承诺，应用方仍需结合地域、冗余类型和 SLA 边界核验。

1.面向 AI Workloads 的弹性高可用架构设计

（1）智能流量整形与多租户自适应 QoS 调度

系统在无状态网关层引入自适应背压（Backpressure）和流量整形，针对检查点（Checkpoint）等突发流量，按租户和存储桶（Bucket）进行限流与降级。同时，采用单元化架构（Cell-based Architecture）划分隔离单元，将单一客户的异常流量和故障影响限制在特定 Cell 内。

（2）轻量级元数据底座：破解 Agentic Bucket 难题

针对十亿级 Bucket 管理需求，系统通过逻辑命名空间和轻量级虚拟 Bucket，降低智能体存储桶（Agentic Bucket）的资源与元数据开销。元数据层采用水平分片（Metadata Sharding），按 Agent ID 分布数据，并结合无状态网关和分布式架构，降低查询时延和单点热点风险。

（3）存算物理隔离

针对融合向量检索能力的 Vector Bucket，基础对象存取和向量索引构建、查询分别运行在独立计算池（Compute Pool）中，实现 CPU 和内存资源隔离。向量索引采用异步 Checkpoint 机制。极端负载下，系统优先保障基础对象存取，并对向量查询实施降级，如返回 Top-K 粗排结果，维持主干系统稳定。

2. 全生命周期的研发与变更管控体系

针对快速迭代的挑战，公共云对象存储需要构建基于“防御性工程”的 DevOps 与 SRE 体系。以下以阿里云 OSS 实践说明其关键环节：

（1）单元化灰度与爆炸半径收敛

任何涉及 AI 新特性的发布，均应遵循“单节点→单 Cell→单可用区(AZ)→全地域”的渐进式灰度策略。配合实时自动化指标对齐，一旦发现新版本导致 SLA 波动，系统应自动阻断变更。

（2）常态化混沌工程(Chaos Engineering)

在 Agentic 和 Vector 等新特性的上线准入标准中，强制引入故障注入测试。通过模拟大面积断网、计算节点 OOM、元数据脑裂等极端场景，验证系统在“异常态”下的自愈能力和数据一致性。研究归纳：基础设施稳定性的共性特征是“资源池化+故障域隔离+弹性供给+自动恢复”。厂商产品形态可以不同，但评价时应统一关注供给确定性、扩缩时延、故障隔离范围和恢复目标。

四、模型服务稳定能力

（一）上线准入前置，降低模型变更风险

模型服务稳定性应围绕上线准入、动态调度、容错恢复和容量治理，构建覆盖全生命周期的保障闭环。模型版本、推理引擎、提示模板、安全策略和依赖组件持续迭代，模型服务变更已覆盖算法、配置、容量和服务策略，可能引发时延波动、质量下降、兼容异常和容量失衡。模型服务应建立持续集成与持续交付（Continuous Integration/Continuous Delivery, CI/CD）准入体系，将依赖检查、兼容验证、性能压测、安全评估、容量预检和回退预案纳入统一流水线。变更上线前应完成自动校验，并通过小流量、分租户、分区域灰度发布控制影响范围。

针对大模型输出的非确定性，应将脱敏后的失败样本、边界样本、长上下文样本和行业样本纳入大语言模型（Large Language Model, LLM）回归测试，验证任务完成率、格式遵循、内容安全、事实一致性和性能变化。线上问题还应沉淀为问题库、评测样例、监控阈值和应急脚本，形成“发现—归因—修复—验证—回归”闭环。

（二）动态调度优化，提升推理服务质量

推理请求在上下文长度、生成长度、实时性和业务优先级等方面差异明显。平台应结合模型能力、实例负载、排队长度、租户等级和区域容量，统一决定请求使用的模型、实例、执行时间和优先级。容量紧张时，可通过排队、动态批处理、上下文压缩、模型替代和异步

处理降低压力；容量充足时，通过弹性伸缩提高资源利用率。当实例出现高尾延迟、错误率上升或容量不足时，应将请求切换至健康实例、替代模型或备用区域。核心业务及 SLA 范围内的请求应优先保障；低优先级或超出约定范围的请求，可采用排队、降级、异步处理或尽力承接。

（三）容错恢复闭环，增强推理链路韧性

模型服务故障可能由版本、实例、依赖、网络、容量或安全策略共同触发，并表现为错误率升高、首 Token 时延增加、生成中断和队列堆积。平台应围绕请求、实例、模型、租户和区域建立异常检测，形成“异常发现—影响评估—快速控制—服务恢复—复盘改进”闭环。处置措施包括实例摘除、版本或配置回退、流量切换、容量扩展、限流降级和备用模型启用。大范围故障应优先保障核心业务，再逐步恢复普通流量。架构侧还应通过故障域隔离、健康检查、自动摘流、幂等重试和熔断保护降低故障影响。长输出、流式响应和工具增强推理场景，应支持请求去重、结果补偿和中断恢复。故障复盘结果应反向更新准入规则、压测场景、监控阈值和演练脚本。

（四）度量评价牵引，支撑稳定性精细治理

模型服务度量体系应覆盖模型、实例、请求、租户、区域和版本。除可用率、错误率、吞吐量、排队长度和端到端时延外，还应监测首 Token 时延、Token 生成速率、生成中断率、上下文溢出率、冷启动时间和模型切换成功率。指标既用于容量规划、变更评估、故障定位

和调度优化，也用于表达服务等级和可用性边界。多租户场景还应关注重点客户、模型、区域和业务，避免平均值掩盖局部风险。度量评价应结合模型质量和用户反馈，监测任务完成率、格式遵循率、内容安全命中率、事实一致性、用户重试率和人工干预率，使治理范围由资源运行延伸至服务质量。

（五）租户隔离与配额管控

MaaS 平台需防止多租户流量相互干扰，重点应对核心租户被挤兑、在线与离线任务争抢资源、瞬时流量击穿配额，以及业务上量准备不足导致 HTTP 429 限流等问题。

表 1 配额管控：从被动限流到分层容量治理

管控维度	机制	触发条件	治理价值
租户级配额	RPM/TPM 双阈值	租户用量达到阈值	防止单租户挤占资源
模型级配额	按模型分池限流	热点模型水位升高	隔离热点或故障影响
优先级队列	高/中/低优先级调度	资源紧张或排队增长	优先保障核心业务
弹性预留	预留容量+动态扩缩	活动或业务上量窗口	缩短扩容响应时间
过载保护	自适应限流+排队水位	队列、错误率或时延超阈	避免级联故障和雪崩

来源：中国信息通信研究院、阿里云计算有限公司联合编制

平台可同时设置每分钟请求数（Requests Per Minute，RPM）和每分钟 Token 数（Tokens Per Minute，TPM）限制，任一指标达到上限即返回 HTTP 429。核心租户宜采用独享配额和资源隔离，普通租

户采用公平调度和弹性扩容，离线任务通过批量调用接口（Batch API）独立执行。

重点客户可建立三类信息表：场景表记录模型、业务及影响范围；用量表记录调用量和 Token 趋势；限流表跟踪 RPM、TPM 使用率，超过 70% 进入关注，超过 80% 推动扩容。对于计划上量的业务，可根据平均输入 Token 数、平均输出 Token 数和峰值每秒请求数（Queries Per Second, QPS）测算配额： $RPM = QPS \times 60$ ； $TPM = (\text{平均输入 Token 数} + \text{平均输出 Token 数}) \times RPM$ 。大型客户还可按“主账号+模型”“工作空间+模型”“API 密钥+模型”设置多粒度配额。为防止瞬时流量击穿分钟级限制，可采用秒级窗口和 Token 增速限制。配额治理应统一观测限流次数、配额水位、扩容提前量、重点客户成功率和在线请求时延，实现从被动限制向主动保障转变，并通过智能告警收敛、根因分析和自动化切流缩短故障发现、定位与止血时间。

研究归纳：模型服务稳定性可抽象为“变更可控、调度可预期、故障可隔离、容量可治理”四项能力。研究评价不应只看平均可用率，还应按模型、租户、区域和版本观察尾延迟、切换成功率及配额水位。

五、AI 应用服务稳定实践

（一）AI 应用稳定性治理框架与指标体系

1. 应用侧稳定性面临的新挑战

Agent 应用的新挑战与四个基本目标

三类新挑战决定了稳定性的四个基本目标；目标再逐层细化为六个衡量维度与七类指标体系——本章即沿“目标→维度→指标”层层展开。

三类新挑战

- 01 链路长、复合失败率高**
一次任务含数十轮推理/工具/检索/校验，任一中间步骤失败即任务失败，请求级 SLA 难覆盖真实感知。
- 02 静默失败更危险**
看似合理但缺证据的答案、工具部分成功、RAG 错误召回、长任务阶段成果未固化即中断。
- 03 失败不总有错误码**
静默失败对业务决策的影响往往大于一次显式报错。

四个基本目标

✓ 任务可完成

有限步骤内到达明确终态，多步任务端到端跑完。

衡量：任务完成率 TCR

🛡️ 结果可信任

证据支撑、引用可追溯、通过安全护栏。

衡量：引用覆盖率 / 幻觉率

🔄 过程可恢复

任意阶段失败后可基于已有状态与阶段成果恢复。

衡量：断点续跑成功率

🔍 链路可追溯

每步决策/调用/检索/拦截可观测、可回放、可审计。

衡量：Trace 覆盖率

来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 7 Agent 应用新挑战

传统模型应用程序接口（API）调用通常对应单次响应，风险主要包括超时、模型不可用、限流、鉴权失败和参数错误，可通过超时控制、重试、熔断和降级等远程过程调用（RPC）治理机制处理。进入智能体（Agent）阶段后，一项任务可能包含多轮模型推理、工具调用、知识检索、状态更新和结果校验，任一环节失败都可能导致任务中断。模型输出缺少证据、工具部分执行成功、检索增强生成（RAG）召回错误及阶段成果丢失等“静默失败”，也难以通过错误码发现。

Agent 应用稳定性应满足四项目标：任务可完成、结果可信任、过程可恢复、链路可追溯。

2.Agent 应用稳定性五层参考架构

建议构建接入网关层、会话与记忆层、编排与工具层、知识与检

索层、结果治理与风险处置层，并以可观测和安全防护体系贯穿各层。



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 8 面向 Agent 应用的稳定性五层参考架构图

接入网关层负责鉴权、协议适配、模型路由、Token 限流、语义缓存和内容安全。主模型出现限流、超时或异常时，应自动故障转移（Failover）至备选模型。流量控制应覆盖每分钟请求数（RPM）、每分钟 Token 数（TPM）、每秒请求数（RPS）和每秒 Token 数（TPS）。

会话与记忆层负责保存历史消息、工具结果和任务状态，防止上下文溢出或丢失。长期记忆用于沉淀用户偏好和业务背景，长任务则通过检查点（Checkpoint）实现断点续跑。编排与工具层可采用推理一行动（ReAct）或规划后执行（Plan-and-Execute）模式，并设置最大轮次、重复动作检测、分支上限和终止条件。工具调用应配置超时、重试、熔断、健康检查、权限审批和沙箱隔离。知识与检索层应覆盖

数据入库、索引构建、检索召回、答案生成和权限隔离，并通过 Recall@K、平均倒数排名（MRR）、归一化折损累计增益（NDCG）、引用覆盖率、证据一致性和幻觉率等指标评价质量。结果治理与风险处置层负责格式、事实、置信度、安全和业务规则校验。证据不足、工具失败或触发安全策略时，应采取澄清、拒答、降级回答、人工接管或异步补偿。

3.AI 应用服务稳定性指标体系

AI 应用稳定性指标应同时覆盖传统可用性与性能，以及 Agent 特有的任务可靠性、内容可信度、成本稳定性和恢复能力。具体稳定性指标体系可参考：

表 2 AI 应用服务稳定性指标体系

指标类别	核心指标	关注问题	典型治理手段
服务可用性	SLA、请求成功率、5xx 率、429 率、模型不可用率	平台接口与模型服务是否可用	多模型路由、限流、熔断、模型 Fallback、区域切流
性能稳定性	TTFT、TGS、TP95/TP99 延迟、端到端任务耗时	用户感知响应是否稳定	语义缓存、Batch 异步化、并发控制、推理资源隔离
任务可靠性	任务完成率、平均执行步数、工具调用成功率、断点续跑成功率、异常终止率	Agent 是否能完成业务目标	任务状态机、Loop 上限、工具健康检查、Checkpoint
质量可信度	幻觉率、引用覆盖率、证据一致性、检索命中率、人工纠错率	输出是否可信、可验证	RAG 评测、引用追溯、Guardrails、低置信度兜底
成本稳定性	单任务 Token 成本、异常重试成本、缓存命中率、Batch 占比	资源消耗是否可控	Token 预算、语义缓存、任务拆分、异步通道
安全合规性	护栏命中率、Prompt 注入拦截率、敏感数据识别率、违规输出率	输入输出是否安全合规	输入输出双向检测、权限隔离、工具审批、敏感信息过滤
运营恢复能力	故障发现时间、MTTR、Fallback 命中率、回滚耗时、切流耗时、复盘闭环率	故障能否被快速发现、止血和复盘	告警分级、故障 SOP、版本回滚、客户可观测、复盘机制

来源：中国信息通信研究院、阿里云计算有限公司联合编制

模型故障切换（Fallback）应监测请求成功率、切换耗时和用户感知失败率；断点续跑应监测任务完成率、恢复耗时和重试成本；RAG 引用追溯应监测证据一致性、幻觉率和人工纠错率。统一指标可推动

稳定性治理由经验驱动转向数据驱动。平台可为关键指标设置发布门禁和容量预警阈值，但应结合业务重要性、模型能力和服务等级协议（SLA）动态调整。

表 3 关键阈值

指标类别	指标名	建议阈值或观测口径	适用场景
任务指标	任务完成率	建议不低于 85%，核心业务应单独设更高基线	Agent 全场景
任务指标	平均执行步数	建议关注突增，复杂任务可先以 15 轮作为初始观察线	Harness、智能体应用
工具指标	工具调用成功率	建议不低于 95%，写入类工具需单独统计	全场景
检索指标	Recall@K	建议不低于 80%，并按知识库类型分层校准	RAG
安全指标	Guardrail 命中率	监控趋势与异常波动，区分误拦截和真实风险	全场景
性能指标	TTFT	交互式场景建议控制在 2 秒以内	在线推理、Agent 对话
性能指标	TGS	建议结合模型类型设定生成速率基线，例如 30 tokens/s 作为观察口径	在线推理
成本指标	单任务 Token 消耗	建议控制在预算上下 20%波动区间内	全场景

来源：中国信息通信研究院、阿里云计算有限公司联合编制

4.AI 应用服务稳定性成熟度模型

AI 应用服务稳定性建设需要分阶段推进。对于大多数企业而言，不应一开始就追求全自动自愈，而应先保证基础可用，再逐步建设可观测、可恢复、可回归和可自愈能力。

AI 应用服务稳定性成熟度模型

先保证基础可用，再逐层建设可观测、可恢复、可回归与自愈能力——从人工经验驱动走向数据与自动化驱动。



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 9 AI 应用服务稳定性成熟度模型

表 4 成熟度模型

成熟度等级	能力阶段	核心能力	建设目标
L1	基础可用	超时、重试、限流、错误码处理、基础监控	避免单次请求失败直接暴露给用户
L2	可观测可恢复	链路追踪、告警、模型 Fallback、版本回滚、故障分级	缩短故障发现与恢复时间
L3	平台韧性	多租户配额治理、容量水位、区域切流、客户可观测、独享资源隔离	降低租户间干扰与区域性故障影响
L4	Agent 工程化稳定	任务状态机、工具沙箱、上下文治理、断点续跑、评测回归、权限审批	保障长链路智能体任务稳定完成
L5	持续自愈	线上 Trace 回流评测集、质量漂移检测、自动诊断、策略回放、复盘自动化	让稳定性治理从人工经验驱动走向数据与自动化驱动

来源：中国信息通信研究院、阿里云计算有限公司联合编制

成熟度模型可作为企业建设 AI 应用稳定性的路线图：L1 建设超时、重试和限流等基础能力；L2 补齐链路追踪（Trace）、告警、模型故障切换（Fallback）和回滚；L3 完善多租户隔离、容量治理和区域容灾；L4 重点治理 Agent 长链路、工具执行和评测回归；L5 通过线上数据回流、自动诊断和策略优化形成持续闭环。

（二）应用服务稳定性实践，保障服务连续性

1. 备选模型 Fallback 与多模型切换

单一模型可能发生限流、超时、内容安全拦截、版本异常或区域

不可用。模型 Fallback 用于在主模型异常时，自动切换至能力相近或业务可接受的备选模型，降低业务中断风险。AI 网关应建立主备模型映射、健康检查和切换阈值。主模型连续超时、返回 5xx 错误、触发限流或健康检查失败时，自动切换至备选模型。

表 5 多模型切换机制

环节	关键机制	触发信号	稳定性目标
健康探测	持续检查主备模型健康度	成功率、5xx/429、TTFT 异常	尽早识别模型不可用或退化
降级切换	切换能力相近的备选模型	连续超时、限流或健康检查失败	降低故障暴露时间
兜底体验	缓存、标准化响应或人工接管	主备模型均不可用	维持最低可用服务

来源：中国信息通信研究院、阿里云计算有限公司联合编制

高价值场景应提前评估备选模型的任务完成率、输出格式、工具兼容性和安全能力。Fallback 还应配合指数退避重试、任务拆分和批量调用接口（Batch API），分别应对短时限流、大 Token 请求和非实时任务。关键指标包括异常期间请求成功率、Fallback 命中率、切换耗时、输出一致性和用户感知失败率。

2.发布灰度与版本回退

提示词（Prompt）、工具、知识库和模型变更，都可能引发任务路径偏移、输出不稳定、成本上升或幻觉率增加。智能体开发平台应管理草稿、线上和历史版本，记录版本 ID、发布人、时间和变更内容，并支持异常时快速回退。发布前应自动评测任务完成率、工具调用成

功率、平均执行步数、格式正确率和内容质量，指标下降超过阈值时阻断发布。重要活动期间，可冻结高风险变更，并对重点客户安排专项监控。关键指标包括评测通过率、回滚次数与耗时、灰度异常发现率和变更故障占比。

3. 区域切流

当区域故障、容量不足或上游依赖异常时，平台应通过全局流量调度，将请求切换至备用区域。切流前需核查备用区域容量、模型可用性、数据合规性和客户路由策略，避免流量集中转移引发新的拥塞或级联故障。

4. 全链路排查、故障分级与处置闭环

MaaS 平台应以请求标识符（`request_id`）为主线，结合日志、链路追踪（Tracing）和监控大盘，依次判断请求是否到达网关、是否触发鉴权或限流、是否进入排队、推理耗时是否异常，以及影响范围属于单请求、单客户、单模型、单区域还是全局。高频错误码应形成标准处置表，明确错误含义、影响范围、监控指标和处理动作。

表 6 监控指标

返回码	错误码	含义	监控指标	处置策略
429	Throttling.AllocationQuota	TPM/TPS 限流	TPM/TPS 水位、429 数量	申请扩容或调整限流统计周期
429	Throttling.RateQuota	RPM 限流	RPM 水位、请求突增率	申请扩容、引导削峰或拆分任务
429	Throttling.BurstRate	增速限流	五分钟 Token 增速	引导客户平稳调用，必要时调整增速阈值
500	InternalError / InternalError.Algo	服务内部错误	5xx 率、模型维度失败率	偶发引导重试，持续出现升级值班处理
400	DataInspectionFailed	内容安全拦截	护栏命中率、误报率	反馈客户确认输入，必要时调整安全策略
400	InvalidParameter	参数不合法	参数错误率、模型维度错误分布	引导客户按目标模型检查参数与资源可访问性

来源：中国信息通信研究院、阿里云计算有限公司联合编制

故障可按四类分级：S 序列表示客户业务影响，P 序列表示平台影响范围，F 序列表示客户暂时无感但存在风险，X 序列表示客户自身原因。多客户受影响时按最高等级处理。

重大或面积性故障应及时升级。恢复后，重大故障宜在 24 小时内复盘，一般故障宜在 7 个自然日内复盘，并将结论更新至告警规则、容量模型、标准操作流程（SOP）和知识库。

表 7 常见错误分类与处置机制

错误类别	典型错误码	常见原因	处置动作
客户端错误	400/参数非法	请求参数或格式错误	参数校验并返回明确提示
认证/限流	401/429	鉴权失败或超配额	刷新凭证、排队或限流重试
服务端错误	500/503	模型或平台服务异常	模型 Fallback、隔离和告警
超时类	504/网关超时	排队或链路耗时超限	熔断、降级和区域切流

来源：中国信息通信研究院、阿里云计算有限公司联合编制

MaaS 故障治理可归纳为：“异常监测—影响判断—故障分级—服务降级—模型 Fallback—区域切流—恢复验证—复盘沉淀。”

（三）Harness 稳定性工程实践，构建通用智能体运行底座

Harness 位于基础模型与业务应用之间，负责 Agent 的执行循环、上下文管理、工具调用、错误恢复、安全权限、记忆、可观测和评测回归。基础模型提供推理能力，Harness 提供运行约束，二者共同构成智能体系统。不同 Harness 会直接影响同一模型的任务完成率和恢复能力。AI 辅助编程（AI Coding）中的循环控制、工具沙箱、上下文治理和阶段成果保存，也可复用于办公助手、数据分析、智能运维和企业客服等场景。

1.Harness 六类核心功能组件

通用 Agent Harness 包括六类组件：执行循环：驱动“推理—行动—观察—状态更新”；工具系统：支持文件、命令、知识和外部服务调用；上下文管理：控制信息保留、压缩和删除；错误恢复：将模型、工具和环境异常转化为可处理状态；安全机制：划分自动执行、人工确认和禁止执行边界；记忆系统：保存跨会话的用户偏好、项目结构和历史决策。

表 8 Harness 核心组件

功能组件	核心作用	主要稳定性风险	关键指标
执行循环	驱动 Agent 多步执行	死循环、无效重复、过早终止	任务完成率、平均步数、异常终止率
工具系统	连接外部服务与执行环境	工具超时、返回异常、非幂等重试	工具调用成功率、工具超时率
上下文管理	控制信息进入、压缩与丢弃	上下文溢出、状态污染、噪声累积	上下文截断率、格式解析成功率
错误恢复	支持重试、降级和断点续跑	局部失败扩大为任务失败	恢复成功率、重复执行成本
安全权限	控制自动化边界	越权执行、高危操作误触发	审批命中率、越权拦截率
记忆系统	持久化短期状态与长期偏好	记忆污染、租户串扰、状态丢失	记忆命中率、隔离违规率

来源：中国信息通信研究院、阿里云计算有限公司联合编制

通用 Agent Harness 六子系统概览



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 10 通用 Agent Harness 六子系统概览图

2.Agent Loop 编排的稳定性约束

Agent 执行循环（Agent Loop）的典型问题包括死循环、无效重复、过度分支和过早终止。编排框架应设置最大执行轮次、上下文长度、分支上限和终止条件。不同智能体平台的参数范围和默认值存在差异。阿里云百炼的 ReAct 最大轮次可配置为 1-50 轮，短期上下文可配置为 0-30 轮，具体取值需根据生产环境实际负载确定。OpenClaw 采用动态迭代预算，范围为 32-160 次，内置 genericRepeat、knownPollNoProgress、pingPong 三种循环检测器，并采用 warningThreshold（默认 10 次）、criticalThreshold（默认 20 次）、globalCircuitBreakerThreshold（默认 30 次）三级熔断机制；上下文管理需显式设置 contextWindow，contextTokens 建议设为 contextWindow 的 75%-85%，压缩预留空间 reserveTokensFloor 建议

为 20000-40000tokens；多 Agent 编排支持 maxSpawnDepth、maxChildrenPerAgent、maxConcurrent 等分支控制参数。Hermes Agent 主 Agent 默认迭代上限约 90 次（可通过 agent.max_turns 或 HERMES_MAX_ITERATIONS 配置），子 Agent 默认约 50 次（受 delegation.max_iterations 控制），连续 3 次无效工具调用后自动终止；上下文窗口有 64000 tokens 的硬性下限要求，采用双层压缩机制，Gateway 层在上下文达 85%时触发作为兜底，Agent 内部层在 50%时触发作为主力；多 Agent 并发默认上限为 3，且禁止子 Agent 递归调用自身。。

Agent Loop: 可控、可终止、可预算



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 11 Agentloop

连续多轮调用相同工具且任务无进展时，应触发策略调整、澄清提问或人工介入。短链路任务可采用推理一行动（ReAct），复杂长

任务可采用规划后执行（Plan-and-Execute）。关键指标包括任务完成率、平均执行步数、重复动作命中率、最大轮次触达率、过早终止率和单任务 Token 成本。

3. 上下文与状态治理

长任务中的历史消息、工具输出和检索结果不断累积，容易引发上下文溢出和状态污染，导致任务目标丢失或错误信息进入后续 Prompt。

系统应优先保留系统指令、任务目标、关键约束和最近工具结果，压缩或删除重复历史和低相关内容。Qwen-Agent 可通过 `max_input_tokens` 控制输入长度，Qwen3-Coder 可通过 `use_raw_api` 和 `thought_in_content` 降低格式混淆。

上下文与状态治理：Checkpoint 保障长程可靠



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 12 上下文与状态治理

中间思考 (Thought)、工具输入输出 (Tool I/O) 和最终结果 (Result) 应分别存储。工具数据可作为结构化调用片段 (Span) 保存，支持审计和回放。短期状态可存入 Tair，工作区文件可存入网络附属存储或对象存储 (NAS/OSS)，长期记忆可存入 Lindorm Memobase。长任务应定期写入检查点 (Checkpoint)，记录执行步骤、阶段成果和关键变量，以支持断点续跑。关键指标包括上下文截断率、工具解析成功率、Checkpoint 写入成功率、续跑成功率、恢复耗时和重复 Token 成本。

4. 工具执行沙箱与权限隔离

外部 API、浏览器、代码环境和依赖服务可能出现超时、格式变化、崩溃、资源耗尽或权限异常。Harness 应将工具统一纳入隔离沙箱，限制单一工具故障的影响范围。

工具治理主要包括三类分级。一是超时分级：查询工具采用秒级超时，代码和数据任务采用分钟级超时，批量任务异步执行；二是重试分级：幂等工具可自动重试，非幂等工具需人工确认；三是权限分级：只读工具可自动执行，写入工具需二次确认，高风险工具进入强隔离沙箱，并配置资源配额、网络白名单和审批策略。

表 9 工具治理分级建议

风险等级	典型工具	权限层级	管控策略	失败补偿
低危只读	文档读取、检索、状态查询	Read	秒级超时、自动重试、语义缓存	自动重试或切换备用读取源
中危写入	写文件、更新记录、创建任务	Write	分钟级超时、Diff 对比、自动备份、人工确认	暂停任务并等待确认
高危执行	执行命令、代码运行、Git 提交、资金操作	Execute	强沙箱隔离、资源配额、网络白名单、审批	转人工处理或进入死信队列

来源：中国信息通信研究院、阿里云计算有限公司联合编制

关键指标包括工具调用成功率、超时率、沙箱崩溃率、越权拦截率、人工确认率和非幂等误重试次数。

5.错误恢复与降级处置

Harness 应将执行错误转化为可识别、可处置、可恢复的工程状态。模型错误可通过格式校验、重采样和模型切换处理；工具错误可通过重试、备选工具和人工确认处理；环境错误可通过实例重建、断点续跑和死信队列处理；业务错误则通过澄清、拒答或人工接管处理。

错误四分类与恢复策略



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 13 错误恢复与降级处置

（四）智能体应用实践，增强业务任务执行韧性

智能体应用层应重点解决任务恢复、阶段成果保存、业务幂等、工具补偿、多 Agent 协作和多租户数据隔离等问题。

1. 任务状态机与业务幂等

复杂任务应建模为状态机，划分待执行、执行中、等待确认、完成和失败等状态，并持久化状态变化、工具结果和决策依据。工具中断后可从最近成功状态恢复；写入类操作应设置幂等键，防止断点续

跑时重复创建工单、发送消息或修改数据。

关键指标包括状态转换成功率、重复执行率、幂等冲突拦截率和任务恢复耗时。



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 14 任务状态与工具健康检查

2.工具健康检查与失败补偿

智能体依赖模型上下文协议（Model Context Protocol，MCP）工具、知识库、审批系统和企业连接器等外部服务。应用侧应监测工具可用性、响应时延、返回格式和权限状态，并在规划阶段避开异常工具。同类工具可设置主备关系，查询类工具可使用语义缓存。失败处置采用“重试—备选工具—人工确认”三级策略：幂等工具可自动重试，非幂等工具需暂停并确认，主备工具均不可用时进入人工处理或异步补偿队列。关键指标包括健康检查通过率、工具切换成功率、缓

存命中率和人工介入率。

3.长任务断点续跑与阶段制执行

持续数分钟至数小时的企业任务应采用异步和阶段制执行。各阶段需明确输入、输出和完成条件，并在结束后保存成果、更新状态；中途失败时，仅从失败阶段恢复。公共云函数计算通常可通过异步任务、指数退避重试、死信队列和实例预热提供容错能力。

长任务续跑、记忆隔离与多 Agent 协作



来源：中国信息通信研究院、阿里云计算有限公司联合编制

图 15 长任务续跑、记忆隔离与多 Agent 协作

关键指标包括阶段完成率、断点续跑成功率、恢复耗时和重复 Token 成本。

4.记忆持久化与多租户隔离

Agent 记忆分为短期记忆和长期记忆。短期记忆保存会话、工具结果和任务状态，可由 Tair 承载；长期记忆保存用户偏好、业务背景

和历史决策，可由 Lindorm Memobase 承载，并通过 project_id 和 ProfileConfig 实现项目隔离。记忆写入、检索和注入提示词（Prompt）的全过程，应使用租户、项目和用户隔离键。敏感信息、临时状态和错误中间结果不应写入长期记忆，并应通过生存时间（Time to Live, TTL）和删除机制及时清理。关键指标包括记忆命中率、误写率、租户隔离违规率、召回时延和用户纠错率。

5. 多 Agent 协作与人工确认边界

多 Agent 协作容易产生角色冲突、资源竞争、消息乱序和状态不一致。系统应明确规划 Agent、执行 Agent 和审查 Agent 的职责、工具权限及资源范围，并通过统一状态机、资源锁、版本控制和消息幂等协调操作。Agent 间通信可通过消息队列或事件总线完成，失败消息进入死信队列。主管 Agent (Supervisor) 负责检查执行 Agent (Worker) 状态，对无进展、重复动作或路径偏移进行终止、重新规划或人工升级。高风险操作、不可逆写入、权限变更、外部消息发送和资金操作必须经过人工确认。确认信息应说明操作对象、影响范围和可回滚性；超时未确认时，应暂停任务或降级为只读输出。关键指标包括协作冲突次数、资源锁等待时间、高风险操作审批率、人工确认超时率和写入回滚次数。

六、全栈可观测能力体系

（一）全栈可观测的技术地基

全栈可观测应贯通基础设施、模型和应用三层数据，形成从异常

发现、关联分析到协同处置的治理闭环。大语言模型（Large Language Model, LLM）应用通常涉及终端、认证、会话、编排、模型路由、外部工具、向量数据库和缓存等组件。为关联复杂链路、保障服务等级协议（Service Level Agreement, SLA）和用户体验，需要建设统一数据语义、低成本采集、上下文透传及链路分析能力。

1. 标准化的数据语义规范

链路数据可按照用户（User）、会话（Session）、链路（Trace）和调用片段（Span）组织，一个 Trace 由多个 Span 组成。Span 类型包括编排链（CHAIN）、检索（RETRIEVE）、向量化（EMBEDDING）、重排（RERANK）、模型调用（LLM）、工具调用（TOOL）和智能体（AGENT）等。Span 属性应记录输入输出、提示模板、模型参数、Token 用量、工具调用及检索文档等信息，用于分析模型性能、调用异常、检索质量和 Token 成本。数据规范应与开放遥测标准（OpenTelemetry）的生成式 AI 语义约定保持一致。

2. 低成本、高质量的数据采集

数据采集应兼顾低侵入、低开销和高覆盖，可基于 OpenTelemetry 对不同语言应用自动埋点。Python 探针可采集模型参数、Token 用量、首 Token 时延（Time to First Token, TTFT）和单 Token 生成时延（Time per Output Token, TPOT），并适配 LangChain、LlamaIndex 和 Dify 等框架；Java 探针可通过字节码增强关联远程过程调用（Remote Procedure Call, RPC）与 LLM 链路；Go 探针可通过编译期插桩自动

采集。在 Kubernetes 环境中，可通过配置文件或初始化容器（initContainer）部署探针，并利用采样、指标预聚合和动态配置控制开销。

3. 标准化协议透传

链路上下文应兼容 OpenTelemetry W3C 协议，贯通用户终端、应用程序接口（Application Programming Interface，API）网关、模型应用、外部依赖和模型服务，实现跨语言、跨组件的完整 Trace 关联。

4. 链路采集与加工

数据侧应建立统一的采集、加工和存储链路。终端及后端应用可直接上报 Trace，网关等组件可将包含 Trace ID 的日志转换为 Trace。在服务端事件（Server-Sent Events，SSE）流式输出场景中，可采用“客户端分段采集、服务端合并还原”的方式，分段上报事件，并在调用结束后记录 TTFT、Token 用量等数据，以降低客户端压力并保留完整链路。

5. LLM Trace 查询与分析

平台应支持通过 Trace ID 还原请求链路，展示模型输入输出、提示词（Prompt）、模型参数、Token 用量和各 Span 耗时，并按状态、耗时和 Span 类型进行筛选。还可利用大模型分析 Trace，识别模型生成、工具调用或下游服务中的异常节点，并提出超时控制、异步执行和链路优化建议。端到端全链路追踪是全栈可观测体系的基础，可通

过统一 Trace 和数据语义，贯通 AI 基础设施层、模型服务层和应用服务层。

（二）AI 基础设施层可观测

1. 定位与目标

AI 基础设施层承载 GPU 及其他加速卡、智算服务器、高速网络、存储、容器和 Kubernetes 调度等资源，上层模型服务的性能与稳定性最终都依赖该层的资源供给。基础设施层可观测主要回答“算力、显存、网络、存储和调度是否成为瓶颈”，为性能优化与容量规划提供依据。与传统基础设施即服务（IaaS）和容器即服务（CaaS）监控相比，该层需要重点关注 GPU 资源成本高、显存与带宽约束突出、多卡通信复杂，以及训练与推理混部产生的资源争抢等问题。

2. 关键观测信号

基础设施层应围绕“算力—显存—通信—存储—调度”开展观测：
算力与显存：GPU 利用率、显存使用率和显存带宽利用率（Memory Bandwidth Utilization, MBU），用于识别内存耗尽（Out of Memory, OOM）及生成性能风险；通信：高速外设互联（Peripheral Component Interconnect Express）、GPU 高速互联（如 UALink、NVLink 等）及远程直接内存访问（Remote Direct Memory Access, RDMA）网络的带宽、时延和丢包；存储：模型权重加载及键值缓存（KV Cache）换入换出的吞吐和时延；调度：Kubernetes 工作负载（Workload）、控制面、流量入口（Ingress）、GPU 调度事件和节点资源状态。

表 10 基础设施层关键指标

观测维度	关键指标	推理阶段关注度	异常影响
显存带宽	显存带宽利用率 (MBU)	高 (Decode)	直接限制生成速度
算力	GPU 利用率	中	计算资源是否饱和
显存	显存利用率	高	OOM 风险
通信	PCIe/NVLink 带宽	多卡场景高	跨卡通信瓶颈
网络	RDMA 吞吐/延迟/丢包	分布式场景高	跨机通信瓶颈
存储	高速存储读写吞吐/延迟	中	权重加载、KV 换出变慢
调度	节点资源异动、调度事件	中	资源争抢、Pod 驱逐

来源：中国信息通信研究院、阿里云计算有限公司联合编制

3.数据采集方式

基础设施层以指标 (Metrics) 采集为主，辅以事件和日志。GPU 及加速卡指标可通过节点采集器暴露至 Prometheus 兼容的时序数据库；Kubernetes 侧采集控制面指标、事件和日志；网络、存储则通过相应设备或服务的 Exporter 采集。建议统一采用 Prometheus 与 OpenTelemetry Metrics 标准，使基础设施指标能够与模型服务和应用服务数据进行跨层关联。

4.可视化、告警与实践

可视化宜采用“集群总览—节点明细—单卡或单 Pod 下钻”的分层结构。顶层展示 GPU 利用率、显存水位、网络与存储吞吐等整体

状态，下层支持定位具体节点、GPU 和容器实例。告警应聚焦直接影响推理服务的信号。例如，显存长期接近上限时触发 OOM 预警，GPU 持续饱和时提示扩容，NVLink 或 RDMA 带宽打满时提示通信瓶颈。当模型推理耗时升高时，可先通过调用链定位推理阶段，再下钻检查 GPU 显存和带宽。若显存仍有余量，但推理引擎的显存分配参数设置较低，可适当调整；若 GPU、网络或存储资源已饱和，则应优先从扩容、调度或架构层面处理。

（三）模型服务层可观测

1. 定位与目标

模型服务层承载“把模型跑起来并对外提供推理能力”的职责，涵盖三类典型形态：模型即服务（MaaS，调用托管的公共大模型 API）、托管模型服务（在平台上部署自有模型）、自建模型服务（在自有 GPU 集群上通过推理引擎部署）。常见的推理引擎有 vLLM、SGLang 等。

模型服务层可观测的目标，是回答“推理性能慢在哪里、资源够不够、调用是否稳定、输入输出是什么”。它是连接底层算力与上层应用的桥梁，也是成本与性能博弈的主战场。

2. 理解推理引擎的内部结构

以 VLLM 为例，现代推理引擎通常包含三类组件：API 服务进程负责 HTTP 请求接入和文本分词；引擎进程负责 KV Cache 管理、请求调度和批处理；GPU 工作进程负责加载模型权重并执行模型计算。

请求进入后，通常经历预填充（Prefill）、排队和解码（Decode）三个阶段。Prefill 负责处理输入提示词并生成首个 Token 所需的 KV Cache；Decode 逐个生成后续 Token；调度排队则决定请求何时获得 GPU 资源。三部分共同影响用户感知时延。

3.关键观测信号

建议从 API Server、模型输入输出、推理过程、推理引擎状态四个维度组织信号。API Server 维度记录请求的路径（如 /v1/completions）、状态码（200/400）、客户端类别、请求处理耗时。模型输入输出维度记录提示词、响应方式（Stream/None）、模型名称、最大 Token 数，以及采样参数 Temperature、Top-K、Top-P、N（返回答案数量）。这些既是质量分析的输入，也是成本归因的基础。

推理过程维度是性能诊断的核心，关键指标如下表。

表 11 推理过程维度关键指标

指标	含义
E2E 时间	推理总时间，从接收请求开始到推理完成
首 Token 时间（TTFT）	从输入 Prompt 到生成第一个输出 Token 的时间，约等于“等待时间+Prefill 时间”，直接影响实时交互体验
Prefill 时间	Prefill 阶段耗时
Decode 时间	Decode 阶段耗时
等待时间	请求在调度器中排队的时间
Token 间隔时间（TPOT）	Decode 阶段每个输出 Token 的延时，衡量生成速度

来源：中国信息通信研究院、阿里云计算有限公司联合编制

推理引擎状态维度反映引擎整体负载与健康度，关键指标如下表。

表 12 推理引擎状态维度关键指标

指标	含义
运行中请求数	正在推理的请求数量
等待请求数	等待被调度执行的请求数量
被抢占请求数	当 KV Cache 不足时被调度出 GPU、释放显存的请求数量
KV Cache 使用率	较高的使用率可能导致性能下降
总提示词 Token 数	所有请求的提示词 Token 总数
总生成 Token 数	所有请求生成的 Token 总数
处理成功请求数	已成功处理的请求总数

来源：中国信息通信研究院、阿里云计算有限公司联合编制

推理性能的核心目标可概括为：最快的首个 Token 响应时间（TTFT）、最高的吞吐量（Throughput）、最快的每个输出 Token 时间（TPOT）。

4.数据采集方式

API 服务、模型输入输出和推理过程可分别创建 Span，并将模型参数、Token 用量和耗时等信息记录为属性，从而与应用调用链关联。

推理引擎内部指标则单独上报至 Metrics 系统。自建 vLLM、SGLang 服务可采集其原生 Prometheus 指标；托管或 MaaS 服务则依赖平台提供的调用日志、审计日志、模型指标和资源监控数据。

5.可视化、告警与实践

仪表盘应同时展示推理性能、资源水位、请求日志和推理链路，并支持按模型、区域和实例下钻。当请求耗时升高时，可先通过 Request ID 关联 Trace，定位模型推理阶段，再依次检查 TTFT、TPOT、正在执行的请求数和等待请求数。若 TTFT、TPOT 正常，但排队请求明显增加，通常说明并发已超过当前处理能力，可通过增加 GPU、扩容实例或优化调度进行处理。TTFT 升高还可能与输入提示词过长、并发排队或 KV Cache 使用率过高有关，应结合提示词长度、显存水位和等待队列综合判断。

（四）应用服务层可观测

1. 定位与目标

应用服务层是最贴近用户的一层，承载 LLM 应用、编排框架（如 Dify/LangChain/LlamaIndex 等）、Agent 与 MCP 工具、RAG 检索、缓存与向量数据库等。这一层的行为最复杂、决策路径最不确定，也最能体现“AI 原生”的特征——Agent 自主规划、多步推理、动态调用工具。应用服务层可观测的目标，是回答“Agent 的决策与工具选择是否合理、错误出在哪一步、Token 花在了哪里、整条链路慢在何处”。

2. AI 原生应用开发的核心痛点

应用侧常见问题包括：Agent 选择了不合适的工具；参数格式或类型错误难以定位；多轮工具调用造成 Token 成本快速增长；Agent 重复调用相同工具并陷入循环。MCP 为模型与外部工具建立了统一交互协议，降低了工具接入成本，但也增加了模型、宿主环境和工具

服务器之间的链路复杂度，需要通过可观测能力控制其性能、稳定性和安全风险。

3.应用服务层应具备的观测能力

应用服务层宜具备以下能力：无代码或低代码接入；可视化展示 Agent 的工具选择与执行过程；通过调用链定位异常步骤；按应用、用户和工具分析 Token 成本；完整追踪“用户—Agent—MCP 工具—外部服务”的调用链。

4.关键观测信号

应用服务层应基于本章前述 Trace 语义记录以下数据：

编排与 Agent 维度，记录执行结构、各步骤输入输出、耗时和工具选择结果，并识别循环调用、过度分支等异常模式。工具与 MCP 维度，记录工具名称、调用参数、输出大小、传输协议、服务地址、状态码和耗时，并统计请求量、错误量和耗时（Rate、Error、Duration、RED）等指标。RAG 维度，记录文档分片、相关性得分和召回质量，用于识别错误召回导致的答案偏差。成本维度则按应用、用户和模型统计 Token 用量，定位高消耗环节。

5.数据采集与一个端到端示例

可通过无侵入探针串联“用户请求—编排框架—大模型—MCP 工具—外部服务”，并在 Trace 中展示工具名称、参数、输出、地址、协议和 RED 指标。

6. 可视化、告警与实践

平台应提供调用链、拓扑、会话和场景分析视图，并支持从 Trace 下钻至 Span，或聚合分析 Token、工具调用和错误趋势。结果异常时，可沿 Trace 检查检索、工具和模型生成过程。例如，若低相关性文档导致错误回答，应将问题定位至 RAG 检索环节，而非直接归因于模型。

（五）贯穿三层的横切能力

三层可观测体系并非相互独立，而应通过若干横向贯通能力实现协同联动，充分发挥**整体效能**。

1. AI 网关：治理与观测的统一入口

AI 网关承担模型路由、鉴权、限流、缓存和多模型切换，也是请求数据的统一采集入口。网关应监测每秒请求数（Queries Per Second, QPS）、成功率、时延、TTFT、Token 消耗、缓存命中、限流、模型切换和内容安全事件，并通过应用程序接口密钥（API Key）或 JSON Web 令牌（JSON Web Token, JWT）识别租户。Metrics 用于记录 QPS、时延、成功率和 Token 用量等聚合指标；Logging 用于记录请求、响应、状态码、模型、租户和缓存命中等明细。

2. 自动化评估：为质量与安全兜底

自动化评估可基于 Trace 或调用日志，利用独立模型评估正确性、幻觉、有害内容、合规性、用户情绪和意图。评估结果应与应用、模

型、版本和具体请求关联，用于质量量化、告警和追踪，并支持按业务规则配置评估模板。

3. 智能告警与治理闭环

全栈可观测应形成“监控—分析—优化—再监控”闭环。平台可针对成功率下降、Token 消耗突增和模型切换失败设置分级告警，并联动扩容、限流或主备模型切换。平台还应定期生成资源、成本、质量和安全报告，为配额调整、模型替换和提示词（Prompt）优化提供依据。可观测助手（Copilot）还可辅助日志查询、Trace 诊断、异常巡检和根因分析，缩短平均修复时间（Mean Time to Repair, MTTR）。

七、发展趋势与建设建议

随着大模型应用规模化落地，公共云 AI 推理稳定性正由单点保障转向体系治理，主要呈现技术智能化、行业分级化、标准协同化和运营常态化四类趋势。

（一）技术智能化：推动稳定性能力自治演进

公共云 AI 推理稳定性正沿“监测—检测—诊断—预测—自愈”演进。微软对 156 起大语言模型（Large Language Model, LLM）推理事件的分析显示，约 74% 的事件可自动检测，但仅 28% 可热修复，自动恢复能力仍需提升。人工智能运维（Artificial Intelligence for IT Operations, AIOps）正向智能体驱动的服务韧性工程（Agentic SRE）发展，由多个智能体（Agent）协同完成故障检测、分析和处置。可观测范围也由资源指标扩展至模型质量、语义漂移和 Token 性能，混沌

工程开始覆盖显存不足、集合通信库（如 NCCL、ACCL、HCCL 等）中断和键值缓存（KV Cache）污染等 AI 特有故障。

建议：分阶段建设智能告警、自动处置、根因分析、预测预警和多 Agent 自治能力；推进 Agentic SRE、AI 混沌工程、模型质量观测和预测式资源调度。

（二）行业分级化：建立差异化服务保障机制

金融、医疗、政务和工业等行业对时延、准确性、安全性和责任追溯的要求不同。金融强调实时风控和可解释性，医疗关注循证校验，政务强调内容安全和人工复核，工业关注低时延和生产连续性。企业对 AI 服务的要求也由模型能力强转向行为可预测、决策可解释、异常可接管、成本可控制，SLA 将成为重要方向。

建议：建立金融级、企业级和通用级等分级保障体系；提供行业化稳定性方案；向租户开放核心指标和健康报告；沉淀可复用的行业实践指南。

（三）标准协同化：完善评测基准与责任边界

国内外 AI 标准化工作正在加快推进。中国信通院主导的 ITU-T AICP 国际标准体系已覆盖总体架构、模型开发、性能评估、高可用和推理优化等方向，并发布《云上 Agent 基准度量模型》《Agent 可靠性能力要求》等标准。但在多云、多服务商和多应用主体共同参与的场景下，推理质量异常和服务故障的责任边界仍不清晰，事故分级、

质量评价和 SLA 表达也缺乏统一规范。

建议：加快制定稳定性分级、质量评价、故障分类、SLA 表达和多主体责任矩阵；将常态化监测扩展至 P95/P99 时延和故障恢复时间；推动国内外标准衔接；建设公开、可复现的推理质量评测工具链，并通过开源参考实现推动标准落地。

（四）运营常态化：强化演练复盘与持续改进

AI 推理除可用率、延迟和错误率问题外，还存在输出非确定、语义漂移和质量退化等“软故障”，需建立面向 AI 场景的服务韧性工程（AI-SRE）体系。服务等级指标（Service Level Indicator, SLI）应扩展至 TTFT、Token 生成速率、KV Cache 命中率、内容忠实度和安全拦截率，服务等级目标（Service Level Objective, SLO）与 SLA 也应逐步纳入模型质量要求。

混沌演练可分层实施：基础级覆盖 GPU 故障和节点宕机，中级覆盖 KV Cache 污染和上下文溢出，高级覆盖模型质量下降和提示词（Prompt）注入攻击。故障恢复标准也应由“服务启动”提升为“性能和质量恢复至基线”。

建议：组建具备基础设施、模型和应用编排能力的 AI-SRE 团队；建立常态化分级演练机制；完善 SLI—SLO—SLA 体系；建设智能应急平台和 Agent 应用持续评估机制；推动稳定性事件匿名共享。

总体来看，公共云 AI 推理稳定性将向 Agentic SRE、行业分级保

障、标准协同和持续运营演进，逐步形成覆盖技术、管理和生态的稳定性治理体系。

CAICT 中国信通院

编制说明

本研究报告自 2026 年 6 月启动编制，由中国信息通信研究院云计算与数字化研究所牵头，联合阿里云计算有限公司共同完成。在前期研究、框架设计、文稿起草、征求意见和修改完善等五个阶段中，编制组围绕公共云 AI 推理服务稳定性建设这一核心议题，结合阿里云计算有限公司的工程实践与产业经验，依托中国信息通信研究院在云计算与人工智能可靠性领域的产业研究积累，系统梳理了公共云 AI 推理服务稳定性建设的技术框架、关键能力与典型实践，经多轮研讨修改完善后形成本报告。

牵头单位：中国信息通信研究院云计算与数字化研究所

联合单位：阿里云计算有限公司

中国信息通信研究院 云计算与数字化研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301311

传真：010-62301311

