CAICT 中国信通院

集智 专题报告

云上人工智能可靠性建设指南

（2026 年）

中国信息通信研究院云计算与数字化研究所

2026年7月

版权声明

前　言

随着人工智能技术加速融入千行百业，云上 AI 系统已从辅助工具演变为驱动业务创新与决策的核心基础设施。大模型、生成式 AI 与智能体（Agent）的广泛应用，在释放巨大生产力的同时，也带来了前所未有的可靠性挑战：模型输出的不可控性、数据漂移引发的性能退化、分布式训练中的故障频发、推理服务的高并发压力，以及伦理安全与合规风险的交织叠加，使得传统以“功能正确”为中心的工程范式难以为继。在此背景下，构建一套覆盖全生命周期、贯穿技术与治理的云上人工智能可靠性体系，已成为保障 AI 系统可用、可信、可控的关键前提，也是企业实现智能化转型可持续发展的基石。

云上 AI 的可靠性建设远非单一技术模块的优化，而是一项涉及基础设施、数据、模型、推理与应用多层协同的系统工程。为此，本指南系统梳理了当前云上 AI 可靠性面临的各类运行风险，提出以“高可用性、高解释性、高鲁棒性、高可恢复性”为可靠性建设目标的建设框架，并分层阐述基础设施、数据、模型、推理、AI 应用等关键域的能力清单与实施要点，旨在为企业提供一套可落地、可度量、可演进的实践路线图。

展望未来，AI 系统的可靠性将不再仅是运维或算法团队的责任，而是深度融入架构设计、开发流程与治理机制的原生能力。随着 Agent 自主能力持续提升、跨行业场景复杂度增加，可靠性建设必将向动态化、场景化、标准化方向演进。

目 录

图 目 录

表 目 录

一、背景与总体概述

（一）云上人工智能可靠性的内涵

随着人工智能技术加速融入云计算基础设施并深度嵌入核心业务流程，以“系统可用性”为核心的可靠性范畴不足以全面覆盖智能系统运行中的新型风险与保障需求。在此背景下，云上人工智能可靠性呈现出多维度、全链条、强耦合的新特征，其内涵亟须在继承传统可靠性工程方法论的基础上进行系统性拓展与重构。

本报告认为，**云上人工智能可靠性是指在保障云平台基础设施高可用、高弹性和高安全性的前提下，保障人工智能系统在其全生命周期内持续输出满足准确性、公平性、鲁棒性且符合业务目标的可信结果的能力。**该定义突破了软件系统“服务不中断即可靠”的局限，将可靠性范畴从系统层面向数据、模型、推理及业务价值等多维度延伸。具体而言，云上人工智能可靠性包含以下三个关键层面：

一是系统可靠性。服务韧性工程（Site Reliability Engineering，SRE）的核心要求，聚焦云服务环境下 AI 系统的高可用架构、弹性扩缩容机制、故障快速恢复能力及资源调度效率，确保推理与训练任务在复杂分布式环境中稳定运行。

二是模型可靠性。针对 AI 模型概率性输出的本质特性，强调模型在面对数据分布偏移、对抗扰动、概念漂移等挑战时的性能稳定性与行为可预期性。重点防范“模型静默退化”“灾难性遗忘”“幻觉输出”等新型故障模式，保障模型精度、公平性、可解释性等关键指标持续满足预设阈值。

三是业务可靠性。将技术指标与业务成效紧密对齐，通过构建涵盖精确率、召回率、延迟、偏差范围等要素的“准确性服务等级协议”（Accuracy SLAs），实现从系统可用到业务可信的闭环管理，确保 AI 输出切实支撑决策质量与用户体验。

云上人工智能可靠性的内涵体现了从“保障系统在线”向“保障智能可信”的范式演进，是构建安全、可控、高效智能基础设施的重要基石。

（二）云上人工智能可靠性工程的核心价值

云上人工智能可靠性工程的核心价值，在于**系统性地保障由人工智能驱动的业务价值能够持续、稳定且可信地交付。**具体体现在保障业务连续性、模型稳定性以及构建规模化可信应用这三个方面。

业务连续性是云上人工智能可靠性的基础保障，它继承并扩展了传统 IT 系统可靠性的核心要求。其目标在于保障承载 AI 能力的基础设施、平台与服务在面临硬件故障、流量洪峰、网络抖动或区域性灾难时，依然能够对外提供不中断的服务响应，解决了 AI 系统“能否工作”的根本问题。

模型的稳定性关注的是 AI 系统“工作是否正确与可信”。与逻辑确定的传统软件不同，AI 模型的行为基于概率与数据驱动，其输出质量并非恒定不变。模型稳定性要求模型在整个生命周期内，其性能、行为与决策逻辑能够保持一致、可预测且鲁棒，确保其关键性能指标（如准确率、召回率等）能够持续维持在业务可接受的范围内。

构建规模化可信应用是在治理与合规方面让 AI 应用获得信任，

一个具备端到端可靠性的 AI 系统，能够向业务方、监管机构与终端用户证明其可预测、可解释、可审计、可追责，从而建立组织内外的信任共识。

云上人工智能可靠性工程的价值，正在于将传统的可靠性边界从基础设施与服务层，向上延伸至模型逻辑与决策行为层，实现对整个智能系统的端到端价值保障。

（三）应用边界与整体发展方向

1.应用范围

本指南适用于部署在云环境中的各类人工智能系统，涵盖从模型服务、AI 平台到 Agent 系统的全栈场景。其内容面向参与人工智能系统设计、建设、运维与治理的各类组织与工程团队，旨在为云上 AI 系统提供系统化、工程化的可靠性建设指导框架，帮助构建具备高可用性、高可解释性、高鲁棒性与高可恢复性的智能服务能力。

模型服务是指将人工智能模型以标准化接口对外提供可访问、可调度、可运维能力的服务化组件。它封装了模型的运行环境与计算逻辑，是模型价值在业务中实现的最终载体，常见的模型服务包括模型推理服务、模型微调服务等。

AI 平台是指为人工智能研发与运营提供一体化、规模化支撑的云基础平台，包括模型（开发/训练/评测/数据管理等）及 Agent 系统（构建/编排/管理等）的研发与运营，提供高效、稳定和易用的工程化环境。

Agent 系统是指基于 Agent 技术构建的、具备自主感知、决策与

执行能力的应用系统。它通过调用模型推理服务与工具，将核心 AI 能力转化为可完成复杂任务、与环境持续交互的智能化应用。

2.整体建设思路

为系统性地构建云上人工智能系统的可靠性，本指南确立以下四大核心建设思路，贯穿系统全生命周期，全方位保障整体运行质量与管控效力。

一是高可用性，通过分布式架构设计、弹性调度与故障转移机制，确保系统核心功能持续可用，最大限度减少服务中断与性能劣化。高可用性不仅是基础设施层面的稳定运行，更延伸至模型服务与 Agent 任务层的业务连续性，构成云上 AI 系统韧性的首要保障。

二是高可解释性，构建可追溯、可审计的模型与 Agent 决策过程，使系统行为逻辑对人类与监管机制均可理解。可解释性是实现可信 AI 的关键，既有助于风险识别与责任划分，也为模型改进与决策优化提供依据，确保系统在高复杂度下仍具备透明性与可控性。

三是高鲁棒性，通过异常检测、冗余机制与防御性建模，增强系统应对输入扰动、数据漂移、资源波动与潜在攻击的抗干扰能力。高鲁棒性意味着系统在不确定性与复杂扰动下依旧能够维持合理、稳定的输出，是 AI 系统从“可用”迈向“可信”的关键特征。

四是高可恢复性，建立智能监测、自愈与灾备恢复机制，确保系统在异常或故障发生后能够快速恢复到正常状态。高可恢复性不仅依赖于容灾架构与备份体系，更体现为 AI 系统在自学习、自调整能力下的动态恢复力，使可靠性从被动防御走向主动修复。

综上，这四项建设目标共同构成云上人工智能系统可靠性工程的核心支柱，支撑其在复杂环境下实现长期的安全、稳定与可持续运行。

图 1 云上人工智能可靠性建设指南整体架构图

二、云上人工智能系统的可靠性挑战

（一）基础设施层挑战

基础设施层是支撑人工智能（AI）系统模型训练与推理的关键底座，也是可靠性风险集中和敏感的环节。伴随大模型参数突破千亿级、计算需求指数级增长，计算架构正向分布式、弹性异构协同模式演进。GPU、NPU、CPU、FPGA 及各类国产加速卡通过云化平台实现统一调度与虚拟化管理，在提升灵活性与资源效率的同时，也显著放大了系统不确定性与运行风险，主要体现在四方面：

一是资源异构化调度一致性风险突出。不同计算类型在性能、架构、显存拓扑等方面差异显著，跨架构、跨可用区甚至跨云平台的任务协同成为常态。调度系统在性能、成本与带宽之间的平衡愈加困难，尤其需要应对显存碎片化和显存容量瓶颈（Out-of-Memory，OOM）

问题。算法偏差或资源映射异常易导致计算结果失准、任务中断，系统可靠性已从单节点稳定转向多层协同保障。

二是弹性机制冲击任务状态连续性。动态扩缩容提升了资源利用率，但 AI 训练长周期、强状态依赖，推理高并发、低延迟，两类任务诉求差异显著，混合池调度易引发资源争抢；扩缩容时机与任务时序冲突时，会造成训练断点续训失败、推理延迟飙升，引发性能骤降或任务中断。

三是网络通信瓶颈成为双重约束。在分布式集群中，节点间的高速互联（如无限带宽、NVLink、融合以太网远程直接内存访问（RDMA over Converged Ethernet，RoCE）是性能与可靠性的关键支撑。然而云环境中链路抖动、流量突发、拓扑变化易引发参数同步延迟、带宽退化，直接影响模型收敛与服务响应；跨云、多区域部署下，网络延迟的不确定性会进一步放大系统失稳风险。

四是软硬件协同易引发级联。固件、驱动、容器、虚拟化层等多层技术栈深度耦合，单点故障易触发连锁反应；版本兼容、驱动适配、补丁同步等问题均可能导致任务失败或节点漂移，稳定性保障已从单一硬件可靠转向综合工程体系。

（二）数据层挑战

在云上人工智能系统的全生命周期中，数据与模型层的可靠性是整个系统可靠性的基石。数据质量决定模型性能上限，而模型稳定性直接影响业务应用效果。

1.数据采集阶段

数据采集阶段是决定整个人工智能系统稳定性的首要环节，这一阶段主要面临以下五个方面的核心挑战。

一是模型数据的多源异构性面临实时性保障的挑战。云上智能系统的数据源高度分散，涵盖业务数据库、日志、物联网设备及第三方应用程序接口（Application Programming Interface，API），数据类型与传输协议差异显著。数据流的突发性与不可预测性，对云端数据处理管道的弹性伸缩与实时接入能力构成持续压力。若数据实时性无法保障，将直接导致模型输入特征缺失或过期，在实时决策场景（如风控、欺诈检测等）中引发系统响应延迟或决策失效。此外，数据源端自身的可靠性波动（如数据库主从延迟、传感器异常、第三方 API 服务不稳定等）会直接传导至云端数据处理链路，若采集流程缺乏有效的故障转移与数据补偿机制，将导致数据流中断，进而破坏模型输入数据的完整性，威胁系统服务的连续性。

二是数据漂移与分布变化的持续性监控挑战。数据漂移是导致生产环境中模型性能“静默衰减”的首要诱因。与硬件故障等显性异常不同，输入数据分布的缓慢变化是隐性的。若缺乏对输入数据的持续分布监测与反馈机制，模型将基于过时的数据模式进行推理，其输出准确性会逐渐降低，直至引发业务指标异常，为系统稳定性维护带来持续压力。

三是数据稀疏性、长尾分布与边缘案例覆盖挑战。训练数据在采集阶段常面临稀疏性与代表性不足的问题，难以全面覆盖实际场景中的所有情况，导致模型在应对未充分覆盖数据时表现不佳。数据分布

通常呈现长尾特征，模型若仅依赖主流数据，将难以处理实践中不断出现的边缘场景与新数据特征，从而影响系统的泛化能力和服务鲁棒性。

四是多样化的数据采集策略对数据隐私合规性构成严峻考验。在《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法规约束下，数据采集的合规性已成为系统稳定运行的先决条件。若在数据采集环节未能有效实现匿名化、去标识化或获取有效授权，将埋下合规隐患。此类风险具有隐蔽性，可能导致业务中断、法律追责乃至品牌声誉受损，对系统的长期稳定运营构成直接威胁。特别是在联邦学习等协同计算场景下，数据虽不出域，但其交互过程中的中间结果也可能泄漏敏感数据分布信息。

五是数据传输及供应链安全带来的可靠性挑战。传输中的数据若缺乏加密或密钥管理存在漏洞，将面临被窃取与遭篡改的风险，严重破坏其完整性。更隐蔽的风险源于“数据投毒”攻击，恶意攻击者通过污染训练数据，系统性误导模型决策逻辑。由于此类攻击在数据采集阶段难以察觉，其危害会持续至模型运行阶段，导致模型行为异常，严重影响系统输出的可靠性。同时，依赖第三方数据标注商或开源数据集引入的供应链风险，若数据源的安全审计与准入机制不健全，则问题数据一旦进入训练管道，将给数据的追溯和彻底清除带来极大的挑战。

2.数据预处理阶段

数据预处理环节是将原始数据转化为高质量训练数据的关键步

骤，其流程的稳定性直接决定模型服务的输出质量与迭代效率。

一是高可用标注平台架构的稳定性与协同挑战。标注任务具有明显的突发性与高并发特征，对标注平台的弹性伸缩与高可用能力提出极高要求。平台服务若出现抖动或中断，将导致标注任务阻塞，延缓模型迭代流程，影响业务需求的响应效率。此外，在大规模标注人员协同场景下，任务调度、进度同步与标准一致性等方面的管理瓶颈，也会转化为系统层面的稳定性风险。

二是自动化清洗流水线的质量、效率与泛化能力挑战。依赖人工的数据清洗方式效率低且一致性差，而自动化清洗流水线则面临高维数据下的“噪声识别”与“语义理解”难题。传统基于规则或统计的异常检测方法在高维场景中易失效，难以区分“噪声数据”与“有效但罕见”的长尾样本。过度清洗可能削弱数据多样性，损害模型的泛化能力；错误的清洗规则会系统性地扭曲数据分布，对模型训练产生深远负面影响。

三是静态预处理规则与动态数据分布的适配性挑战。数据预处理规则（如缺失值填充、异常过滤阈值等）通常基于历史经验静态设定。然而，业务环境与用户行为的变化会导致数据分布发生“概念漂移”。静态规则如不能及时调整，其所输出的数据将无法反映真实场景的最新状态，导致模型输入失真，成为系统稳定性的潜在隐患。

四是数据版本管理与预处理过程的可复现性挑战。模型训练与迭代强烈依赖于特定版本的数据集。预处理过程中的参数调整与规则变更必须与生成的数据版本严格对应。缺乏有效的数据版本管理与流程

可复现机制，将导致模型回退时无法准确定位数据根源，显著增加问题排查的复杂度与时间成本，削弱模型生产流程的整体可维护性与稳定性。

（三）训练层挑战

随着大语言模型的参数量剧增，单个模型的预训练所需卡规模呈爆炸性增长，大模型基础设施复杂度极高，涉及动辄千卡、万卡 GPU 集群、海量的光互联模块，以及极其庞大的计算图与并行策略。在超大规模训练场景下，系统故障已成为常态。

一是长时间高强度运行中，各种硬件问题频发。最显著的困难来自平均无故障时间（Mean Time Between Failures，MTBF）的急剧缩短，实际的大规模集群不仅会经历显存错误校正码（Error-Correcting Code，ECC）、设备离线、高带宽内存（High Bandwidth Memory，HBM）退化或高速外围组件互联（Peripheral Component Interconnect Express，PCIe）链路不稳定等常规问题，更面临光模块（Optical Transceivers）这一高频耗材在高温下的链路抖动与失效挑战。这些硬件故障在长周期训练中会被概率学放大，任何单点故障都可能导致整个任务中断。此外，静默数据损坏（Silent Data Corruption，SDC）是更深层的隐患，硬件计算单元的微小错误可能不报错但导致数值偏差。同时，集群运行在高负载条件下，散热、电源与机架环境稍有不足就会造成 GPU 频繁热降频（Thermal Throttling），引发“木桶效应”，使得训练速度受限于最慢的设备（Straggler），难以保持稳定。硬件压力还体现在存储与输入输出（Input/Output，IO）系统上，当模型规模

达到数百亿或千亿参数时，检查点（Checkpoint）的高频写入对文件系统吞吐量形成显著冲击，IO 瓶颈不仅拖慢训练，更导致有效训练时间大幅降低。

二是大模型训练中，通信瓶颈和同步延迟极易导致 GPU 闲置，严重降低整体训练效率。大模型训练高度依赖 3D 并行（数据、张量、流水线并行），任何通信链路的短板都会立刻重创整体资源利用率（模型浮点运算利用率（Model FLOPs Utilization），MFU）。常见的情况包括梯度同步过程变慢，分布式梯度同步原语（AllReduce）频繁超时，以及某些节点在通信中发生“长尾延迟”，导致所有 GPU 都在等待某几张慢卡。同时，底层网络设备如无限带宽交换机可能出现优先级流控（Priority Flow Control Storm，PFC 风暴）、队列积压或丢包现象，使得对丢包极度敏感的远程直接内存访问（Remote Direct Memory Access，RDMA）网络性能急剧下降。通信库本身也具有不确定性，例如，英伟达集合通信库（NVIDIA Collective Communications Library，NCCL）的版本兼容性问题，当不同节点上的驱动、固件或拓扑配置不一致时，容易出现死锁、进程挂起或环形拓扑结构建立失败的问题。随着模型规模变大，通信量呈非线性增长，许多集群甚至出现资源闲置等通信的状况，严重制约了扩展效率。

三是主流训练框架在超大规模下鲁棒性不足。很多团队在实际训练时会遇到分布式策略在特定输入形状（Input Shape）、激活配置或优化器设置下出现死锁、内存泄漏、流水线气泡不一致的问题。由于训练框架的调用路径非常复杂，许多并行策略在大规模集群上难以完

全覆盖所有边界情况，导致训练过程中出现随机性较强的异常。此外，Checkpoint 机制的容错成本极高，模型参数被切分到不同节点，写入过程容易因为网络抖动或存储压力导致失败。一旦 Checkpoint 写入失败或损坏，训练可能需要回退数十小时，造成昂贵的资源浪费。更进一步的挑战是，训练系统通常缺乏端到端的全链路可观测性，当出现损失值（Loss）异常不收敛、性能掉速或某个进程节点（Rank）出现由硬件静默数据损坏（SDC）引发的数值错误时，很难快速定位“第一案发现场”，也不容易判断是通信拥塞、IO 争抢还是框架内部算子的非确定性导致的问题。

（四）推理层挑战

大模型推理部署阶段的挑战源于其与传统服务截然不同的资源特性和运行模式：**模型体量庞大、计算访存密集、请求负载高度动态，且对延迟和稳定性要求严苛。**

一是大模型推理部署面临计算容量无法适配外部流量高并发波动的核心难题。与训练环境的相对稳定不同，大模型的推理服务往往需要应对瞬时激增的用户请求，例如，热点事件、突发业务场景、节奏性访问或外部流量引导带来的不可预测的负载波动。由于大模型推理的计算成本远高于传统服务系统，且模型权重文件巨大（动辄数百 GB），系统很难像微服务那样实现秒级弹性扩容。一旦流量突破预估区间，若扩容速度跟不上，便会出现排队延迟暴涨、首令牌延迟（Time to First Token，TTFT）恶化甚至集群雪崩等现象。此外，大模型节点冷启动耗时久，进一步降低扩容效率，加剧容量不足的问题；同时流

量突发导致的请求积压，还会向缓存、调度、路由等下游链路传导压力，形成系统性连锁效应。

二是大模型推理因输入输出长度高度可变，易引发显存碎片化，直接冲击调度系统的效率。大模型的显存占用与输入序列长度、输出生成长度呈强相关关系，用户请求所包含的上下文长度从几十Token，到数万 Token 都有可能，这导致系统难以提前准确预判显存消耗。即使采用了连续批处理技术，一旦同时出现若干超长上下文请求，其庞大的预填充计算量会抢占计算资源，导致其他正在解码的短请求出现明显的队头阻塞，使生成出现卡顿。在极端场景下，显存的动态分配若缺乏分页管理，会产生大量碎片，导致系统在名义显存未满时就发生内存溢出（OOM），使得整体系统吞吐急剧下降。对于需要多轮回答或大型输出生成任务的场景，输出长度的不确定性迫使系统只能根据“最坏情况”预留资源，进而导致利用率下降、成本飙升。

三是大模型推理深受模型架构特性与硬件“访存墙”带来的性能瓶颈制约。由于大模型参数规模巨大，推理过程本质上是在计算密集型的预填充（Prefill）阶段与显存带宽密集型的 Decode 阶段之间交替，这对 GPU 的显存带宽、计算单元以及权重加载路径提出了极高且矛盾的要求。模型过大导致单卡难以承载，需要采用张量并行或流水线并行，而这一并行策略本身会引入额外的跨节点通信开销，使推理延迟进一步上升。同时，权重加载、多模型复用（如低秩自适应（Low-Rank Adaptation，LoRA））、键值缓存（Key-Value Cache，KVCache）管理等环节都可能影响每次响应的时效性。特别是 KVCache，虽然

能加速增量解码，但在高并发场景中会消耗海量显存，若调优不当，就会出现缓存频繁换入换出（Swapping）、重复计算或碎片化严重等问题，使推理延迟不可控。

（五）AI 应用层挑战

云原生与大模型深度融合的背景下，人工智能应用已从单一模型服务演进为涵盖 Agent、生成式 AI、多模态系统、自主决策引擎等多样形态的复杂体系。与传统应用不同，AI 应用是基于概率模型与动态上下文进行推理与行动，其行为逻辑并非固定规则驱动，而是由模型参数、输入语境、外部工具依赖及交互上下文共同塑造。当前，OpenClaw 作为开源 Agent 应用，正快速渗透企业办公、业务流程自动化、工业场景智能化等各类领域，凭借自主化推理、动态工具调用、多步任务执行的核心能力，为企业效率提升与业务创新带来全新可能。与此同时，OpenClaw 等新一代 Agent 在规模化落地过程中，token 稳定性失控、大模型调用框架抖动、多模块协同失序、场景化适配不足等稳定性难题集中凸显，让企业在 Agent 落地中面临“功能好用但运行不稳”的现实困境。

一是 AI 应用决策与行为的不可预测性。AI 应用的决策过程通常基于大语言模型等概率生成机制，输出结果具有随机性与上下文依赖性，导致相同行为在不同时间、环境或输入条件下难以完全复现。这种非确定性偏差使得 AI 应用在执行关键任务时可能产生逻辑不可控性，以及生成式 AI 应用的幻觉与事实一致性风险。尤其在多 Agent 协作场景中，概率决策间的叠加效应进一步放大了系统的不稳定性。

二是 AI 应用依赖链路的级联脆弱性，多模态系统的对齐与融合不确定性。AI 应用通常通过外部工具调用来完成任务，而这些依赖构成了跨层次、高耦合的动态执行链条。任何一环的性能波动、接口变更或权限异常，都可能导致任务失败或异常反馈，形成级联式故障传播。随着 AI 应用的生态扩展，依赖链越长、耦合度越高，可靠性保障的边界也越模糊。

三是对 AI 应用评估、监控与治理的体系性缺失。传统监控体系难以捕获 AI 应用的逻辑稳定性与决策可靠性，面向在线服务的 AI 应用要求低延迟响应，但面临模型压缩与精度损失、冷启动与缓存失效的风险。AI 应用的运行状态受知识源更新、上下文管理、工具调用质量等多维因素影响。缺乏统一的可观测性框架与行为评估指标，使得 AI 应用在运行过程中存在治理盲区，难以及时发现与纠正潜在的逻辑错误或行为偏差。

四是传统监控难以适配，AI 可观测与治理存在明显短板。不同于传统在线服务的确定性规则，AI 应用基于概率模型推理生成，受模型架构不确定性、数据驱动不完全性、交互环境开放性影响，存在结果不确定、上下文敏感等固有特性。当前传统监控难以捕获其逻辑稳定性与决策可靠性风险，且缺乏统一的可观测性框架与适配全生命周期的评估体系（含多维度指标、全链路追踪），无法应对 Agent 阶段对幻觉、合规性、准确性等核心要求；同时，AI 运行状态还受知识源更新、工具调用质量等多维因素影响，最终导致治理盲区，难以及时发现纠正输出偏移、语义失准等漂移问题。

五是对于 AI 应用安全与伦理的新型风险。AI 应用的高自治性与工具调用能力，使其具备跨系统操作与外部交互能力。一旦身份验证、调用约束或上下文控制不当，可能造成越权执行、数据泄露甚至逻辑攻击。同时，AI 应用的生成决策过程缺乏完全可解释性，可能引发偏见放大、伦理越界等信任危机。这要求可靠性工程不仅关注技术稳定性，也需纳入可信与伦理治理机制。

三、云上人工智能系统的可靠性设计

（一）可靠性工程的设计原则

随着人工智能技术与云基础设施的深度融合，云上人工智能系统的可靠性设计，不再仅限于传统意义上的“服务不中断”与“可恢复”，而是针对云上人工智能系统在概率性推理、不确定性输出、模型持续演化及 Agent 自主决策等特征，扩展到基础设施、数据、模型与应用协同演化的多层稳定性与可信性，体现 AI 系统可靠性设计的专属性与差异性。同时以可用性与恢复性为核心目标的传统可靠性工程体系，也已经无法完整覆盖云上人工智能系统在基础设施、数据、模型与应用场景中的可靠性需求。

为构建具备韧性、自治性与可信性的 AI 系统，本指南提出以下六项可靠性工程设计原则，作为技术架构与实施策略的指导思想。

本文所提出的可靠性工程原则，综合参考了云计算与分布式系统可靠性工程实践（如 SRE）、云架构设计框架（如 AWS Well-Architected Framework）、可观测性理论（如 Distributed Systems Observability），以及可信人工智能相关规范，并结合云上人工智能系统的技术特征进

行扩展与抽象形成。

1.不确定性可验证原则

云上人工智能系统的核心特征在于其基于概率分布的推理机制，模型输出天然存在不确定性与波动性。因此，可验证性不再仅关注结果正确性，而需覆盖输出置信度、分布稳定性与行为一致性。系统应构建面向模型与 Agent 的评估体系，包括置信度校准、漂移检测、对抗测试与基准评测，实现从“结果验证”向“分布验证”与“行为验证”的演进。

2.可观测性原则

云上 AI 系统不仅需要观测系统性能，更需要观测模型与 Agent 的推理过程与决策路径。传统日志与指标无法解释智能行为，必须引入对模型输入输出、上下文状态、推理链（Chain-of-Thought）、工具调用轨迹等信息的采集与分析能力，构建覆盖“基础设施—数据—模型—Agent”的多层语义观测体系。

3.多层自愈与模型弹性原则

多层自愈与模型弹性原则要求系统能够在面对环境波动、资源异常或任务失效时具备自主故障诊断与自愈能力，以维持持续可用的运行状态。云上 AI 系统的失效不仅来源于基础设施，还包括模型退化、推理异常与 Agent 决策失效。自愈能力需从基础设施层扩展至模型与 Agent 层，包括模型降级（如切换小模型）、推理重试策略、Prompt 修复、策略回退及多模型冗余机制，实现从“资源自愈”向“智能行为自愈”的扩展。

4.可演化性原则

可演化性原则强调系统应在持续学习与模型更新的过程中保持稳定与一致性，实现动态演化下的长期可靠性。云上 AI 系统需持续进行模型更新与数据迭代，但模型演化可能引入性能回退或行为漂移。因此必须构建面向模型的版本治理体系，包括离线评测、在线 A/B 测试、灰度发布与自动回滚机制，并结合数据版本与特征分布监控，实现模型演化过程中的可控性与可追溯性。

5.跨层耦合协同原则

云上 AI 系统的性能与可靠性由计算资源调度、数据质量、模型结构与 Agent 策略共同决定，各层之间存在强耦合关系。例如计算资源波动会影响推理延迟，进而改变 Agent 决策路径。为此，应通过跨层资源调度、状态同步与策略协同，打通基础设施层的资源弹性、模型层的推理稳定与 Agent 层的决策自治，构建系统级一致性保障机制。该原则要求系统具备跨层感知、全局优化与联动响应能力，从而避免因层间耦合不当或状态不一致导致的链式失效，确保全局稳定性与业务连续性。

6.可信与安全约束原则

可信与安全约束原则强调，人工智能系统的可靠性不仅关乎“技术稳定”，更关乎“行为可信”与“社会可接受性”。云上 AI 系统具备自主决策与工具执行能力，其风险不仅在于系统故障，更在于错误决策带来的放大效应。因此需在系统设计中引入安全约束机制，包括输出安全过滤、权限隔离、工具调用审计、人类在环（Human-in-the-loop）

控制及公平性评估，确保模型行为在可控边界内运行。

（二）可靠性工程分层技术架构

云上人工智能系统的可靠性保障应构建在“基础设施层—数据层—模型层—推理服务层—AI 应用层”的分层技术架构之上。该架构的核心思想是利用分布式系统的冗余性与自治能力，将单点故障风险分散至全局，通过各层的分布式协同机制，构建端到端的容错与自愈体系。各层应具备相对独立的可靠性控制能力，并通过跨层协同机制形成端到端的防御体系和演进能力，统一支撑高可用性、高可解释性、高鲁棒性与高可恢复性目标，如图 2 所示是可靠性工程分层技术架构图，以下是各层次的具体建设方法。

图 2 可靠性工程分层技术架构图

1.基础设施层设计

基础设施层应构建覆盖计算、存储、网络三大维度的纵深韧性架构，利用分布式架构消除单点故障，支撑大规模 AI 系统在高并发场景下的稳定运行。

计算层（硬件、固件与调度协同）设计。构建统一异构计算资源

池，屏蔽底层硬件差异，按任务重要性、服务等级协议（SLA）分级调度。核心任务绑定高保障资源池，避免混部争抢；弹性任务搭配竞价实例与检查点（Checkpoint）机制，保障任务进度可恢复。部署硬件遥测代理，实时采集温度、错误校正码（ECC）错误率、PCIe 重传率等指标，异常时自动标记故障节点，通过断点续训将任务迁移至健康节点，实现故障透明自愈。同时支持虚拟机监控程序（Hypervisor）热升级、虚拟机实时迁移，容器平台集成拓扑感知调度，基于非统一内存访问（NUMA）、高速互联拓扑优化部署，提升算力利用率。

存储层（硬件介质与数据服务）设计。训练集群本地缓存层采用非易失性内存主机控制器接口规范固态硬盘（Non-Volatile Memory Express Solid State Drive，NVMe SSD）或计算快速链路内存池（Compute Express Link，CXL），提供微秒级输入输出（I/O）响应；共享存储后端部署全闪存分布式文件系统（如 Lustre、WekaFS）或对象存储（如 MinIO/ 简单存储服务（S3）），支持远程直接内存访问（RDMA）直通访问，消除协议栈瓶颈。存储系统应默认采用分布式多副本或纠删码冗余，关键数据跨可用区同步复制，基于仲裁机制保障写入强一致，全链路启用校验和防范静默数据损坏。同时按数据访问热度自动分层存储，结合业务标签配置生命周期策略，平衡性能与成本。

网络层（硬件设备与通信协议）设计。数据中心内部采用 Clos 或 Fat-Tree 无阻塞架构，核心/汇聚交换机双活部署，链路聚合控制协议（Link Aggregation Control Protocol，LACP）与双向转发检测

（Bidirectional Forwarding Detection，BFD）快速检测确保故障收敛<50ms，消除单点故障。关键训练集群强制部署融合以太网远程直接内存访问（RoCEv2）或无限带宽，启用优先级流控（Priority Flow Control，PFC）与显式拥塞通知（Explicit Congestion Notification，ECN）协同流控，实现微秒级延迟与接近零丢包。网络服务质量（QoS）策略优先保障分布式梯度同步原语（AllReduce）、梯度同步等关键流量。通过单根输入输出虚拟化（Single Root I/O Virtualization，SR-IOV）等技术实现网络资源的高性能、低中断切换，避免因底层维护和扩容对上层 AI 工作负载造成可感知影响。

2.数据层设计

在数据层，应围绕“数据可用、数据可信、数据可追溯”目标构建高可用、一致性及流转的高效性分布式数据架构，为模型训练与推理提供稳定可靠的数据基础。

数据采集安全可信设计。通过统一合规网关进行数据的敏感信息识别（如身份证、银行卡号等）、脱敏或拦截，满足通用数据保护条例（General Data Protection Regulation，GDPR）、网络安全法等监管要求，实现合规准入控制。数据传输时，强制使用传输层安全协议（Transport Layer Security，TLS）/ 双向传输层安全协议（mutual Transport Layer Security，mTLS）与双向身份认证，防止中间人攻击与未授权访问。记录采集时间、地理位置、设备标识（ID）、操作主体等关键元数据，可通过链式存证等方式实现数据溯源与不可抵赖。

数据处理平台高可用设计。基于容器编排平台（Kubernetes）与

微服务设计，通过消息队列（如 Kafka）实现任务缓冲。支持双活/多活部署，故障时保障任务连续性。通过消息队列（如 Kafka、Pulsar）实现采集、标注、清洗任务的异步流转，提升系统弹性与吞吐能力。对关键场景实施多标注者交叉验证、抽样复核、一致性评分（如 FleissKappa），确保标注准确率达标；未达标任务自动回流，形成质量闭环。

自动化数据质量流水线设计。数据预处理流水线应支持端到端自动化清洗流程，覆盖缺失值处理、异常值识别、格式规范与特征衍生等环节。通过 SLA 量化管控，定义明确的服务等级协议，如处理延迟、吞吐波动率、错误率，并在指标越界时自动触发告警与降级策略。

动态数据健康监测设计。建立动态数据漂移监测能力，针对关键特征定期或实时计算柯尔莫哥洛夫–斯米尔诺夫检验（Kolmogorov–Smirnov 检验，KS 检验）统计量、相对熵（Kullback–Leibler 散度，KL 散度）等指标，用于衡量生产数据分布与训练基线分布的差异。设置分级响应策略，当漂移指标超过预设预警阈值时，应自动触发重采样、重标注或模型重训练流程；当检测到整体数据缺失率超过设定比例或关键特征偏移达到严重阈值（如多倍标准差）时，系统应进入保护态，通过限流、降级或人工复核避免异常数据直接影响业务。

数据资产治理与追溯设计。对数据资产进行细粒度版本管理，对训练集、验证集、线上样本集进行版本快照，记录数据来源、处理规则、时间窗口、质量评分等元信息。并维护基准数据集（Golden Dataset），用于不同模型版本之间的性能对比及回归验证，为可靠性评估和审计

提供依据。

面向幻觉抑制的数据工程设计。数据层应针对大语言模型容易产生的“事实冲突”与“无中生有”等幻觉特征，建立全生命周期的幻觉防御机制。在预训练数据采集阶段，除通用的合规性过滤外，应增加事实性校验逻辑实现数据源头事实性核验与过滤，同时维护一套针对幻觉评估的数据集，覆盖事实矛盾、逻辑错误等多种幻觉类型，在模型发布流水线中嵌入自动化的幻觉率检测，实现对幻觉风险的量化监控与风险熔断。

3.训练层设计

在训练层，需利用分布式并行策略解决大模型训练挑战，并保障训练过程的鲁棒性，应围绕“训练稳定性、推理一致性、决策可解释性与过程可复现性”构建系统化的工程机制，确保模型在复杂环境与持续演化中的可靠行为。

分布式训练架构的韧性设计。构建高效且稳定的分布式训练框架是实现大规模模型训练的基础。面对超大规模模型（如参数量超过 100 亿），需要采用先进的分布式训练模式，例如参数服务器（Parameter Server）或环形全量归约（Ring-AllReduce）。通过梯度压缩技术减少通信开销，并利用异步 Checkpoint 机制最小化故障恢复时的进度损失。此外，平台须具备实时监控各训练节点上的图形处理器（GPU）使用情况、远程直接内存访问（RDMA）网卡状态及存储性能的能力。一旦检测到节点故障等异常情况，能够自动重新分配资源，保证训练任务的连续性和收敛性。

由于大模型训练的资源投入大、训练周期较长，通常需要投入数十卡到数百卡甚至数千卡，且训练周期需要一周以上甚至数周，所以大模型训练的“断点续训”工程能力是保障训练架构韧性的核心能力。不论是基于原生工具 PyTorch 等的自有“状态保存”与“状态恢复能力”，还是基于商业化大模型训练平台的断点续训能力，在大模型训练阶段发生批量服务器故障等原因导致的训练中断场景下，快速完成故障恢复及续训，是训练连续性保障的重要措施。

这种设计不仅提高了训练效率，也增强了系统的容错能力，确保即使在硬件或网络不稳定的情况下，也能维持训练的稳定进行。具体分布式系统稳定性建设可参考中国信通院《分布式系统稳定性成熟度模型》，该标准覆盖稳定指标、故障预防、故障感知与分析、预案能力、故障改进、安全管理及流程机制 7 大能力域。

模型版本管理与过程可复现性设计。为确保模型迭代过程中的透明性和可控性，对每次发布的模型都应形成完整的训练谱系。这意味着不仅要对训练代码和超参数配置进行版本控制，还要涵盖所使用的数据集版本以及运行环境的具体信息。固定随机种子并限制并行度波动范围，使得重复训练可以在可接受的误差范围内产生一致的结果。当遇到重大性能偏差时，基于留存的训练谱系可以快速回溯至问题发生的场景，便于深入分析原因并采取相应的改进措施。这样的做法不仅能提高研发效率，还能增强模型更新过程中的一致性和稳定性，确保每一次迭代都是基于可靠的数据和技术基础之上完成的。

可解释性与推理一致性保障设计。对于那些支撑关键业务决策的

模型来说，其输出结果应当是可以被理解和验证的。为此，应在模型内部集成解释机制，例如采用沙普利加性解释值（Shapley, SHAP）等方法输出关键特征的贡献度与敏感性分析，支持对单次预测与整体行为进行多维解释。平台宜支持影子训练或影子推理机制，在新旧模型并行运行时，对输出差异、置信度变化和业务影响进行系统性对比，当差异超过预设阈值时应自动阻断升级或回滚模型版本。

强化模型安全与鲁棒性设计。考虑到潜在的安全威胁和输入扰动可能对模型造成的负面影响，必须在模型层面上强化安全防护措施。具体包括但不限于对抗样本检测、防投毒训练以及嵌入模型水印等手段。特别是对于涉及高风险业务的模型更新，强制执行安全与公平性的评估流程，通过模拟多样的攻击场景和测试用例来检验模型在不同条件下的表现。只有通过了严格的鲁棒性测试，确认其能够在面对异常输入或恶意干扰时依然保持稳定性能的模型，才允许正式部署上线。此举极大地增强了模型对外部威胁的抵抗能力，同时也保障了其在实际应用中的可靠性和安全性。

内生真实性对齐设计。模型训练阶段应通过优化损失函数和反馈机制，从算法层面构建对事实准确性的内生约束。关注事实性增强训练，通过调整损失函数权重促使模型在生成关键事实 Token 时更加聚焦，减少因注意力稀释导致的推理幻觉。在强化学习阶段，推行过程监督理念，不仅奖励最终正确答案，更对推理路径中的每一个正确事实步骤进行正向激励，通过优化奖励函数设计，防范模型为追求高奖励而产生“奖励黑客”或“阿谀奉承”现象。

4.推理层设计

推理服务层是面向最终业务的直接承载，其稳定性、响应效率与成本可控性直接决定了智能应用的用户体验与系统韧性。为构建面向高并发、低时延、强连续性要求的智能服务底座，推理层需以“端到端高可用、精细化资源治理、安全可控演进”为核心原则，通过分布式服务化架构，应对高并发访问请求，保障服务的高吞吐与低延迟。

单元化与多活架构设计。将服务按地域、可用区或业务域划分为自治的服务单元，每个单元包含独立部署的模型实例、本地缓存及必要的依赖资源（如向量数据库、特征存储）。这种隔离式架构不仅降低了跨区域调用带来的网络延迟，更在局部故障（如单可用区断电、网络分区）发生时，能够通过全局流量调度系统自动将请求切流至健康单元，实现分钟级故障隔离与业务无感恢复，从根本上提升系统的容灾能力和地域亲和性。具体应用多活架构设计可参考中国信通院《应用多活成熟度模型》，该标准对架构可靠度、应用成效度、组织建设度等方面进行了成熟度的规范。

服务质量监控与弹性伸缩机制。平台应实时采集并分析涵盖性能、资源与业务语义及多租户、输入输出长度分层的多维指标，包括各租户、各输入输出长度区间的每秒查询率（Query Per Second，QPS）、首令牌延迟（TTFT）、每输出令牌耗时（Time per Output Token，TPOT）、图形处理器（GPU）流式多处理器（SM）利用率、显存带宽使用率等关键服务等级指标（SLI）。基于这些指标，结合历史负载模式、业务优先级及租户等级、请求长度特征，动态触发水平扩缩容策略，在保

障 SLA 的同时避免资源浪费。针对突发流量或异常请求可能引发的雪崩效应，系统需内置多层次防护能力：通过令牌桶（TokenBucket）或漏桶（LeakyBucket）算法实施精细化限流，结合租户等级与请求长度差异化配置阈值；对异常下游依赖启用熔断，对非关键请求引入排队缓冲，并在资源紧张时启动优雅降级（如按租户优先级、请求长度差异化切换至轻量模型、返回缓存结果），确保核心业务链路始终可用。

渐进式发布与灰度控制体系。新模型或服务版本支持金丝雀发布机制，通过小比例真实流量验证新模型或新版本服务，当错误率、业务指标或合规性指标超出基线一定范围时，应自动回滚至稳定版本。同时，应支持暗流量测试（ShadowTraffic）能力，在不影响用户的前提下对新模型进行并行评估，对输出差异、资源消耗与成本影响进行量化对比，为大规模发布提供依据。

解码干预与自校准设计。推理服务层应集成实时拦截与干预机制，确保模型在生成过程中遵循事实一致性。比如集成了事实性核采样（Factual-nucleusSampling）算法，根据生成内容的置信度动态调整采样概率，平衡生成的多样性与事实准确性，或引入“层对比解码”（DoLa）策略，通过对比模型不同层级间的未经过归一化的概率（logits）差异来放大事实性知识的权重。另外，还可以内置自校准与验证链（CoVe）设计，支持模型在生成初步回复后，自动拆解其中的事实命题并生成验证问题进行自我纠偏，实现推理结果的后处理增强。

5.AI 应用层设计

AI 应用层是云上人工智能系统的高层抽象，应在保障业务安全与合规的前提下，通过微服务化与全局观测性，保障业务逻辑的连续性与可观测性，实现自主决策与持续演化。

行业边界的严格定义。建立明确的行为边界与操作约束，对高风险操作（如删除关键资源、跨系统修改配置等）实行硬性禁止或强制多人确认机制。对具备工具调用能力的 Agent，应通过白名单、黑名单和粒度化权限控制限制其可调用的接口范围，对潜在高危命令（如操作系统级删除命令“rm-rf”等）进行静态和动态双重拦截。

内容安全需贯穿全链路。建立覆盖输入、推理、输出和执行环节的伦理风险治理体系，识别并处置违法、暴力、歧视、隐私泄露、欺诈、偏见及虚假信息等风险。输入侧拦截恶意提示或隐私探查请求，在多轮对话中防范渐进式越界；输出侧对生成内容进行安全与公平性评估，超标即拦截、降级或转人工。定期开展红队测试，持续优化检测规则，并对关键操作全程审计，确保可追溯、可问责。

决策过程应具备不确定性感知。Agent 在处理高风险决策时，应具备不确定性量化与阈值控制能力，例如通过蒙特卡洛随机失活（Monte Carlo Dropout，MCD）对决策置信度进行估计。当连续多次输出置信度低于阈值或结果分歧过大时，自动切换至规则引擎、专家策略或人工介入模式，避免低可信度决策直接影响业务，守住安全底线。

演进机制依托沙箱与人机协同。Agent 更新与策略演化应优先在沙箱环境中完成，沙箱环境应尽可能还原生产依赖与流量特征，对行

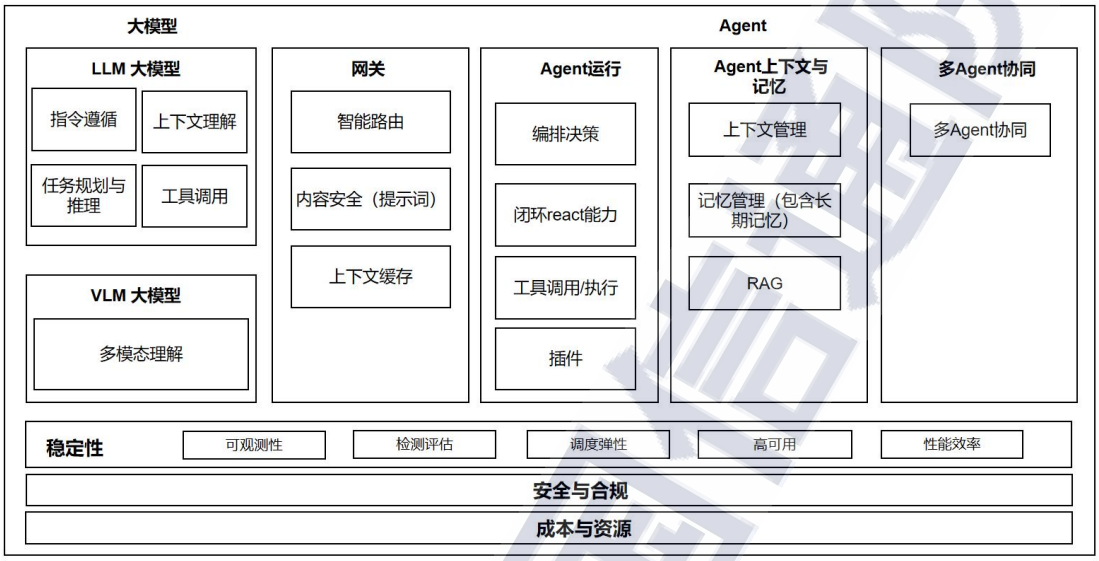
为偏差、资源消耗、调用链路安全性进行全面评估，当偏差超过预设阈值时应冻结更新并回滚至稳定版本。对于关键决策，应建立人类在环机制，通过工单系统或审核系统对 Agent 的高风险操作进行抽查或全量审核，并将审核结果作为学习反馈回流至模型与策略训练过程，形成“Agent—人类—系统”三者之间的闭环治理体系。

外部工具调用的安全可控设计。针对外部公开工具/API 调用场景，建立分级授权+请求校验+结果核验全流程管控机制：通过工具白名单明确可调用范围，禁止调用未备案服务；对请求参数、调用频率进行严格校验；对返回结果开展格式与真实性核验，异常时自动重试或切换备用接口，确保调用安全可靠。

基于外部工具和系统级反馈结合的幻觉抑制。引入检索增强生成（Retrieval-Augmented Generation，RAG），将问答模式从“闭卷考试”转为“开卷考试”，通过引入实时、权威的外部知识库来补充模型的参数化记忆，从源头上降低误导性信息的产出比例。同时构建集规划、执行、反思于一体的 Agent 框架，利用多 Agent 协作（如“生成者—审核者”模式）来隔离并纠正单点生成的幻觉。同时要设计不确定性感知与风险熔断的能力，实时监测模型生成的语义不确定性（如激活熵值），当置信度低于预设阈值时，自动触发降级逻辑，由开放式生成切换为受限检索模式或提示人工介入，将生成行为从“概率预测”转向“事实锚定”。

具体 Agent 可靠性建设可以参考中国信通院《Agent 可靠性能力要求》，该标准聚焦如何让 Agent 更稳定地运行，从大模型（包括 LLM

大模型、VLM 大模型）、网关（衔接层）、Agent 的运行、上下文与记忆、多 Agent 协同等模块，对整个 Agent 运行流程中所需的可靠性要求进行规范。



来源：中国信通院《Agent 可靠性能力要求》

图 3 Agent 可靠性能力要求框架图

（三）可靠性工程管理体系

在云上人工智能系统中，可靠性不仅关乎技术稳定性，更涉及数据、模型、服务与应用全生命周期的协同治理。要系统性提升 AI 系统的可用性、一致性、可解释性与抗风险能力，必须构建一套覆盖组织、流程、技术与度量的可靠性管理体系。本体系通过合规与安全、可观测性、灾难应急管理三大板块，构建从预防到快速恢复的闭环管理机制。

1.灾难应急管理体系

灾难应急管理体系是 AI 系统可靠性的核心保障，通过“事前预防、事中应急、事后优化”的全流程管理，确保人工智能系统具备快速响

应与恢复能力。

恢复时间目标（Recovery Time Objective，RTO）/ 恢复点目标（Recovery Point Objective，RPO）是灾难应急管理的前提。企业需根据业务影响分析对 AI 系统进行分级管理，设定差异化的容灾恢复标准，并定期评估更新 RTO/RPO 指标。这不仅是后续架构设计的基础，也是故障恢复的底线标准。同时，容灾系统的建设需要投入大量的成本，需要基于大模型支撑的业务场景的重要程度，进行灾难管理的业务影响分析（Business Impact Analysis，BIA），确定分层级的 RTO/RPO 目标。例如，如果大模型支撑的 AI 应用，嵌入了在线电商业务交易、商业广告智能推荐等联机事务处理（OLTP）类关键业务系统，则需要考虑建设小时级的 RTO、分钟级的 RPO；如果大模型支撑的 AI 应用，仅支持月度销量分析等联机分析处理（OLAP）类非关键业务系统，则可暂时延缓容灾系统建设，采用天级 RTO、小时级 RPO 标准。

多层容错容灾体系设计确保系统具备高可用性。在遵循“深度防御”原则的基础上，构建覆盖基础设施、数据、服务与模型的全方位、多层次弹性架构，确保任一单点或局部故障均不会导致系统级失效。

1）基础设施层的容错容灾设计：网络层面，实施多链路接入与动态路由协议，确保在主链路中断时可自动、无缝切换至备用链路，保障网络连通性。计算层面，通过服务器集群与资源池化技术，实现硬件故障场景下业务的自动迁移与无感知恢复，提升计算资源的整体可用性。存储层面，采用分布式存储架构，并结合数据冗余技术（如

多副本、纠删码），避免因单点或多点存储故障导致数据丢失或服务中断。

2）数据层的容错容灾设计：依据业务确定的恢复点目标（RPO），部署跨可用区（AZ）或跨地域（Region）的数据同步或异步复制机制，确保关键数据具备异地备份与快速恢复能力。制定并执行完善的数据备份和恢复策略，包括全量、增量及日志备份，并定期开展备份数据恢复演练，验证备份的有效性与可恢复性。

3）服务与模型层的容错容灾设计：服务高可用，采用多活或主备部署模式，配合负载均衡设备，实现服务节点的自动故障检测与流量切换，实现服务高可用。在多个区域部署模型推理服务副本，并建立可靠的版本同步机制，实现模型容灾。当主区域服务不可用或性能严重下降时，可自动将请求路由至备用副本。

4）弹性与降级机制：弹性伸缩方面，根据实时负载指标（如 QPS、GPU 利用率等）动态调整资源规模，以应对业务流量波动。服务降级方面，预设降级策略，当核心依赖服务（如特征库、复杂模型）发生故障时，可自动启用简化模型、规则引擎或返回缓存/默认结果，保障核心业务的基本可用性。

5）跨层级协调机制：建立跨层级的容灾协调机制，确保在区域性故障等场景下，数据、服务与基础设施各层的恢复动作能够有序联动，形成整体恢复能力。

主动演练与验证机制有助于暴露隐患并验证预案的有效性。主动演练与验证是检验系统韧性、完善应急响应机制的核心手段。应建立

常态化的演练体系：**一是开展混沌工程**，在生产环境受控范围内注入通用故障（如实例宕机、网络延迟）及 AI 特有故障（如数据污染、分布漂移），验证系统容错能力与监控告警有效性；**二是组织红蓝对抗演练**，全过程实行“演练前评估—演练中记录—演练后复盘”的闭环管理；**三是实施综合性容灾演练**，按年度计划定期验证数据恢复、服务切换、模型回滚等关键能力，并将结果纳入可靠性评分卡驱动改进。同步构建高效、自动化的应急响应机制，制定覆盖各类故障场景的应急预案与操作级 Runbook，确保工程师可快速执行；建设智能应急平台，基于多维监控指标自动分级告警（P0–P2），对通用故障执行重启、扩缩容等自动化恢复，对 AI 特有异常（如模型漂移、置信度骤降）触发版本回滚或降级策略，并结合历史故障库实现根因推荐与辅助决策。每次事件后须在 24 小时内完成复盘，输出含改进项、责任人和时限的《故障复盘报告》，并将经验沉淀至知识库，形成“演练—响应—复盘—优化”的持续提升闭环，真正实现“以练促防、以战促改”。

合规与安全、可观测性、灾难应急管理三者相辅相成，共同构成了云上人工智能系统可靠性的铁三角。合规与安全是前提，划定了系统运行的边界；可观测性是手段，提供了洞察系统内部的能力；灾难应急管理是机制，确保了故障发生时的快速响应与恢复能力。只有将这三者有机结合，并融入从设计到运维的全生命周期，才能构建出真正可靠、可信的云上人工智能系统。

2.合规与安全体系

合规与安全体系为 AI 系统的可靠运行提供制度保障和技术防护，是确保系统符合监管要求、抵御安全风险的基础支撑。需从法律遵循、算法透明、安全防御及供应链管控四个维度构建全方位的治理能力。

合规性框架建设是确保 AI 系统合法运行的基础。首先，在严格遵循《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等上位法的基础上，须将“模型可解释性”作为合规准入的核心要素，建立分级分类的算法治理机制。AI 系统应具备决策逻辑的追溯能力，拒绝“黑盒”运作。对于可能影响用户权益的自动化决策，系统必须提供清晰、易懂的决策依据或影响因素说明，保障用户的知情权和申诉权。其次，构建可信评测机制，将可靠性、安全性、公平性、隐私保护等维度纳入上线准入与运行评审，依据《生成式人工智能服务管理暂行办法》，生成式内容服务应具备来源追溯与内容标识能力。对于生成的文本、图像或视频，系统应能解释其生成的参考来源或数据依据，并建立防伪造机制，避免产生误导性不可解释内容。确保所有记录可审计、可追溯。最后，实施严格的版本发布评审流程，在业务高峰期或重要保障期执行变更冻结策略，控制人为操作风险。

纵深安全防御体系为 AI 系统提供全方位的安全防护。内生安全管理重点在于保障模型与数据安全。采用对抗训练提升模型鲁棒性，使用差分隐私、同态加密技术保护训练数据和推理隐私，防范数据投毒、模型窃取、对抗攻击等风险。应用安全管理方面，加强生成式 AI 的内容过滤与毒性检测，建立提示词注入攻击防护机制，并对敏感操

作实行二次确认。此外还需强化技术措施落地。首先是构建数据安全管控机制。实施数据分类分级管理，强化访问控制、加密传输与存储、敏感数据脱敏与最小化采集等保护措施。其次是加强平台的安全管控。遵循零信任架构、最小权限原则，严格管理密钥与凭证。最后是强化供应链的安全管控。对第三方模型、开源组件进行漏洞扫描和安全评估，建立漏洞监控和应急响应机制。

3.可观测性体系

可观测性体系是 AI 系统可靠性的感知中枢，通过实现从基础设施到业务输出的全链路观测，为系统状态认知与问题定位提供数据支撑。

全链路、多维度数据采集机制为系统状态认知提供基础。构建覆盖基础设施层、模型层和业务层三个维度的数据采集体系，打通从底层硬件资源到上层业务应用的数据流，实现端到端的数据贯通。**基础设施层**需采集服务器 CPU、内存、GPU 利用率、显存、RDMA 网络带宽占用、存储 I/O 等关键指标，实时监控计算资源的负载情况与健康状态。**模型服务层**需重点收集模型调用量、响应时延、并发请求数、错误率等运行指标，以及模型输入输出的完整日志，包括提示词、生成结果、Token 数量等细节信息。**AI 应用层**需跟踪用户交互行为数据、业务流程执行状态、功能调用成功率及用户反馈数据，如满意度评分、纠错记录等。

针对生成式 AI 的特性，需特别关注模型训练与推理过程中的数据漂移指标，如训练数据与推理数据的分布差异、特征重要性变化等，

为后续模型优化与问题定位提供数据支撑。同时，数据采集需覆盖时间、空间、业务逻辑等多个维度，时间维度上实现毫秒级采样与持续追踪，空间维度上关联多节点、多服务的协同数据，业务维度上区分不同场景、不同用户群体的特征数据，确保构建起立体、全面的观测数据基座。

统一的可观测与智能分析平台提供端到端的可观测能力。建设统一可观测平台，集成多源数据采集、流式处理、智能分析等功能，提供端到端的可观测能力。

首先，采用分布式架构实现高性能数据处理，通过自定义仪表盘实现多维度数据可视化。通过自定义仪表盘将基础设施健康状态、模型性能指标、业务运行数据及 AI 风险指标进行多维度、立体化的展示，支持下钻分析和时间范围对比，帮助运维、算法与业务人员快速掌握系统全局状态。

其次，引入机器学习算法实现智能异常检测与根因定位，自动识别性能异常、质量下降等问题。通过历史数据训练的基线模型（如孤立森林、LSTM 时间序列预测）结合动态阈值策略，能够自动识别 CPU 突增、推理延迟异常、预测准确率骤降、幻觉率超标等潜在问题，并基于预设的规则引擎或强化学习模型进行根因定位推荐。

最后，平台应提供交互式分析能力，支持用户进行深度数据探查与关联分析。平台应提供类似 SQL 的查询语言或可视化拖拽式分析工具，允许用户自定义指标组合、构建复杂的关联规则，甚至进行 A/B 测试效果的量化评估。具体可观测性能力建设，可参考中国信通院《云

计算智能化可观测性能力建设成熟度模型》，该标准规范和指导云计算环境下的智能可观测性建设实践，为企业实施云环境下的智能化可观测能力建设提供指导。

建立模型可解释性与溯源机制。强化模型可解释能力，在模型开发阶段优先选择可解释性强的算法，对深度学习模型应用 LIME、SHAP 等解释工具。通过特征重要性排序、局部决策边界可视化等方式，直观呈现模型决策依据。模型溯源机制则要求对模型生命周期中的所有关键节点进行全程记录与追踪，构建完整的模型血缘关系图谱。具体而言，需详细记录模型训练数据的来源、采集时间、预处理逻辑及质量校验指标，确保数据的可追溯性；同时，完整保存模型训练过程中的超参数配置、训练日志、版本迭代记录以及每次部署上线的审批流程。当模型在生产环境中出现性能漂移或决策偏差时，可通过溯源机制快速定位问题环节，判断是训练数据分布发生变化，还是某次版本更新引入了缺陷。

（四）可靠性工程分层评估体系

基于可靠性工程的整体技术架构与管理体系要求，本章节构建了覆盖基础设施、数据、训练、推理及 Agent 的分层评估框架。该框架旨在为人工智能系统可靠性建设提供能力评估和监测的能力基线。同时通过有效的度量与管理，提升云上人工智能系统在全生命周期中的稳健性与可信度，保障其稳定、可靠地对外提供服务。

1.基础设施层

随着 AI 模型规模计算集群的持续增长，AI 基础设施已从“支撑

性资源”演变为影响系统连续性的核心载体。其可靠性已从单点可用性，转向存储、网络、硬件及平台组件的全栈协同韧性。尤其在千卡乃至万卡级智算集群中，组件数量与拓扑复杂度显著提升，局部性能波动或故障都可能被放大，导致训练中断或服务降级。

本节从**存储、网络、硬件、平台组件**四个层面出发，聚焦关键指标，并阐明其在 AI 场景下的必要性与独特价值。

（1）存储层

AI 工作负载对存储系统提出了前所未有的挑战：训练阶段需高吞吐并发读取大规模数据集，推理阶段要求低延迟加载模型权重，而训练检查点（Checkpoint）的保存与恢复要求高带宽和高可靠性。传统通用存储指标难以刻画上述场景下的真实瓶颈，需构建适配 AI 特征的评估框架。本节围绕吞吐性能、元数据性能、大文件传输可靠性及缓存效率等维度构建针对性度量体系。

表 1 存储层度量指标

指标	指标解释/计算方法（扩展）	选择该指标的原因
聚合读写吞吐（GB/s）	通过行业标准 I/O 测试工具或生产环境 I/O 监控采集多客户端并发读写带宽；通常取典型分位值（如 P95）	大模型训练需高频加载 TB 级数据集，吞吐不足会成为训练瓶颈。该指标反映存储系统在高并发下的真实服务能力
元数据操作延迟（ms）	监控 AI 典型元数据操作（如文件打开、属性查询）等系统调用的 P99 延迟，或使用元	AI 训练常涉及海量小文件（如 ImageNet），元数据性能差会导致“数

	数据性能测试工具测量每秒元数据操作数（MOPS）	据饥饿”，降低数据供给连续性
大文件分片上传/下载成功率	成功率=成功完成的分片请求数/总分片请求数；需覆盖断点续传、重试逻辑	大文件（如检查点）任何分片失败都可能导致训练中断或状态不一致，必须保障端到端可靠性
缓存命中率	命中率=缓存命中的数据访问次数/总数据访问次数；可通过多级缓存系统内置指标获取	高命中率可显著减少后端存储压力和 I/O 延迟，对推理服务响应时间（P99latency）有直接影响
存储服务可用率	可用率=（总时间-不可用时间）/总时间；“不可用”定义为 API 错误率>5%或时延>阈值持续 5 分钟	作为 SLA 核心指标，直接决定用户能否正常提交训练任务或加载模型，是稳定性的底线

来源：中国信息通信研究院

（2）网络层

AI 训练高度依赖节点间高效通信，尤其是基于 AllReduce 的梯度同步，对网络带宽、延迟和丢包极为敏感。参数面普遍采用 RDMA/RoCE/NVLink 等高速互联技术，而数据面仍依赖 TCP。在超大规模集群中，网络拥塞、PFC 风暴或慢收敛等问题极易引发级联故障，需重点监控通信效能、无损传输保障与故障恢复速度。

表 2 网络层度量指标

指标	指标解释/计算方法（扩展）	选择该指标的原因

集合通信带宽（GB/s）	使用分布式通信测试框架在实际 GPU 集群上测量多卡/多机带宽，如 AllReduce 等典型集合通信操作；取 P95 值	大模型训练依赖频繁梯度同步，通信带宽不足会严重拖慢训练速度，甚至导致超时失败。该指标是衡量参数面网络效能的“黄金标准”
无损网络丢包率	通过交换机 Telemetry（如 gNMI）、网卡统计（rdmastatistic）或 eBPF 程序采集网络流量丢包比例，适用于 RoCE、InfiniBand 等要求无损传输的网络	RDMA 对丢包极度敏感——即使 0.1%丢包也会触发重传，使吞吐下降 90%以上。零丢包是 AI 网络的硬性要求
PFCPause 帧频率	统计单位时间内优先级流控机制触发次数；来自交换机或网卡计数器	频繁优先级流控表明网络拥塞，虽可避免丢包，但会引发“PFCStorm”导致全网停顿。需监控以优化拥塞控制算法（如 DCQCN）
TCP 重传率	通过/proc/net/snmp 或 eBPF 采集 TCP 重传段数/总发送段数	数据面（如数据加载、日志上传）依赖 TCP，高重传率反映网络质量差或配置不当，影响整体任务效率
网络故障收敛时间	从链路/设备故障发生到路由重新收敛、业务流量恢复的时间；可通过业务流探针或网络设备 Telemetry 数据综合判定	在大规模集群中，慢收敛会导致部分节点“失联”，引发训练任务卡死或资源浪费

AllReduce 同步完成时长（ms）	单次梯度同步任务从发起至所有节点完成数据交换的总耗时	该指标直接反映梯度同步的实际效率—即便底层网络参数达标，若节点调度不均、拓扑拥塞，仍会导致同步耗时飙升，直接拖慢训练迭代速度，是原有指标体系中缺失的结果型核心指标

来源：中国信息通信研究院

（3）计算层

AI 基础设施以加速器为核心，其规模可达千卡甚至万卡级别，硬件故障概率显著放大。同时，高密度集成架构（如多加速器共享供电/散热）下组件高度耦合，单点故障可能影响整体可用性。此外，AI 负载对硬件稳定性要求远高于通用计算，需系统性评估可靠性。

本节围绕单加速器可靠性、关键错误检测、高密度架构可用性三大维度定义指标项，明确度量方法，为硬件系统韧性评估提供依据。

表 3 计算层度量指标

指标	指标解释/计算方法（扩展）	选择该指标的原因
加速器月化故障率（MFMR）	MFMR=（当月故障 GPU 数/总 GPU 数）×100%；“故障”包括 XID 错误、驱动崩溃、显存 ECC 不可纠正错误等导致加速器无法继续执行计算任务	加速器是 AI 集群最昂贵且最易损组件。千卡集群下，即使 0.5%MFMR 也意味着每月 5 块卡故障，显著影响训练连续性

关键错误事件发生频率	从系统日志或硬件管理接口日志中提取关键错误事件（如 XID 错误）次数/月	关键错误事件（如 XID 错误）通常由 GPU 硬件异常（如电源、显存）引起，会导致进程崩溃，是 GPU 健康的核心预警信号
高密度集成架构整体可用率	高密度集成架构内任一组件故障即视为不可用；可用率=正常运行时间/总时间	高密度集成架构内部高度耦合，单点故障可能使整柜失效，更能反映架构整体健壮性。
计算节点可用性	可用于调度的计算节点占比	保障资源可被调度
服务器故障率	单位时间内故障服务器占比	衡量硬件基础设施的稳定性
GPU 卡可用时长比例（SLO）	集群内异常卡的可用时长累计/集群所有卡的总时长	大规模集群，精准衡量整体服务连续性；量化集群内 GPU 资源的整体服务保障能力
GPU 卡不可用时长	GPU 卡因硬件故障（如显存损坏、板卡短路）、环境异常（如散热失效、供电不稳）等问题，从异常被发现到修复完成并重新纳入集群调度的累计持续时长	定位运维流程瓶颈，优化故障响应策略

来源：中国信息通信研究院

（4）平台组件层

AI 场景下平台组件面临大规模任务调度、资源动态分配与 SLA 保障要求，需适配 AI 负载特征（如高并发启动、长周期运行）。调

度器、操作系统、镜像中心、可观测系统等平台组件需在通用能力基础上，强化 AI 负载适配性、资源保障能力与专属可观测性。

表 4 平台组件层度量指标

指标	指标解释/计算方法（扩展）	选择该指标的原因
任务调度延迟（P95）	从提交 Job 到所有 Pod Running 的时间分布第 95 百分位；适用于大规模 AI 任务。	能够反映调度器并行处理能力
任务调度成功率	调度成功的任务占比	从任务视角分析，可直接反映硬件资源是否可用以及反映资源碎片问题
大规模镜像拉取 P95 时延	从任务实例创建到启动完成的时间，针对大模型镜像采样。	能够反映镜像分发效率。镜像拉取是任务启动的关键路径，慢拉取会阻塞整个训练队列，尤其在弹性扩缩容时影响显著
硬件驱动/内核崩溃率	单位时间内因硬件驱动或系统内核导致的节点宕机次数。	驱动不稳定会引发整机重启，造成训练任务全量丢失，比应用层错误更严重
AI 核心指标可观测率	已采集的 AI 专属指标数/应采集总数（如利用率、通讯耗时、显存碎片率等）。	若无法观测关键 AI 指标，故障根因分析将严重滞后，延长 MTTR
高优任务抢占生效时间	从高优先级任务提交到低优先任务资源释放的耗时。	在共享集群中，需保障紧急任务（如线上推理扩容）能快速获取资

		源，体现平台 QoS 保障能力

来源：中国信息通信研究院

2.数据层

模型的知识受限于其训练数据，如果数据完整性、正确性存在问题，模型会继承并可能会放大这些缺陷，作为人工智能可靠性工程的“源头与底座”，需要坚持把数据质量放在首位，严把入口关、过程关、出口关，完善血缘追踪与来源核验、标注规范与质检机制，严格落实合规要求，确保数据来源可追溯、使用有依据、结果可信。评估指标体系可围绕“质量稳定性—分布一致性—覆盖充分性—合规可控”的逻辑进行开展。重要的关注评估指标包括标注一致率、高风险样本覆盖率、PII 合规通过率、数据时效性等指标（关键业务场景宜采用更高阈值），同时应建立指标的持续性监测能力，相关指标与建议基线可参考下表。

表 5 数据层度量指标

指标	指标解释/计算方法（扩展）	选择该指标的原因
标注一致率（Cohen/FleissKappa 或一致标签占比）	衡量不同标注人员对同一样本标签的一致性。Kappa 权重用于剔除随机一致概率影响。	控制数据来源偏差，保障模型学习目标不被噪声偏置，直接影响模型性能上限
标准准确率	指标注标签与样本真实标签的匹配比例，核心衡量标注结果的精准度，需区分人工标注准确率（针对人工标注数据）和	标注准确率是“源头中的源头”—数据是模型的基础，而标注的准确

	算法标注准确率（针对自动标注数据）	性直接决定数据的有效性
数据漂移度 PSI	对比线上实时分布与训练分布的变化；PSI=Σ(Observed－Expected%)×ln(...)	生产环境变化导致模型失效（数据断层/温度不足），此指标用于快速触发重训策略
分布差异 KS/Wasserstein	KS 检验 P-value 或 Wasserstein 距离衡量关键特征的概率分布差异。	模型推理阶段鲁棒性的根基是不改变分布假设，此指标用于防止概念漂移
高风险样本覆盖率	高风险业务类别在训练集占比	异常/极端场景缺失会导致重大故障（如金融欺诈未检出），必须严格保障覆盖性
数据清洗吞吐波动	清洗/特征流水线吞吐变化（相对标准差或变异率 CV）	保障训练/推理的数据供应链稳定性，避免性能回退
数据时效性延迟（实时性）	“产生→可用”端到端延迟统计	新鲜度不足导致模型与现实脱轨，是业务错判主要来源
PII 合规通过率	脱敏、最小化、访问审计等检查项通过率	隐私违规导致不可恢复法律与声誉损失

来源：中国信息通信研究院

3.训练层

训练层的稳定性保障面临与传统在线系统显著不同的技术挑战，模型训练过程具有典型的离线批处理特征，主要体现在计算资源的高效利用和训练过程的连续性保障方面，应以作业稳定性、分布式效率

与能效为核心，典型的评估观测指标包括计算资源利用率、有效训练时长占比等指标。

表 6 训练层度量指标

指标	指标解释/计算方法（扩展）	选择该指标的原因
有效训练时长	（总训练时间－无效耗时）/总训练时间×100%	在单个大型任务预训练场景中，需要综合考虑数据保存和回退、调度损失以及其他各类损失的情况
作业成功率	成功结束任务/全部任务	训练失败意味着资源浪费&迭代停滞，故须保持高稳定
Checkpoint 成功率	周期性存储成功占比	保证可恢复性，是训练不中断的“保险丝”
训练 MTBF	平均无故障持续训练时间	直接衡量分布式训练稳定性，节点故障会全局失败
任务中断恢复耗时（s）	指训练任务因硬件故障、网络中断、资源抢占等异常情况暂停后，从触发恢复机制到任务重新进入正常迭代状态的全流程耗时	该指标是保障训练连续性的核心量化参数，直接反映系统故障容错与恢复能力
GPU 利用均衡度（CV）	利用率变异系数	削峰填谷，提升加速比；避免 rank 卡顿导致同步阻塞
AllReduce 效率	通信带宽利用率、同步等待比	分布式训练瓶颈在通信层，此指标直接对应吞吐能力

梯度异常告警时延	异常出现→告警时间	“训练崩坏”是级联崩溃，需秒级发现与介入
吞吐/迭代时长回归	当前与基线比较	通过回归监测避免隐性性能退化、资源浪费

来源：中国信息通信研究院

4.推理层

推理层是人工智能模型对外提供服务的关键环节，服务典型特点包括：1）响应时间长（特别是大模型推理），较传统应用高 1—2 个数量级，冷启动耗时显著 2）资源消耗模式特殊，表现为显存占用大，计算密集型特征明显；3）服务模式多样，典型场景包括在线推理服务的实时请求—响应模式（含流式处理）、离线批处理推理的批量数据处理模式等。应以服务可用性与端到端时延为主要指标，整体服务稳定性在关注成功率基础上，更关注输出质量、响应延迟等体验指标。

推理层可靠性评估指标以服务成功率为基础，结合其典型服务特点，需额外关注首 Token 延迟、每 Token 生成延迟（TimeperOutputToken，TPOT）和整体延迟等指标，综合考虑多个维度衡量服务质量，另外建议同时考虑资源成本效率、安全合规等维度指标。除以上典型评估关注指标外，其他可参考的评估指标如下表所示。

表 7 推理层度量指标

指标	指标解释/计算方法（扩展）	选择该指标的原因
服务可用性（按分钟）	不可用时间/总时间	直接对应 SLA，任何停机即业务损失
端到端时延（P50/P95/P99）	包含网络、模型、依赖服务延迟	用户体验主要指标，对交互式模型至关重要

TTFT/每 Token 延迟	响应首 Token 和生成速度	控制 LLM 服务卡顿，是关键用户体验指标
错误率	5xx/功能失败/工具调用失败占比	服务链路故障会引发连锁中断，必须阈值告警与自动熔断
事实性/相关性评分	检索匹配率+人工/自动评分结合	降低 AI 幻觉导致的业务错误风险
有害输出率/越权调用率	减少违规、攻击与隐私泄露	安全合规“零容忍”，影响不可逆
成本指标（每请求成本）	Token 成本与硬件利用	直接影响大规模部署与商业化能力
漂移检测时长	性能退化到发现&定位耗时	推理阶段往往隐性失效，需快速止损
缓存命中率	KV 缓存命中占比	减少重复计算，显著改善时延与成本
混沌测试通过率	故障注入试验成功比	验证抗脆弱能力，防范未知故障模式
服务冷启动耗时（ms）	指推理服务从接收到首个请求，到完成模型加载、初始化并输出首个有效响应的总耗时；针对大模型，需区分模型权重加载耗时和上下文初始化耗时两个子项	该指标是衡量服务弹性和资源调度效率的关键参数

来源：中国信息通信研究院

5.AI 应用层

Agent 具备完成复杂任务的能力，在可靠性评估时重点关注其行为的可控性、准确性、安全性和高效性，确保其安全、可靠和有效运行。AI 应用层的可靠性评估是一个涉及功能、性能、安全和成本的综

合性功能，可以从以下四个维度入手：

1）任务完成度与准确性，建立一套标准任务评测集，持续度量在这些任务上的成功率和最终结果的准确性，在具备基本要求的前提下，可进一步关注用户意图满足率；

2）规划与推理稳定性，应对 Agent 的行动计划或者思维链进行监控，评估其逻辑的连贯性与执行的稳定性，指标可使用无效步骤率、推理循序检测率、平均任务步数、自主回复成功率等；

3）工具使用可靠性与安全性，严格监控所有工具调用的成功率、参数准确性，并不得允许任何未授权的操作，在指标上使用工具调用成功率、工具错误处理率、越权/危险操作尝试率等指标。

4）性能与成本效率，度量任务的端到端延迟与资源消耗成本，可以使用指标包括端到端任务延迟、首次行动时间、单位任务成本、有效成本占比等。

以 AI 应用支撑的智能客服场景为例，可设置几项典型业务指标，如：AI 应答采纳率>95%、AI 应答错误率<5%、AI 应答满意度>90%。

表 8 AI 应用层度量指标

指标	指标解释/计算方法（扩展）	选择该指标的原因
用户意图理解准确率	Agent 对用户输入指令的真实意图的识别符合度	意图理解是 Agent 完成任务的前置核心环节，若意图识别错误，后续的任务规划、工具调用、推理执行都会偏离目标

来源：中国信息通信研究院

云上人工智能系统可靠性度量评估应从业务目标规划的需求出发，按照自顶向下的设计逻辑，指标覆盖“目标（SLA）—承诺（SLO）—观测（SLI）”三级，并根据本评估建议中的相关分层设计针对性指标，按“关键/重要/一般”业务等级设定分级阈值与处罚/豁免条款。**另外，在人工智能系统的可靠性建设的同时，也需要考虑通用的可靠性建设经验，可参考中国信通院《分布式系统稳定性建设指南》**的相关建议完整构建云上人工系统可靠性工程的评估能力，如运维管控、安全设计、演练设计等。

四、云上人工智能可靠性实施路径与建设规划

（一）成熟度分阶段建设路径

为引导云上人工智能系统的可靠性建设从基础能力逐步迈向体系化、智能化，构建三阶段成熟度演进模型，包括**阶段 1：基础高可用、阶段 2：弹性自愈和阶段 3：智能预判。**该成熟度模型覆盖基础设施、数据与模型、推理服务、AI 应用以及可靠性管理与治理等维度，通过分阶段目标和能力项建设，使组织能够在业务发展和技术条件允许的前提下，以可控节奏持续提升云上人工智能系统的可靠性水平。

组织在规划云上人工智能可靠性建设时，应结合自身业务重要性、技术基础与人才储备，明确当前所处成熟度阶段与目标阶段，制定分阶段建设计划和里程碑，逐步实现从“基础高可用”向“弹性自愈”再到“智能预判”的演进路径。如下表所示：

表 9 云上人工智能可靠性成熟度等级与建设重点

成熟度阶段	目标定位	关键能力要点	典型关注指标举例

阶段 1 基础高可用	保证核心服务基本不中断、故障可恢复	冷备集群/双机房、基础资源监控、人工切换与恢复	服务可用性、基础资源利用率、故障恢复时间
阶段 2 弹性自愈	实现主要故障场景下的自动扩缩与自愈	自动扩缩容、训练容错流水线、金丝雀发布、混沌工程	P95/P99 时延、错误率、自动恢复成功率
阶段 3 智能预判	通过智能分析与预测实现全局最优	AIOps、故障预测、预测性维护、智能根因分析与调优	预测命中率、业务影响降幅、可靠性评分

来源：中国信息通信研究院

1.阶段 1：基础高可用

阶段 1 主要面向处于起步或探索阶段的云上人工智能系统，目标是在有限投入下实现“服务基本不中断、故障可恢复、问题可感知”的基础高可用能力。

在基础设施层，应至少构建单地域多可用区部署或同城双中心架构，对关键组件（如模型推理服务、数据存储集群）配置冷备集群或热备实例，具备人工干预或半自动的故障切换能力；对训练集群，应至少具备节点级故障恢复和作业重启机制。

在数据与模型层，应建立基础的数据备份与恢复机制，对关键训练数据、模型参数和特征库实行周期性备份，并保障备份的完整性与可恢复性。模型发布应遵循最小变更原则，重要变更应通过人工评审和小流量试运行进行验证。

在模型推理层，应部署基础监控系统，采集 CPU、内存、磁盘、网络等资源指标，以及服务存活状态和接口错误率等运行指标，构建

最小可用的告警体系，对服务中断、错误率异常和资源耗尽等典型故障实现及时告警。

在 AI 应用层，应明确基本的服务等级目标，如服务等级协议（Service Level Agreement,SLA）、服务等级目标（Service Level Objective,SLO）和服务等级指标（Service Level Indicator，SLI）的初始基线，并建立值班、故障记录和复盘机制，对重大故障形成经验沉淀。

达到阶段 1 时，组织应能够在常见基础设施和服务故障场景下保障核心业务不发生长时间中断，具备基础的监测、告警与人工恢复能力。

2.阶段 2：弹性自愈

阶段 2 面向已具备基础高可用能力、业务规模和复杂度持续提升的云上人工智能系统，目标是从“人工可恢复”迈向“系统自愈与弹性伸缩”，在保证可靠性的同时提升资源利用率和运维效率。

在基础设施层，应实现自动化扩缩容与跨集群调度能力，根据负载和队列状态动态调整推理实例和训练任务资源分配，避免资源长时间空转或拥塞。在 GPU 集群中，应支持节点级和微架构级故障的自动隔离和迁移，对抢占式或竞价实例应提供优雅退出机制，避免训练过程中的状态丢失和推理服务的大规模抖动。

在数据与模型层，应逐步引入机器学习运维（Machine Learning Operations，MLOps）能力，对数据采集、预处理、训练、评估与发布建立统一流水线。流水线应支持自动化的检查点保存、失败重试、作

业回滚以及模型性能回归测试，当发现训练失败、评估指标异常或数据漂移越界时，系统应自动阻断发布并触发告警与人工干预。

在推理服务层，应完善弹性伸缩策略与熔断限流机制，并对关键指标（例如 P95/P99 延迟、错误率、业务转化率等）绑定服务等级目标（SLO），当 SLO 越界时自动驱动扩容、降级或灰度回滚。应引入金丝雀发布和蓝绿部署策略，将变更风险控制在可接受范围内，并定期开展混沌工程实验，通过主动注入故障检验冗余与自动恢复机制的有效性。

在 AI 应用层，应将 Agent 视为可管理的工程实体，对其策略、工具调用链路和外部依赖进行显式建模，建立行为监控与异常审计能力。当 Agent 出现连续异常决策或高风险操作时，应自动收紧其行为边界，将其退回到低权限模式或静态规则模式，并通知运维与业务人员进行复核。

在组织与治理层，应逐步引入 SRE 方法，将可靠性目标量化为错误预算、变更频率与事故率等指标，并在规划阶段、变更阶段和运行阶段持续对齐业务价值与可靠性投入。同时应当在关注传统 SRE 的“可用、性能、容量”等能力的同时，增加关注数据模型层、推理服务层的“正确、无害、无偏、可控、可审计”等能力。

平台应支持对如下 AI 特有 SLI 的持续评测：

1）时延：首 Token 延迟、时延（P50/P95/P99）、每 Token 延迟；

2）质量：事实性/相关性、检索匹配率（RAG）、一致性（同问多答的差异）；

3）安全：毒性率、越权工具调用率、提示泄漏率；

4）成本：每千 Token 成本、单位查询 GPU 时长。

指标越界应联动流量治理（限流/熔断/降级）与版本回滚。

达到阶段 2 时，组织应基本实现核心业务的自动扩缩、训练作业的失效恢复与模型发布的风险可控，MTTR（Mean Time To Recovery，故障平均恢复时间）显著下降，可靠性工作从“事后恢复”部分转向“事前预防与事中自愈”。

3.阶段 3：智能预判

阶段 3 面向在云上全面推进人工智能应用，并希望实现“智能自治与全局最优”的组织，目标是通过智能运维和故障预测能力，把云上人工智能可靠性从规则驱动提升至数据驱动与模型驱动，实现从被动响应向主动预判的转变。

在基础设施层，建立基于历史告警、日志与运行指标的预测性维护能力，对 GPU、加速网络、存储阵列等关键组件构建健康评分模型，通过异常模式识别与趋势分析，提前发现潜在故障并在用户感知前完成资源迁移或替换。

在数据与模型层，引入持续评估与在线学习能力，对数据质量、模型性能和业务指标进行联合建模，形成漂移检测、性能退化预测与自动调度的闭环。对高价值场景应构建多模型冗余与专家模型池，当主模型预测结果置信度不足或被识别为不可靠时，由备份模型或规则体系接管决策，以降低单模型故障对业务的影响。

在推理服务层，应通过端到端可观测性与智能根因分析技术，对

日志、指标、调用链、用户反馈等多模态数据进行融合分析，自动定位性能瓶颈与故障根因，并给出修复建议或直接下发自动化变更。对大规模多租户环境，应在成本、时延与可靠性之间进行动态平衡调度，实现跨区域、跨集群的全局优化。

在 AI 应用层，应建设 AI 应用全生命周期的验证与治理体系，对 Agent 的策略演化、工具调用和跨系统动作建立可度量指标，包括决策稳定性、合规性、伦理风险与用户信任度等。当系统检测到行为模式异常或伦理风险上升时，应自动收紧权限、提高审核比例或中止高风险任务。

在组织与治理层，应实现可靠性指标与业务指标的深度联动，将 SLA、SLO 与业务收入、用户满意度、合规风险等指标结合评估，形成以业务价值为导向的可靠性投资决策体系。可靠性治理应实现从项目级、系统级向组织级、生态级扩展，通过跨团队协同与统一工具链支撑多业务、多平台的统一可靠性管理。

（二）关键能力建设清单

构建高可靠的云上人工智能系统，需围绕基础设施层、数据层、训练层、推理层四大核心层级，系统性布局关键能力，精准应对各阶段的核心挑战。数据可信是模型可靠性的原始基础，对于数据来说，数据层面临的核心挑战是数据的质量与合规性问题。在实际模型训练阶段，面临的核心矛盾是昂贵的计算资源与有效训练市场之间的矛盾。在模型应用阶段，聚焦与服务可靠性与用户实时体验的平衡。推理层则聚焦服务可靠性与用户实时体验的平衡，在应对区域故障、模型漂

移和流量突增的同时，实现低延迟、高可用、低成本的服务交付。

1.基础设施层

基础设施层是支撑高可靠、高性能 AI 训练与推理的基石，需围绕**硬件可靠性、弹性调度、网络效率与数据持久化**四大维度构建端到端能力。

首先，**应建立硬件交付前的全链路可靠性压测验收流程**，对整机服务器、GPU 加速卡、高性能网卡（如 RoCEv2）等核心组件进行长时间高负载压力测试，确保其在典型 AI 工作负载下的稳定性。同时，在运行阶段部署细粒度的硬件健康监控体系，实现故障的秒级检测与精准告警，并与上层调度平台联动，自动迁移任务、释放异常节点，快速恢复因硬件失效导致的训练或推理中断。

其次，面向超大规模 AI 场景，**需构建异构计算统一资源池**，将 CPU、GPU、TPU 等计算单元纳入统一调度平面，通过智能调度器实现跨架构资源的秒级弹性扩缩容，满足突发性训练任务或推理流量的敏捷需求。

第三，**为加速大模型分布式训练，须打造超低延迟高性能网络底座**，全面部署 RDMA（Remote Direct Memory Access）网络，端到端通信延迟控制在 2 微秒以内，并原生支持 AllReduce 等集合通信操作，高效完成千亿级参数的梯度同步。最后，**依托高可用分布式存储系统**，实现关键训练数据、模型 Checkpoint 与日志的三副本跨可用区（AZ）持久化存储，提供不低于 99.999%的数据可用性与强一致性保障，确保在单点或区域级故障下数据不丢、服务不中断。通过上述能力协同，

基础设施层方能为上层 AI 应用提供坚实、敏捷、可靠的运行环境。

2.数据层

数据层作为模型智能的源头活水，其核心使命是确保高质量、高可信、高合规的数据供给。为此，需构建一套融合智能治理、隐私安全与闭环质量控制的综合能力体系。

首先通过**智能数据湖引擎**实现跨业务数据孤岛的解耦与动态融合，建立基于元数据治理的异构数据协同框架，结合生成式样本补全机制消除场景覆盖盲区；同步**构建自动标注纠偏系统**，融合多模型交叉验证与领域知识规则库，形成闭环的标注污染识别与修复能力，确保数据生产的本质可靠性。

其次，基于隐私计算网关架构，采用联邦学习与同态加密技术构建数据“可用不可见”的隐私保护范式，实现算法层与基础设施层的协同防护；通过合规审计中台建立敏感数据全链路监控机制，结合实时风险阻断与区块链溯源能力，形成可验证的合规性保障体系。

3.训练层

训练层作为大规模模型能力生成的核心引擎，必须突破传统训练范式的稳定性、扩展性与效率瓶颈，构建面向超大规模、高可靠、自适应的智能训练体系。**从容错恢复、并行扩展、运行优化三个维度系统化打造关键能力。**

首先，为应对分布式训练中因硬件故障、网络抖动或节点驱逐导致的长时间任务中断问题，应**建设训练中断预测与高效恢复机制。**通过实时监测节点健康状态（如 GPU ECC 错误率、网络丢包率、存储

I/O 延迟），结合历史故障模式构建轻量级预测模型，在故障发生前主动触发保护措施；同时，全面落地增量快照技术（Delta Checkpoint），仅保存相对于上次检查点的参数变化量，并融合三级压缩策略（量化压缩+稀疏编码+通用无损压缩），将检查点（Checkpoint）体积降低 80%以上、写入耗时缩短至秒级。该机制与 Kubernetes 跨可用区（AZ）智能调度深度集成——当检测到某 AZ 节点异常时，调度器可自动将训练任务迁移至健康 AZ，并从最近增量快照快速恢复，实现分钟级故障切换，最大限度减少训练进度损失。

其次，针对千亿乃至万亿参数模型远超单机算力极限的挑战，需**构建智能化三维并行训练架构**。系统应支持自动切分策略，根据模型结构、数据规模与硬件拓扑，动态组合数据并行、流水线并行与张量并行，实现计算、通信与内存负载的全局最优分配。在此基础上，引入拓扑感知通信优化机制：优先利用服务器内部 NVLink 高速互联进行 GPU 间梯度同步，跨节点通信则通过 RoCE RDMA 网络进行聚合归约，并结合梯度分桶、通信重叠计算等技术，显著降低 All Reduce 通信开销。该能力使千卡集群的训练效率线性加速比提升至 85%以上，持续推动大模型训练速度边界。

最后，为减少对人工干预的依赖、提升训练鲁棒性，需**建立实时监控与自适应调节闭环**。平台持续采集关键训练指标，如梯度范数（用于检测梯度爆炸/消失）、GPU SM 利用率、显存带宽饱和度、Loss 震荡幅度等，一旦发现异常模式（如梯度范数突增 10 倍、GPU 利用率骤降至 20%以下），即触发自适应调节机制：自动降低学习率以稳定

收敛、冻结异常层参数或在 K8s 层面迁移任务至新节点。该机制不仅能有效抑制训练发散，还可将大规模训练任务的整体失败率降低 70%以上，显著提升资源利用效率与工程交付确定性。

4.推理层

推理层作为连接模型能力与终端业务价值的最终出口，必须兼顾高可用、高适应性与高成本效益。为此，需围绕**服务韧性、模型稳定性与资源协同三大核心维度**，构建智能化、自适应的推理服务体系。

首先，为应对区域性基础设施故障（如可用区断电、网络分区）或突发流量冲击导致的服务中断风险，推理层应**建设基于 Envoy 的动态流量调度体系。**该体系实时采集各可用区（AZ）实例的延迟（P99 TTFT/TPOT）、错误率、GPU 利用率等 SLI 指标，通过加权轮询算法动态调整路由权重——当某 AZ 性能劣化或错误率超标时，自动降低其流量比例甚至完全切流；同时，结合常态化混沌工程主动测试（如模拟 AZ 级网络隔离、注入高延迟），验证多活架构的容灾切换能力。在此机制支撑下，系统可在秒级内完成故障 AZ 的流量迁移，并弹性承载高达 10 倍的突发请求，确保关键业务连续可用。

其次，针对模型上线后因数据分布漂移（如用户行为突变、传感器校准偏移）引发的预测偏差或业务误判问题，推理层需**嵌入实时数据健康监测与自动响应闭环。**通过在线计算输入特征与训练基线之间的 KL 散度（Kullback–Leibler Divergence）或 PSI（Population Stability Index），量化分布偏移程度；一旦指标超过预设阈值，系统立即触发分级响应：轻度漂移启动增量训练流水线，利用最新样本微调模型；

重度漂移则自动将流量切换至更鲁棒的备用模型版本（如轻量版或规则引擎），并在后台并行执行全量重训练。该机制将“模型退化—业务受损”的被动响应转变为主动防御，显著提升线上决策的长期可靠性。

最后，为优化跨地域部署的资源效率与用户体验，推理层应**实施智能分层计算策略**。对计算密集型模块（如大语言模型的注意力层、向量编码器），采用动态分片技术，按请求复杂度或设备能力决定是否卸载至云端高性能 GPU 集群；同时，在边缘节点（如 CDN POP 点、区域网关）部署缓存预热与模型轻量化机制，对高频请求的上下文、常用提示模板及低秩适配器（LoRA）进行本地缓存。该架构不仅大幅降低骨干网带宽消耗（尤其在多模态生成场景），还将边缘侧首 Token 延迟压缩至百毫秒级，实现“近用户、低时延、低成本”的智能服务交付。

5.AI 应用层

AI 应用层是模型能力落地终端业务的核心载体，核心目标是保障任务精准达成、用户意图有效满足，需围绕意图理解、任务执行、工具协同、安全合规四大关键维度，构建高可靠的应用服务体系。

其一，搭建多维度意图理解引擎，融合语义解析、上下文关联与领域知识校验能力，精准识别模糊指令与隐含需求，从源头规避意图偏差导致的任务执行错误；同步实时监测意图理解准确率，确保复杂业务场景下稳定达标 95%以上。

其二，构建自适应任务规划框架，支持复杂目标的子任务拆解与

动态优先级调度，配套子任务失败重试、替代方案匹配等容错机制，保障端到端任务完成率不低于 99%。

其三，建立安全可控的工具调用中台，实现工具注册鉴权、参数合法性校验、全链路操作审计的全流程管控，同时支持多工具协同调度与结果融合，有效规避越权操作、恶意输入等业务风险。

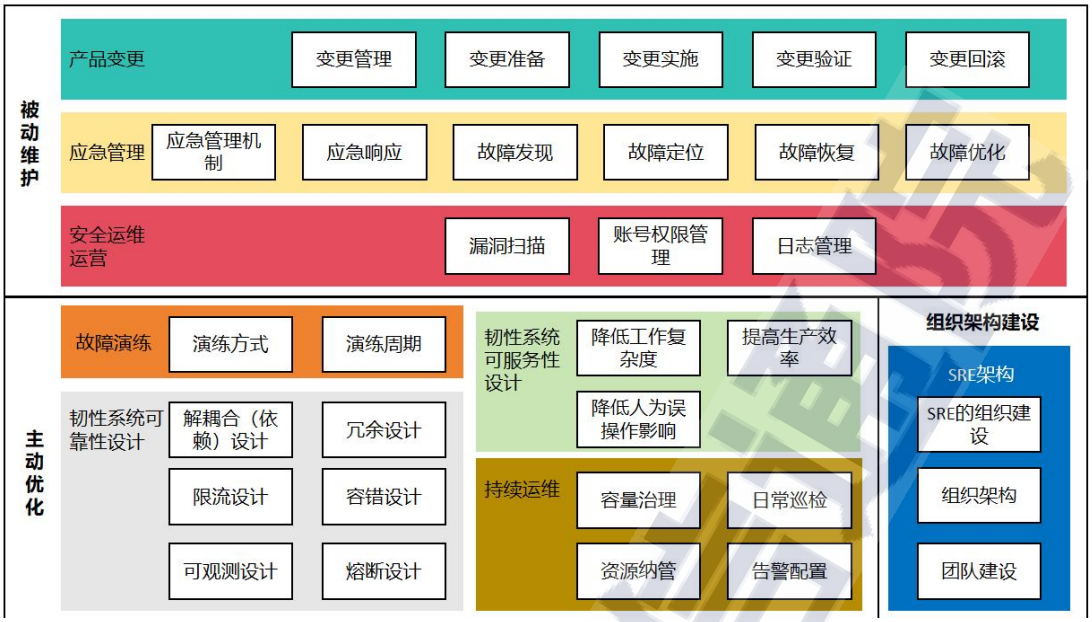
其四，打造全链路体验与效果评估闭环，实时监测任务响应时长、意图满足率等核心指标，联动业务目标（如转化率提升、差错率下降）持续优化模型与应用策略，实现可靠服务与业务价值的协同提升。

（三）优化组织与治理机制

1.SRE 团队能力建设路线

SRE 系统架构通过强调可靠性、自动化运维、快速故障恢复、弹性扩展、持续改进以及业务与技术的紧密结合，提高系统的稳定性、可维护性和灵活性。它利用实时监控和自动化工具，迅速发现和解决问题，最大限度地减少系统故障时间，保证业务连续性。同时，SRE 系统架构具备弹性，能够根据负载需求进行快速扩展或收缩，确保系统在高压力和低负载下都能有效运行。通过持续改进和优化，SRE 不断提升系统性能和用户体验，促进业务的可持续发展。

随着云上人工智能系统的复杂度持续提升，传统的被动运维模式已难以满足高可用、高可靠、高弹性服务的保障需求。SRE 能力正从以“故障响应”为核心的被动维护，向以“系统韧性设计”和“主动优化”为导向的演进路径迈进。如图所示：



来源：中国信通院《服务韧性工程（SRE）能力要求》

图 4 服务韧性工程（SRE）能力要求

在被动维护层面，涵盖产品变更管理、应急管理及安全运维三大模块，分别对应变更的全生命周期管控、故障发生后的响应与恢复流程，以及漏洞扫描、权限管理和日志审计等基础安全运营活动，体现的是对系统运行状态的事后响应与保障。而在主动优化层面，通过韧性系统设计（如解耦、冗余、限流、可观测性）、故障演练（包括演练方式与周期）以及持续运维（容量治理、告警配置、资源配管等），推动系统从“可修复”向“自愈”和“防患于未然”演进，强调在设计阶段即构建高可用与高可靠能力。同时，组织架构建设部分，明确指出 SRE 需依托科学的组织结构、规范的协作机制与团队能力建设，才能实现理念落地与长期可持续发展。整体呈现了从“救火式运维”到“预防性治理”的系统性演进逻辑，体现了现代 SRE 从技术实践向组织化、体系化能力升级的核心趋势。

2.混沌工程常态化实施框架

混沌工程是一种通过主动引入故障和异常来测试系统韧性和可靠性的方法。其核心理念是通过在实际环境中模拟各种可能的故障场景，验证系统在这些极端条件下的表现，从而发现潜在的脆弱点和改进空间。

在 AI 原生时代，随着人工智能技术的广泛应用，AI 服务的可靠性和稳定性变得尤为重要。传统的混沌工程方法需要进一步演进，以应对 AI 服务的复杂性和不确定性。AI 服务通常涉及复杂的模型推理、数据处理和分布式计算，任何一个环节的故障都可能导致服务中断或性能下降。因此，AI 服务能力演练成为确保 AI 系统可靠性的关键环节。通过混沌工程，可以模拟各种可能的故障场景，如模型推理失败、数据传输中断、资源抢占等，验证 AI 系统在这些情况下的表现，确保其在真实环境中的稳定性和可靠性。

混沌工程实施应以安全可控为前提、场景覆盖为基础、闭环优化为驱动，通过明确边界约束、构建标准化注入体系与持续迭代评估机制，系统性提升云上 AI 系统的韧性验证能力与应急响应水平。一是明确实施边界与安全基线，严格遵循业务运行规范及系统安全要求，聚焦核心交易、关键链路等核心业务场景，提前设定资源隔离、故障影响范围、回滚机制等硬性约束条件，全程规避对生产环境稳定性及业务连续性的负面影响。二是构建标准化故障注入体系，全面覆盖基础设施（如服务器、网络）、AI 应用层（如人脸稽核服务）、数据层等多维度故障场景，设计可复现、可量化的故障注入流程，制定操作

权限管控、执行前校验等规范，确保故障注入过程安全可控。三是建立闭环评估与优化机制，通过全链路监控工具实时采集系统指标、业务数据，结合预设基准值分析故障影响范围与系统韧性表现，形成问题整改清单与优化方案。同时持续迭代故障场景库与实施流程，将实践经验转化为标准化规范，不断提升系统抗风险能力与 SRE 团队应急处置效率。

3.可靠性文化培育机制

“可靠性文化”建立的核心是推动组织认知从“被动应对故障”向“主动预防风险”深度转变，以“全员参与”与“持续改进”为双轮驱动，构建上下协同、全员共建的文化生态。

理念宣贯夯实认知根基。通过内部培训、案例分享会、专题讲座等多元形式，普及可靠性核心知识与实践价值。拆解典型故障案例的根源与影响，传递“隐患即事故”的风险意识，让“追求可靠、重视预防”的理念渗透到研发、运维、产品等各岗位日常工作中。

实践赋能提升全员能力。搭建混沌演练平台，开展故障演练、风险复盘、可靠性评审等常态化活动。鼓励跨部门协作参与混沌工程实验、系统韧性测试，让员工在实战中掌握风险识别、隐患排查与应急处置技能，将文化认知转化为实操能力。

激励机制激发参与热情。建立专项激励体系，对主动发现重大隐患、提出有效优化方案、在可靠性建设中表现突出的团队与个人给予表彰奖励。形成“人人重视可靠、主动参与改进”的正向循环，最终筑牢系统可靠根基，护航业务长期稳定发展。

五、未来展望

（一）AI 驱动下的可靠性技术：从单点保障走向智能协同

随着运维领域 Agent 的发展，未来必将进一步向 AI 驱动的运维模式转变，**系统可靠性建设将突破传统边界，从单一组件保障转向全链路协同、人机智能互补、多 Agent 协作的复合型能力构建**，通过技术创新与模式升级，实现可靠性体系的迭代与进化。

1.可靠性边界外扩

未来，可靠性建设将不再局限于局部系统的稳定性保障，而是聚焦跨域、跨层级、跨场景的协同可靠性，构建“全链路感知—智能决策—协同执行”的一体化保障体系。核心方向包括：一是打通端侧设备、边缘节点与云端服务的可靠性数据链路，实现跨层级的故障预警与风险传导防控；二是针对多场景融合的复杂业务（如智能运维、自动驾驶、工业互联网），建立端到端的可靠性指标体系与协同校验机制，确保各环节的技术标准、接口规范与安全策略的一致性；三是依托 AIOps 技术，实现从“单点修复”到“全局协同优化”的升级，通过智能调度、资源动态分配与故障协同处置，提升系统在复杂协同场景下的容错能力与抗干扰能力。

2.人机智能互补

尽管 AIOps 与自动化愈合技术将日益成熟，但在处理极端复杂的未知异常和涉及重大伦理决策的场景时，人类的经验与价值判断仍不可或缺。未来的可靠性体系将更加强调构建高效、可信的人机协同

（Human-in-the-Loop）监督与干预机制，将其作为保障系统安全的最后一道防线。建设重点将从纯粹的自动化转向增强人类专家的决策能力，一是发展深度可解释性（XAI）的诊断界面，当 AI 系统检测到异常时，不仅提供告警，更能以人类可理解的方式（如决策归因图、自然语言摘要）呈现根因假设，辅助专家快速研判。二是建立分级干预与授权通路，针对不同风险级别的 AI 自主操作（如模型回滚、流量切换、数据清理等），设计从自动执行、人类确认到强制人工接管的多级干预流程，确保重大决策的审慎性与可追溯性。这标志着可靠性建设的终极目标并非完全取代人，而是构建人与 AI 能力互补、责任共担的共生系统。

3.多 Agent 协作

随着 AI 应用从单一辅助工具向自主 Agent（Autonomous Agents）网络演进，未来的可靠性挑战将从“单体智能”转向“群体智能”的交互稳定性。当多个 Agent 在云端进行自主协作、交易或谈判时，可能会产生预料之外的“涌现行为”，例如死锁、循环交互甚至群体性恐慌（类比金融市场的闪崩）。未来的建设方向需探索建立 Agent 交互协议标准与社会化可靠性契约，在 Agent 之间设立明确的沟通规范与冲突解决机制。同时，研发针对多 Agent 系统的“宏观监视器”，不仅监控单个 Agent 的健康度，更要监控整个 Agent 生态系统的宏观稳定性，防止局部决策偏差在群体交互中被非线性放大，引发系统性崩溃。

（二）企业 AI 可靠性管理：迈向治理、价值与韧性新形态

在云原生与 AI 融合背景下，**企业 AI 可靠性建设必将向全域治理、价值驱动、韧性底座的新阶段演进**，转向全流程可控、成本精益平衡、资产合规可持续的体系化能力构建，通过治理框架升级与运营模式创新，**实现企业级 AI 可靠性体系的整体跃迁**。

1.智能原生治理框架

企业管理体系中，人工智能的引入正推动治理模式从“被动支撑 AI 应用”向“智能内生于治理架构”深刻演进。未来的 AI 可靠性管理，不再仅是技术团队的专项任务，而是企业战略、风险控制、合规治理与运营效率协同升级的核心组成部分。企业需以成熟的云原生治理框架为基础，在组织制度、流程规范与责任机制中系统性嵌入 AI 全生命周期管理要求——包括模型开发伦理准则、数据使用合规边界、版本变更审批机制、服务可靠性承诺（SLA）以及模型退役与审计追溯等关键环节，确保 AI 能力在可控、可信、可问责的前提下赋能业务。

这一转型要求企业在架构治理上实现更高水平的融合：一方面，将 AI 纳入统一的企业技术治理体系统一管理，明确业务部门、数据团队、算法工程师与合规官之间的协作界面与权责分工；在治理架构中增设“专门的 AI 结果审核团队”，清晰界定其与业务部门、算法团队、合规官的协作界面及权责分工。另一方面，通过标准化流程（如将模型上线纳入变更管理审批等）、自动化控制（如在发布流程中嵌入公平性与安全检查等）和常态化验证（如定期开展 AI 红队演练与

韧性评估等），把抽象的政策要求转化为可执行、可度量、可审计的操作实践。同时搭建从数据输入、决策执行到业务影响的全链路监控与反馈体系，使管理层能够实时掌握 AI 系统的运行健康度与潜在风险，实现从“事后追责”到“事前预防、事中干预”的治理跃迁。

2.成本与业务价值

未来，云上人工智能的可靠性建设将不再是单纯的技术投入，而是与业务成本、效率、创新速度紧密耦合的战略决策。极致的可靠性往往伴随着高昂的计算资源、冗余资源和人力成本。因此，建立一套量化“可靠性投入产出比”的财务运营（FinOps）框架将至关重要。一方面，通过精细化成本归因，将可靠性成本（如多活架构、数据备份、混沌工程演练等）与具体的业务场景、模型服务乃至单次 API 调用的价值进行关联，使技术决策能够基于清晰的业务影响评估。另一方面，发展可靠性等级与成本的动态寻优能力，允许业务根据不同阶段、不同场景的重要性，在平台预设的可靠性策略（如“成本优先”“延迟敏感”“最高可用”等）之间灵活切换。这将推动可靠性建设从“一刀切”的最高标准模式，转向基于业务价值和风险容忍度的、数据驱动的、动态的、精益化的资源配置与投资模式，实现技术韧性与商业效益的最优平衡。

3.企业经营韧性

在云原生+AI 深度融合的趋势下，AI 系统的可靠运行越来越取决于其所依赖的数据、知识与内容资产是否具备稳定、可持续的合规来源与可控边界。合规性在数据采集环节已成为稳定运行的先决条件，

一旦失守将带来业务中断、法律追责与声誉损失。

展望未来，这一约束将进一步从采集是否合规扩展到知识与内容资产是否可持续运营。一方面，企业将更加频繁地引入外部数据、知识库、第三方模型能力与工具生态，以支撑更广泛的业务智能化。另一方面，生成式能力在业务输出层的影响持续扩大，使得内容质量、来源可核验、使用权边界与风险可控成为运营层面的硬约束。与传统系统不同，AI 的输出本身可能成为业务结果与外部承诺的一部分，一旦在关键时点触发合规争议、内容风险或权属不清，可能被迫下线、暂停或大范围回收，从而形成新的可靠性中断形态。未来规划应从更宏观的治理视角，把知识与内容资产纳入可靠性底座，将其作为与计算资源、数据同等重要的生产要素进行统一治理，形成可持续的资产准入、变更与退场规则。在组织层面明确业务、数据、算法与合规之间的责任边界与协作机制，使可靠性建设不仅“可用”，更能“可持续”。这一方向与本指南所强调的在制度与流程中嵌入数据使用合规边界等要求一致，但更强调面向长期运营的资产治理与外部风险承压能力，将可靠性从工程指标进一步提升为企业级经营韧性。

（三）跨行业 AI 可靠性：标准与生态协同新阶段

面向未来，跨行业云上人工智能的可靠性治理生态将从静态合规向动态适配、场景驱动的深度融合演进，亟须构建一套“通用底座+弹性适配”的分层标准体系，以兼顾技术统一性与行业特殊性。在基础设施、数据模型、推理服务等基础层应推动形成跨行业共识的统一技术规范与接口标准。

1.Agent 统一接入协议的规范

未来，规范 Agent 接入形式是跨行业 AI 可靠性治理的基础方向，核心是建立统一的 Agent 接入协议，实现 skills、MCP、A2A 等接入形式的标准化升级。趋势上，需明确运维 Agent 的标准化能力边界，涵盖场景适配与运维内容两大核心；对 skills 建立分层分责体系，明确权责与审核机制；规范 MCP 作为数据获取标准通道的交换契约，统一云厂商数据要素规范，适配多云场景发展需求；依托运维平台建立 A2A 通用可观测数据规范，为多类型 Agent 高效接入奠定基础，推动接入流程简化、兼容性提升。

2.底层协议的可靠性协同与统一

底层协议的统一与协同是未来跨行业 AI 可靠运行的核心支撑，也是技术发展的必然趋势。当前协议杂乱、适配性差的问题，将逐步通过构建统一底层协议体系解决，重点聚焦资源调度、模型元数据、推理指标、数据血缘等核心协议的标准化。未来，将推动各类协议无缝协同，统一可观测数据传输交互规范，破解多云、多平台数据结构不一致难题，降低接入适配成本，为跨行业 AI 规模化、高质量应用提供稳定的协议支撑，助力技术落地效率提升。

3.AI 可靠性治理生态构建

构建多方协同治理生态是跨行业 AI 可靠性发展的长远趋势，核心是明确政府、云厂商、企业、科研机构的角色分工，形成“标准—测试—认证—应用”的闭环体系。未来，政府将强化顶层政策引导，云厂商提供标准化技术支撑，企业落实主体责任，科研机构赋能前

沿技术研发；同时完善 skills 分层分责机制，通过标准化协作协议界定各方权责，将合规要求转化为可量化指标，最终形成全链条协同生态，推动跨行业 AI 向高可用、高鲁棒、可解释、可持续的方向迭代升级。

编制说明

本建设指南自2025年启动编制，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重要阶段，均面向人工智能可靠性领域的技术提供方、产品服务方、行业应用方开展了深度访谈、意见征集等工作。参与编制的单位说明如下：

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：小米科技有限责任公司、阿里云计算有限公司、蚂蚁科技集团股份有限公司、小红书科技有限公司、北京百度网讯科技有限公司、北京同创科技有限公司、京东科技信息技术有限公司、中信建投证券股份有限公司、四川农商联合银行、中移（苏州）软件技术有限公司、中国移动通信集团浙江有限公司、中国移动通信集团广西有限公司、中国移动通信集团有限公司、荣耀终端有限公司、联通数字科技有限公司、南京星邺汇捷网络科技有限公司。

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

网址：www.caict.ac.cn