ICS 03.120.01
CCS L 05

T/CICC

中 国 指 挥 与 控 制 学 会 团 体 标 准

T/ CICC 35012—2025

复杂智能系统可靠性技术要求

Technical requirements for reliability of complex intelligent systems

2025-09-12 发布　　2025-09-12 实施

中国指挥与控制学会　发布

目　　次

前　言

本文件按照GB/T 1.1-2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国指挥与控制学会提出并归口。

本文件起草单位：北京航空航天大学、杭州市北京航空航天大学国际创新研究院（北京航空航天大学国际创新学院）、可靠性与环境工程技术重点实验室、北京航空航天大学可靠性工程研究所、中国船舶集团有限公司综合技术经济研究院、中国兵器工业软件工程与评测中心、中国电力科学研究院有限公司、中国航空综合技术研究院、中国科学院声学研究所、中国电子科技集团公司第十研究所、中国航空系统工程研究所、中国船舶集团有限公司系统工程研究院、四川治为科技有限公司、长龙航空维修工程有限公司、华威大学、中国农业大学、浙江荷湖科技有限公司。

本文件主要起草人：杨顺昆、吴梦丹、郝程鹏、杨诚、刘虹晓、翟亚宇、王栓奇、周怡婧、侯展意、段峙宇、司昌龙、林聪、安冬、仇树茂、赵星宇、马欣瑞、张靖、李思远、齐晓琳、文佳、王若、马庆、欧阳荷清、赵宇熙、冯吉开、曾康、杨懿、黄海驰、赵诣深、张昱昊、张耀星。

复杂智能系统可靠性技术要求

1 范围

本文件规定了复杂智能系统可靠性的核心对象、技术要求、全生命周期及主要活动和分析、设计、训练、测试、验证以及不确定性量化方法。

本文件适用于在开放、不确定环境中运行的复杂智能系统的可靠性要求、测评与验证活动。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41867—2022 信息技术 人工智能 术语

GB/T 42018—2022 人工智能平台 计算资源规范

GB/T 45958—2025 网络安全技术 人工智能计算平台安全框架

ISO/IEC 5259-1:2024 分析和机器学习的数据质量——第1部分：概述、术语、框架（Data quality for analytics and ML — Part 1: Overview, terminology, framework）

3 术语与定义

GB/T 41867—2022、GB/T 45958—2025、GB/T 42018—2022和ISO/IEC 5259—1确立的以及下列术语和定义适用于本文件。

3.1

复杂智能系统 complex intelligent systems

由感知、认知、决策与执行等功能模块构成，采用机器学习等方法，在不确定、开放环境下执行任务的人机环管协同系统。其复杂性体现在多源异构数据、动态场景、要素耦合以及全生命周期演化。

3.2

智能系统可靠性 reliability of intelligent systems

在规定的使用环境与时间周期内，智能系统在面临数据分布漂移、对抗干扰或组件故障时，维持其功能性能在可接受范围内的能力。

[来源：GB/T 41867—2022，3.4.4，有修改]

3.3

AI平台 platform for AI

为AI应用提供计算、存储、网络与开发运维能力的软硬件与服务集成，包括但不限于CPU、GPU、AI加速器、系统软件、中间件、框架与接口等。

[来源：GB/T 45958—2025，3.1，有修改]

3.4

智能系统不确定性量化 intelligent system uncertainty quantification

针对智能系统输出及其决策过程中的认知不确定性与数据不确定性，进行建模、估计与校准的技术活动与结果。

3.5

智能系统故障容限 intelligent system fault tolerance

智能系统在部分部件、模型子模块或输入数据异常的情况下，通过冗余、降级或自愈机制维持基本功能与安全阈值的能力。

3.6

人工智能可靠性　artificial intelligence reliability

AI系统在规定条件与时间内实现预期功能，并在可接受风险水平下运行且不导致系统失效的能力，涵盖数据、模型、平台与人机协同等要素。

3.7

运行稳定性　operation stability

指系统在连续运行过程中，性能指标（如吞吐、时延、误差）随时间的波动保持在设计容差内的特性。

3.8

模型可靠性　model reliability

模型在规定条件与时间内输出满足功能/性能要求且不引发系统功能失效的能力，体现于泛化能力、鲁棒性、校准性与可解释性/可理解性等维度。

3.9

模型退化　model degradation

在给定任务与评估口径下，模型性能因时间、数据分布、硬件/软件环境变化而出现系统性下降的现象。

3.10

训练　training

在给定的训练数据及约束条件下，通过优化算法调整模型参数以最小化（或最大化）预定目标函数的过程。

[来源：GB/T 42018—2022，3.11，有修改]

3.11

推理　inference

在模型参数固定的情况下，模型对给定输入生成输出结果（如预测、分类、决策或评分）的过程。

注：推理通常在部署与运行阶段进行，与训练阶段的参数更新相区分，可包含不确定性估计与置信度输出。

[来源：GB/T 42018—2022，3.12，有修改]

3.12

训练集　training set

用于估计和更新模型参数的数据集。

注：训练集应与验证集、测试集在样本层面互不重叠；对时间序列或主体相关数据，应采取时间/主体隔离以防信息泄漏。

[来源：GB/T 41867—2022，3.2.34，有修改]

3.13

测试集　test set

用于在模型训练与选择完成后，对模型在未见数据上的性能进行客观评估的独立数据集。

注：测试集仅用于最终评估，不应用于任何形式的训练或调参决策。

[来源：GB/T 41867—2022，3.2.3，有修改]

3.14

验证集　validation set

用于模型开发阶段选择模型结构、调参与早停等的独立数据集，不参与模型参数的最终训练。

注：验证集应与训练集、测试集互不重叠，避免信息泄漏。

[来源：GB/T 41867—2022，3.2.35，有修改]

3.15

数据漂移　data drift

数据分布随时间或场景变化而发生的统计性质改变，导致训练分布与运行分布不一致。

[来源：ISO/IEC 5259—1 术语框架，有修改]

3.16

泛化能力　generalization ability

模型在与训练数据分布相同或相近的未见样本上保持预期性能的能力。

注：可通过测试集性能、交叉验证、分布外评估等方式间接衡量。

3.17

鲁棒性　robustness

在扰动、噪声或分布偏移条件下维持功能/性能的能力。

[来源：GB/T 41867—2022 3.4.9 有修改]

3.18

对抗鲁棒性　adversarial robustness

智能系统在受限范数、物理可实现等对抗扰动下维持性能与安全阈值的能力。

3.19

校准　calibration

模型输出概率与实际命中频率一致性的性质。

3.20

分布外　out-of-distribution

偏离训练数据分布的输入样本或场景。

4　缩略语

下列缩略语适用于本文件：

CPU	中央处理器（Central Processing Unit）
GPU	图形处理器（Graphics Processing Unit）
AI	人工智能（Artificial Intelligence）
ASIC	专用集成电路（Application-Specific Integrated Circuit）
I/O	输入/输出（Input/Output）
API	应用程序编程接口（Application Programming Interface）
SDK	软件开发工具包（Software Development Kit）
CNN	卷积神经网络（Convolutional Neural Network）
MTBF	平均故障间隔时间（Mean Time Between Failures）
CoD	停机成本（Cost of Downtime）
POFOD	按需故障概率（Probability of Failure on Demand）
KL	Kullback–Leibler 散度（Kullback–Leibler Divergence）
KS	Kolmogorov–Smirnov 检验（Kolmogorov–Smirnov Test）
MFTI	模型平均无故障间隔时间（Mean Failure-free Time Interval）
MSE	均方误差（Mean Squared Error）
MAE	平均绝对误差（Mean Absolute Error）

R-squared 决定系数（Coefficient of Determination, R^2）
DB 戴维斯-布尔丁指数（Davies–Bouldin Index）
P@K 前K位精度（Precision at K）
AP 平均精度（Average Precision）
MAP 平均精度（Mean Average Precision）
R@K 前K位召回（Recall at K）
FMEA 失效模式与影响分析（Failure Modes and Effects Analysis）
FMECA 失效模式、影响与危害度分析（Failure Modes, Effects, and Criticality Analysis）
FRACAS 故障报告、分析与纠正措施系统（Failure Reporting, Analysis, and Corrective Action System）
DRACAS 缺陷报告、分析与纠正措施系统（Defect Reporting, Analysis, and Corrective Action System）
HAZOP 危险与可操作性分析（Hazard and Operability Study）
HACCP 危害分析与关键控制点（Hazard Analysis and Critical Control Points）
LOPA 分层保护分析（Layer of Protection Analysis）
CCF 共因失效（Common Cause Failure）
RBD 可靠性框图（Reliability Block Diagram）
MTTR 平均修复时间（Mean Time To Repair）
CTMC 连续时间马尔可夫链（Continuous-Time Markov Chain）
DTMC 离散时间马尔可夫链（Discrete-Time Markov Chain）
FDIR 故障检测、隔离与恢复（Fault Detection, Isolation and Recovery）
SLA 服务等级协议（Service Level Agreement）
BIA 业务影响分析（Business Impact Analysis）
RTO 容灾恢复目标时间（Recovery Time Objective）
RPO 容灾恢复点目标（Recovery Point Objective）
MC Dropout 蒙特卡罗 Dropout（Monte Carlo Dropout）
ANOVA 方差分析（Analysis of Variance）
QUIC 快速UDP互联网连接（Quick UDP Internet Connections）
LSTM 长短期记忆网络（Long Short-Term Memory）
RS 随机平滑（Randomized Smoothing）
LDPC 低密度奇偶校验码（Low-Density Parity-Check Code）
GAN 生成对抗网络（Generative Adversarial Network）
NLL 负对数似然（Negative Log-Likelihood）
EWC 弹性权重固化（Elastic Weight Consolidation）
GNN 图神经网络（Graph Neural Network）
SLO 服务等级目标（Service Level Objective）
FGSM 快速梯度符号法（Fast Gradient Sign Method）
PGD 投影梯度下降（Projected Gradient Descent）
VI 变分推断（Variational Inference）
KD 知识蒸馏（Knowledge Distillation）
ER 记忆回放（Experience Replay）
RL 强化学习（Reinforcement Learning）
FTA 故障树分析（Fault Tree Analysis）

R+FGSM	随机初始化的FGSM（Random Initialization + FGSM）
NF	标准化流/归一化流（Normalizing Flow）
BNF	巴科斯－诺尔范式（Backus–Naur Form）
PEG	概率事件图（Probability Event Graph）
VAE	变分自编码器（Variational Autoencoder）
LPIPS	学习感知图像块相似度（Learned Perceptual Image Patch Similarity）
BN	批量归一化（Batch Normalization）
LN	层归一化（Layer Normalization）
NaN	非数值（Not a Number）
LTL	线性时序逻辑（Linear Temporal Logic）
CTL	计算树逻辑（Computation Tree Logic）
CBMC	C 程序有界模型检验器（C Bounded Model Checker）
CW	Carlini–Wagner 攻击（Carlini–Wagner Attack）
APM	应用性能管理（Application Performance Management）
RCA	根本原因分析（Root Cause Analysis）

5 智能系统可靠性核心对象

5.1 数据对象

5.1.1 范围与构成

数据对象是指支撑智能系统开发、验证与运行的各类数据集合，通常包括：

a） 训练数据（training data）；

b） 验证数据（validation data）；

c） 测试数据（test data）；

d） 运行时数据（operational data）。

相关元数据、标签数据、数据生成与采集流程、数据质量记录与追溯信息数据数据对象的组成部分。

5.1.2 边界与接口

数据对象的边界包括：

a） 数据来源；

b） 采集条件；

c） 预处理与增强策略；

d） 分割与抽样策略；

e） 版本与追溯标识；

f） 访问与控制策略。

数据对象与模型对象的接口包括：

a） 特征与标签规范；

b） 数据模式与输入分布；

c） 数据质量指标与约束。

数据对象与平台对象的接口包括：

a） 数据存储与传输协议；

b） 带宽与时延约束；

c） 缓存与容错策略。

5.2 模型对象

5.2.1 范围与构成

模型对象包括但不限于：

a） 模型架构；

b） 参数与权重；

c） 超参数；

d） 神经元与层次结构；

e） 激活函数；

f） 损失函数；

g） 不确定性与校准策略；

h） 后处理策略。

与模型相关的训练流程、检查点、版本与变更记录、模型签名与完整性校验信息亦数据模型对象的组成部分。

5.2.2 边界与接口

与数据对象接口包括：

a） 输入/输出张量规范；

b） 特征工程/预处理约定；

c） 标签定义与一致性要求。

与平台对象的接口包括：

a） 计算图执行约束（内存/算力/延迟）；

b） 并行与加速策略；

c） 模型部署形态（云/边/端）；

d） 模型加载与热更新策略。

5.3 AI平台

5.3.1 硬件平台

硬件平台构成元素包括：

a） 处理器（CPU/GPU/AI加速器/ASIC）；

b） 存储/缓存；

c） 网络接口；

d） 传感与I/O；

e） 边缘设备；

f） 供电与散热等。

边界包括：

a） 资源能力与约束（算力、内存、宽带、时延、能耗、温度）；

b） 冗余与容错机制的设计接口。

5.3.2 软件平台

软件平台构成元素包括：

a） 操作系统；

b） 驱动；

c） 容器与编排；

d） 运行时与加速库；

e） 模型服务框架；

f） 中间件；

g） 监控与日志组件。

边界包括：

a） 版本与依赖；

b） 接口与兼容性；

c） 资源调度与隔离；

d） 可靠性特性（回滚、熔断、限流）对接点。

5.3.3 接口与交互

接口与交互构成元素包括：

a） 应用结构（API/SDK）；

b） 数据结构（消息/流/文件）；

c） 人机交互接口；

d） 安全与访问控制。

边界包括：

a） 协议；

b） 吞吐与时延指标；

c） 错误与异常处理；

d） 接口契约与兼容性。

6 智能系统可靠性定性要求

6.1 数据可靠性定性要求

6.1.1 训练数据覆盖性

明确训练数据对环境干扰、异常与组合场景的覆盖，以提升模型对异常输入的容错性与输出稳定性。具体要求包括：

a） 应包含可能影响系统可靠性的环境干扰样本及其组合干扰样本；

b） 应包含噪声与异常样本，应在数据版本记录中表示覆盖口径与比例；

c） 宜对边界/极端/异常样本设定最低覆盖比例并形成记录。

6.1.2 测试数据极端与边界场景

验证系统在极端、复杂与边界工况下的可靠性表现。具体要求包括：

a） 应覆盖极端、复杂与边界工况；

b） 应包含多因素组合干扰与异常场景样本；

c） 宜建立场景库与覆盖率度量，并输出测试报告与缺陷闭环记录。

6.1.3 运行时数据质量

应对输入数据中的环境干扰与异常进行识别、过滤与校正；应配置在线数据质量监测与告警；应对输出进行可靠性评估与误差分析。在运行期持续保障输入数据质量与输出可信性，具体要求包括：

a） 应对输入数据中的环境干扰与异常进行识别、过滤与校正；

b） 应配置在线数据质量监测与告警，并对输出进行可靠性评估与误差分析；

c） 宜设定质量阈值与处置策略（如隔离、降级、人工复核）。

6.2 模型可靠性定性要求

6.2.1 模型验证与测试

通过离线与上线前验证降低不确定性，确保稳定性和容错能力。具体要求包括：

a） 应采用K折交叉验证或等效方法，验证在不同数据子集上的稳定性与方差；

b） 应开展压力与极限输入测试（噪声、缺失、异常样本）以检验容错能力；

c） 宜在上线阶段采用分组对照测试或影子流量比对，验证新旧模型的可靠性差异。

6.2.2 不确定性量化与处置

量化预测不确定性并进行校准，支持风险可控的决策。具体要求包括：

a） 应输出置信度或不确定性度量，并进行校准（如温度缩放、预期校准误差控制）；

b） 应对低置信度预测设定处置策略（自动复核、人工干预或保守决策），并记录闭环；

c） 宜定期复核校准性能与阈值有效性。

6.2.3 退化监测与预警

及时发现并处置性能退化，防止服务质量下降。具体要求包括：

a） 应建立退化指标与监测机制（如准确率波动、响应时间增长、预期校准误差上升、漂移度量超阈）；

b） 应设置告警阈值与自动化处置策略（再训练、回滚、限流/降级），并记录结果；

c） 宜开展根因分析与回归验证，确保持续改进。

6.2.4 版本控制与回滚

确保模型可追溯与快速恢复，降低上线风险。具体要求包括：

a） 应严格记录模型版本、训练数据版本与超参数配置；

b） 部署时应保留至少一个稳定版本，支持故障或退化时快速回退；

c） 宜采用灰度发布与受控放量策略。

6.2.5 环境适应性与资源约束

在资源与场景变化下维持可靠性与一致性。具体要求包括：

a） 应对分布变化具备适应策略（再训练、迁移学习或阈值调优），并设定触发条件；

b） 应针对不同运行环境提供模型优化方案（量化、裁剪、蒸馏），并验证不低于接受阈值；

c） 宜进行云/边/端一致性验证。

6.2.6 冗余与多样性

通过模型冗余与多样性提升关键任务场景的可靠性。具体要求包括：

a） 关键任务场景应采用冗余或异构多样性策略（如CNN与Transformer组合），并明确融合机制（投票、加权等）；

b） 应评估冗余对可用度与延迟/资源的影响，确保系统级目标达成。

6.3 平台可靠性定性要求

6.3.1 系统容错能力

提升系统在故障与异常条件下的持续服务能力。具体要求包括：

a） 应建立异常检测与实时监控；

b） 应设计自动恢复流程以处理非致命错误；

c） 宜采用心跳检测与故障注入演练，验证容错有效性。

6.3.2 资源管理

防止资源枯竭与资源竞争引发的不可用。具体要求包括：

a） 应对CPU、内存、磁盘与网络设定阈值与配额/隔离；

b） 应实现动态资源分配与调度；

c） 应建立资源回收机制；

d） 宜开展容量规划与长期趋势分析。

6.3.3 性能稳定性

保证持续运行过程中性能在设计容差内。具体要求包括：

a） 应确保性能波动不超过设计容差；

b） 应进行持续压力测试；

c） 宜建立性能基线并定期比对；

d） 宜监测温度/能耗等物理指标以预防热退化。

6.3.4 依赖管理

降低第三方依赖导致的级联故障风险。具体要求包括：

a） 应对第三方组件与外部服务实施版本与安全管控，明确兼容矩阵；

b） 应制定并实现降级策略，保障依赖不可用时核心功能可用；

c） 应实现熔断与限流机制；

d） 宜配置离线缓存/本地应急能力。

6.3.5 日志与监控

保障可追溯性与异常处置的及时性。具体要求包括：

a） 应建立日志记录与追溯机制，记录关键操作与异常；

b） 应实现实时监控与告警，并定期复盘与阈值再校准。

6.3.6 更新与维护

在变更中维持高可用与可恢复性。具体要求包括：

a） 应支持受控的热更新/滚动更新；

b） 应制定并验证回滚方案；

c） 应建立定期维护与体检机制；

d） 宜在维护窗口执行合规与安全基线审计。

6.3.7 环境适配性

适配异构环境并缓解网络不确定性影响。具体要求包括：

a） 应完成不同操作系统、硬件与驱动/库版本的兼容性验证；

b） 应评估网络波动、抖动与丢包的影响并制定缓解策略；

c） 宜提供环境自检工具，输出自检项与阈值清单，并纳入发布前检查。

7 智能系统可靠性定量要求

7.1 运行可靠性指标体系

7.1.1 平均故障间隔时间（MTBF）

用于度量修复型系统在统一故障判据下，相邻两次故障之间的平均时间间隔，计算方法见公式（1）。

$$MTBF = \frac{a_{MTBF}}{b_{MTBF}} \times 100\% \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（1）$$

式中：

MTBF——平均故障间隔时间，单位为小时（h）；

a_{MTBF}——预测期有效运行总时间（剔除计划停机），单位为小时（h）；

b_{MTBF}——预测期内按统一口径统计的故障次数。

7.1.2 停机成本（CoD）

用于量化系统不可用期间造成的直接与间接经济损失，计算方法见公式（2）。

$$CoD = a_{CoD} \times b_{CoD} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（2）$$

式中：

CoD——停机成本，单位为CNY；

a_{CoD}——停机时间，单位为小时（h）；

b_{CoD}——每小时的成本损失，单位为CNY。

7.1.3 软件按需故障概率（POFOD）

用于衡量软件在一次“按需请求”触发时发生故障的概率，适用于事务性接口与批作业场景，计算方法见公式（3）。

$$POFOD = \frac{a_{POFOD}}{b_{POFOD}} \times 100\% \quad \cdots\cdots (3)$$

式中：

$POFOD$——按需故障概率；

a_{POFOD}——发生故障的按需请求次数，单位为次；

b_{POFOD}——按需请求总次数，单位为次。

7.2 数据可靠性指标体系

7.2.1 覆盖完整性

用于度量数据集对操作参数空间的覆盖程度；采用分箱法将空间离散为超立方体集合时，计算方法见公式（4）。

$$Cov_{\varkappa} = \frac{a_{Cov}}{b_{Cov}} \times 100\% \quad \cdots\cdots (4)$$

式中：

$Cov_{\varkappa}$——覆盖完整性；

a_{Cov}——包含数据点的超立方体数量，单位为个；

b_{Cov}——超立方体总数，单位为个。

7.2.2 数据代表性

7.2.2.1 样本分布一致性

7.2.2.1.1 KL散度

用于衡量样本分布Q相对于目标分布P的信息散度，计算方法见公式（5）。

$$D_{KL}(P||Q) = \sum_i P(x_i) log \frac{P(x_i)}{Q(x_i)} \quad \cdots\cdots (5)$$

式中：

$D_{KL}(P||Q)$——KL散度；

$P(x_i)$、$Q(x_i)$——在x_i处的概率或概率密度。

7.2.2.1.2 KS检验统计量

用于比较经验分布函数与目标分布函数的最大差异，计算方法见公式（6）。

$$D_n = sup_x|F_n(x) - F(x)| \quad \cdots\cdots (6)$$

式中：

D_n——KS统计量；

$F_n(x)$——样本经验分布函数；

$F(x)$——目标（理论）分布的CDF。

7.2.2.2 类别分布一致性

用于检验样本类别分布与目标（或理想）分布的一致性，计算方法见公式（7）。

$$\chi^2 = \sum_c (O_c - E_c)^2 / E_c \quad \cdots\cdots (7)$$

式中：

χ^2——卡方统计量；

O_c——类别c的观测频数；

E_c——类别c的期望频数。

7.2.3 数据准确性与方差

用于度量数据与真实值或设计真值的一致程度，计算方法见公式（8）。

$$\sigma^2 = 1/\mathrm{n}\sum_{i=1}^{n}(x_i-\mu)^2 \quad \cdots\cdots (8)$$

式中：

σ^2——方差，单位为被测量单位的平方；

μ——数据的总体均值；

x_i——数据集中第i个样本的值；

n——数据总数。

7.2.4 数据可追溯性

用于度量数据从来源到使用的可跟踪与可复现程度，计算方法见公式（9）。

$$T_R = N_{tr}/N_{to} \quad \cdots\cdots (9)$$

式中：

T_R——可追溯记录完备率；

N_{tr}——具备完整版本/处理记录的数据项数；

N_{to}——评估范围内的数据项总数。

注：不同敏感级别可设置差异化阈值。

7.2.5 数据独立性

用于验证训练集、验证集与测试集在样本层面互不重叠并在抽样上独立，计算方法见公式（10）。

$$\rho(A,B) = |A \cap B|/\min\{|A|,|B|\} \quad \cdots\cdots (10)$$

式中：

$\rho(A,B)$——数据集A,B的重叠度；

$|\cdot|$——集合基数。

7.3 模型可靠性指标体系

7.3.1 可靠度

7.3.1.1 任务成功比率

用于度量系统（或模型）在时间区间$[0,t]$内无故障运行的概率，计算方法见公式（11）。

$$R(t) = \frac{a_{Rt}}{b_{Rt}} \times 100\% \quad \cdots\cdots (11)$$

式中：

$R(t)$——任务成功比率；

a_{Rt}——任务成功次数；

b_{Rt}——总任务次数。

7.3.1.2 模型平均无故障间隔时间（MFTI）

用于度量相邻两次“由模型算法引发的故障”之间的平均工作时间，计算方法见公式（12）。

$$MFTI = \frac{a_{MFTI}}{b_{MFTI}} \times 100\% \quad \cdots\cdots (12)$$

式中：

$MFTI$——模型平均无故障间隔时间，单位为小时（h）；

a_{MFTI}——模型层面的有效运行时间，单位为小时（h）；

b_{MFTI}——模型层面故障次数，单位为次。

7.3.1.3 可用度

用于估算修复型系统的稳态平均可用度，计算方法见公式（13）。

$$Avail = \frac{a_{avail}}{b_{avail}} \times 100\% \quad \cdots\cdots (13)$$

式中：

$Avail$——可用度；

a_{avail}——正常运行时间，单位为小时（h）；

b_{avail}——总时间，单位为小时（h）。

7.3.1.4 任务成功率

用于衡量特定任务场景下成功完成任务的比例，计算方式见公式（14）。

$$TSR = \frac{a_{TSR}}{b_{TSR}} \times 100\% \quad \cdots\cdots (14)$$

式中：

TSR——任务成功率；

a_{TSR}——成功任务数；

b_{TSR}——总任务数。

7.3.2 模型性能

7.3.2.1 分类模型评估

7.3.2.1.1 准确度

用于衡量整体预测正确性，计算方法见公式（15）。

$$Accuracy = \frac{a_{Acc}}{b_{Acc}} \times 100\% \quad \cdots\cdots (15)$$

式中：

$Accuracy$——准确度；

a_{Acc}——正确的预测数量；

b_{Acc}——预测总数。

7.3.2.1.2 精确度

用于衡量阳性判定的准确性，计算方法见公式（16）。

$$Precision = \frac{a_{Pre}}{b_{Pre}} \times 100\% \quad \cdots\cdots (16)$$

式中：

$Precision$——精确度；

a_{Pre}——正确预测的阳性观察值；

b_{Pre}——总预测阳性观察值。

7.3.2.1.3 召回率

用于衡量阳性判定的覆盖性，计算方法见公式（17）。

$$Recall = \frac{a_{Recall}}{b_{Recall}} \times 100\% \quad \cdots\cdots (17)$$

式中：

$Recall$——召回率；

a_{Recall}——正确预测的阳性观察值；

b_{Recall}——实际阳性观察值。

7.3.2.1.4 F1分数

用于在精确度与召回率之间取得平衡，适合正负样本代价相近的分类场景，计算方法见公式（18）。

$$F1 = 2 \times Precision \times Recall/(Precision + Recall) \quad \cdots\cdots (18)$$

式中：

$F1$——F1分数；

Precision——精确度，按式（16）计算；

Recall——召回率，按式（17）计算。

7.3.2.2 回归模型评估

7.3.2.2.1 均方误差（MSE）

用于度量预测误差的平方平均值，计算方法见公式（19）。

$$MSE = (1/n) \times \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \cdots\cdots (19)$$

式中：

MSE——均方误差，单位为目标量单位的平方；

y_i——第i个真实值；

$\hat{y}_i$——第i个预测值；

n——样本数。

7.3.2.2.2 平均绝对误差（MAE）

用于度量平均绝对偏差，计算方法见公式（20）。

$$MAE = (1/n) \times \sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \cdots\cdots (20)$$

式中：

MAE——平均绝对误差，单位为目标量单位；

y_i——第i个真实值；

$\hat{y}_i$——第i个预测值；

n——样本数。

7.3.2.2.3 决定系数（R-squared）

用于衡量模型对数据变异性的解释比例，计算方法见公式（21）。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \cdots\cdots (21)$$

式中：

R^2——决定系数；

y_i——第i个真实值；

$\hat{y}_i$——第i个预测值；

$\bar{y}$——真实值的样本均值；

n——样本数。

7.3.2.3 聚类模型评估

7.3.2.3.1 DB指数

通过“簇内紧密度/簇间分离度”的相对关系评估聚类质量，计算方法见公式（22）。

$$S_i = (1/|C_i|)\sum_{x \in C_i} dist(x, \mu_i),\ R_{ij} = \frac{S_i + S_j}{d(\mu_i, \mu_j)},\ DB = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i} R_{ij} \quad \cdots\cdots (22)$$

式中：

C_i——第i个簇，$|C_i|$其样本数；

μ_i——簇i的中心；

$dist(\cdot,\cdot)$——距离函数；

K——簇数；

DB——DB指数（越小越好）。

7.3.2.4 排序模型评估

7.3.2.4.1 精准度类指标

用于衡量前K名结果的相关性质量。

a） Precision at K（P@K）的计算方法见公式（23）。

$$P@K = \frac{a_{P@K}}{K} \times 100\% \quad \cdots\cdots (23)$$

式中：

$a_{P@K}$——前K个项目中的相关项目数；

K——项目总数。

b） Average Precision（AP）的计算方法见公式（24）。

$$AP = \frac{1}{|R|}\sum_{k=1}^{N} P(k) \cdot rel(k) \quad \cdots\cdots (24)$$

式中：

$|R|$——样本中真实的相关项总数；

N——检索或排序列表的长度；

$rel(k)$——指示函数，当排序列表中第k个项目是相关项时取1，否则取0；

$P(k)$——截至位置k的截断精确率。

c） Mean Average Precision（MAP）的计算方法见公式（25）。

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP_q \quad \cdots\cdots (25)$$

式中：

Q——查询的数量；

AP_q——第q个查询中相关项的排名位置。

7.3.2.4.2 召回率类指标

用于衡量在前K名中覆盖到多少相关项。

a） Recall at K（R@K）的计算方法见公式（26）。

$$R@K = \frac{a_{R@K}}{K} \times 100\% \quad \cdots\cdots (26)$$

式中：

$a_{R@K}$——前K个项目中的正确的相关项目数；

K——项目总数。

b） Mean Reciprocal Rank（MRR）的计算方法见公式（27）。

$$MRR = \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{rank_q} \quad \cdots\cdots (27)$$

式中：

Q——查询的数量；

$rank_q$——第q个查询中相关项的排名位置。

7.3.2.4.3 排序相关性类指标

用于考虑相关性的等级与位置折扣的综合指标，计算方法见公式（28）。

$$IDCG_k = \sum_{i=1}^{k} \frac{rel(i)}{\log(i+1)}，\ NDCG_k = \frac{DCG_k}{IDCG_k} \quad \cdots\cdots (28)$$

式中：

DCG_k——实际的累计增益；

$IDCG_k$——理想的累计增益。

7.3.3 泛化能力

7.3.3.1 验证集准确度

用于衡量模型在训练外数据上的正确性，用于监测过拟合，计算方法见公式（29）。

$$Accuracy_{val} = \frac{a_{Accuracy}}{b_{Accuracy}} \times 100\% \quad \cdots\cdots (29)$$

式中：

$Accuracy_{val}$——验证集准确度；

$a_{Accuracy}$——正确预测的样本数；

$b_{Accuracy}$——验证集总样本数。

7.3.3.2 平均交叉验证准确率

用于降低单次数据划分的偶然性，评估模型平均泛化性能，计算方法见公式（30）。

$$\overline{Acc} = \frac{1}{K}\sum_{i=1}^{K} Accuracy_i \quad \cdots\cdots (30)$$

式中：

$\overline{Acc}$——K折交叉验证的平均准确度；

$Accuracy_i$——第i折验证集的准确度；

K——折数。

7.3.3.3 训练损失

训练损失反映模型在训练数据上的拟合程度，计算方法见公式（31）。

$$E_{train} = \frac{1}{N_{train}}\sum_{i=1}^{N_{train}} L(y_i, \hat{y}_i) \quad \cdots\cdots (31)$$

式中：

E_{train}——训练集上的平均损失；

N_{train}——训练集样本量；

y_i——第i个样本的真实标注；

$\hat{y}_i$——模型对第i个样本的预测输出；

$L(y_i, \hat{y}_i)$——样本损失函数（对预测与真实值的偏差进行度量，如交叉熵、均方误差等）。

7.3.3.4 验证损失

验证损失反映模型在未参与训练的数据上的泛化能力，计算方法见公式（32）。

$$E_{val} = \frac{1}{N_{val}}\sum_{i=1}^{N_{val}} L(y_i, \hat{y}_i) \quad \cdots\cdots (32)$$

式中：

E_{val}——训练集上的平均损失；

N_{val}——训练集样本量；

y_i——第i个样本的真实标注；

$\hat{y}_i$——模型对第i个样本的预测输出；

$L(y_i, \hat{y}_i)$——样本损失函数（对预测与真实值的偏差进行度量，如交叉熵、均方误差等）。

7.3.4 鲁棒性

7.3.4.1 标准鲁棒准确率

在指定扰动条件下，模型预测正确的比例，计算方法见公式（33）。

$$RobustAcc = (1/M)\sum_{i=1}^{M} \mathbf{1}\{\hat{y}_i^{(*)} = y_i\} \quad \cdots\cdots (33)$$

式中：

$RobustAcc$——鲁棒准确率；

M——预测样本数；

$\hat{y}_i^{(*)}$——样本i在某一固定扰动条件（某种损坏或某种攻击）下的预测；

$\hat{y}_i$——真实标签；

$\mathbf{1}(\cdot)$——指示函数，成立了为1，否则为0。

7.3.4.2 平均鲁邦准确率

用于衡量多个扰动条件的总体表现，计算方法见公式（34）。

$$\overline{RobustAcc} = (1/J)\sum_{j=1}^{J} RobustAcc_j \quad \cdots\cdots (34)$$

式中：

$\overline{RobustAcc}$——平均鲁棒准确率；

J——扰动种类数；

$RobustAcc_j$——第j种扰动下的鲁棒准确率。

7.3.4.3 全部通过率

同一样本在“所有”扰动下都预测正确的比例，计算方法见公式（35）。

$$PassAll = (1/M)\sum_{i=1}^{M}\prod_{j=1}^{J}\mathbf{1}\{\hat{y}_i^{(*)} = y_i\} \quad \cdots\cdots (35)$$

式中：

$PassAll$——全部通过率；

M——测试样本总数；

$\hat{y}_i^{(*)}$——样本i在第j种扰动下的预测；

y_i——样本i的真实标签；

$\mathbf{1}\{\cdot\}$——指示函数，条件为真取1，否则取0。

7.3.4.4 稳定性降幅

衡量扰动带来的性能下降程度，计算方法见公式（36）。

$$Drop = Accuracy_{clean} - \overline{RobustAcc} \quad \cdots\cdots (36)$$

式中：

$Accuracy_{clean}$——无扰动（干净数据）上的准确率。

8 智能系统可靠性分析技术

8.1 故障与风险识别

可采用FMEA/FMECA、FRACAS/DRACAS、场景树/事件树、HAZOP、HACCP、风险矩阵、LOPA、共因失效识别（CCF）等方法枚举失效、顶事件与过程偏移，并结合智能系统的“数-模-人-环”要素生成风险登记与共因路径。包括：

a） 枚举失效模式与影响，形成失效模式库并标注原因/后果；

b） 定义顶事件与触发条件，编制场景清单与事件链；

c） 识别过程偏离与关键控制点，建立风险登记册与接受阈值；

d） 识别共因失效与耦合路径，绘制依赖图谱。

8.2 结构与路径建模

可采用RBD、串并联系统模型、重要度分析等方法构建系统串并联与冗余模型并计算可用度/可靠性与瓶颈；对智能系统，需纳入“数据与模型链路”的结构等价单元和反馈环。包括：

a） 建立系统结构模型，表示串并联与冗余切换关系；

b） 录入单元可靠性参数（MTBF/MTTR），设定计划停机与切换时间口径；

c） 计算系统可靠度/可用度，识别瓶颈单元与重要度排序；

d） 对感知-决策-执行闭环以环路延迟、抖动、丢包率为参数，判定稳定性边界与安全降级区。

8.3 因果概率与时序评估

可采用FTA、最小割集、马尔可夫链（CTMC/DTMC）、半马尔可夫、贝叶斯网络、β因子模型、Petri网等方法求解顶事件概率与割集并评估切换维护策略的时序指标；对智能系统，需显式引入“模型输出不确定性”和“数据-模型依赖”的概率项。包括：

a） 构建顶事件的因果模型，求解最小割集与顶事件概率；

b） 为切换/维修/降级策略建立时序模型，计算稳态可用度与期望停机时长；

c） 对共因与依赖进行联合概率建模，生成情景推演结果。

8.4 检测-隔离-恢复-验证

可采用故障注入测试、FDIR评估、混沌工程等方法执行故障注入与混沌实验并测量检测/隔离/恢复性能；增加“模型退化与漂移检测”与“再训练触发”的验证闭环。包括：

a） 设计并执行故障注入与混沌实验，覆盖软件/硬件/网络/数据异常；
b） 记录检测率、虚警率、隔离时延、恢复时延与SLA违约事件；
c） 比对模型预测与实测结果，给出差异与修正参数；
d） 综合近窗口内的错误率、置信度校准误差、漂移统计量（如PSI/KL）、延时分位，作为再训练/回滚触发量；
e） 为人机协同建立“确认链路可靠性测试”，验证双人复核、撤销窗口与审计完整性。

8.5 业务与性能影响评估

可采用BIA、RTO/RPO建模、离散事件仿真打了个方法量化停机与稳健性影响并确定阈值与降级策略。包括：

a） 量化停机与数据丢失成本，设定并校验RTO/RPO；制定降级/切流/人工接管策略；
b） 构建数据损坏基准，统一扰动预算并批量评测模型稳健性；
c） 基于代价矩阵选择阈值，给出收益-风险权衡。

8.6 不确定性量化与校准

可采用偏差-方差分解、MC Dropout、深度集成、Sobol方差分解、ANOVA、Morris等方法拆分不确定性来源并生成校准/覆盖等置信度指标；加入漂移检测与对抗鲁棒性评测，并将阈值与运维动作绑定。包括：

a） 区分认知/偶然不确定性，进行方差与敏感性分解，定位主要贡献因子；
b） 估计预测后验或近似后验，评估并校准概率输出；
c） 生成回归或计数任务的预测区间与覆盖率，设定不确定性驱动的复核/降级阈值；
d） 采用PSI、KL、MMD等统计量对输入/特征/输出分布进行监测，定义告警等级与响应策略；
e） 当ECE超过阈值或覆盖率低于目标时，自动提高阈值/切换保守策略/触发再训练；
f） 记录每版模型的不确定性画像与校准曲线，确保回滚与比对的一致性。

9 智能系统可靠性设计技术

9.1 数据采集可靠性设计

可采用多模态融合、联邦一致性校验、孤立森林/深度异常检测、区块链存证/审计链等方法保证采集稳定、可互证、可追溯、可重校准。包括：

a） 部署具备本地异常过滤与特征提取的智能传感器；在关键点配置“异构阵列+多模态融合”互证；
b） 启用跨设备一致性校验（阈值触发隔离）；
c） 上线滑动窗口漂移检测与自动重校准；
d） 建设多传感器时间同步与采集事件溯源链。

9.2 数据传输可靠性设计

可采用智能分片/优先级、MPTCP/QUIC、传输健康度模型、同态加密、差分隐私、零知识校验等方法保障低时延、可预警、可切换且隐私可验证的链路。包括：

a） 启用智能分片与优先级队列，可采用强化学习链路选择优化时延和丢包；
b） 部署健康度预测与预警，联动动态路由/多路径切换；
c） 对隐私数据启用加密与密文域完整性校验。

9.3 数据存储可靠性设计

可采用LSTM热度预测、纠删码RS/LDPC、Gorilla压缩、端到端哈希/签名等方法让数据恢复。包括：

a） 按访问热度做智能分层与资源匹配；

b） 训练样本采用“多副本+纠删码”，推理结果启用时序压缩且元数据独立冗余；

c） 全链路哈希/签名校验与周期性完整性扫描，失败即隔离并自动恢复。

9.4 数据处理可靠性设计

可采用对抗样本检测、GAN对抗库、数字孪生、低置信触发等方法在在进入训练/推理前拦截错误数据并闭环验证。包括：

a） 训练前运行对抗样本检测与清洗；构建攻击样本库用于鲁棒训练；

b） 在数字孪生中模拟异常数据与设备失效，低置信度触发人工复核与回退；

c） 根据训练反馈动态调整增强策略与强度，保持分布一致性。

9.5 N版本设计

可采用多样化集成、投票/加权融合、主备/热备切换等方法提高模型抗故障与稳定性。包括：

a） 组建多个异构模型版本；

b） 配置投票/加权融合与主备切换；

c） 开展蓝绿/灰度演练。

9.6 正则化设计

可采用L1/L2、Dropout、标签平滑、Mixup/CutMix、早停、权重衰减等方法控制过拟合并提升鲁棒泛化。包括在训练流程启用多种正则并记录影响。

9.7 损失函数设计

可采用交叉熵、大间隔损失、NLL、EWC、知识蒸馏等方法依据任务与可靠性诉求构造损失函数。包括：

a） 选择基础任务损失；

b） 按需加入鲁棒性、不确定性、持续学习、可解释性损失并定权重。

9.8 内部架构设计

可采用模块化与接口契约、参数范围/冗余校验、多模态对齐、健康度评分等方法通过模块化、双通路与自适应计算构建可升级且可容错的内部架构。包括：

a） 划分输入/特征/核心/输出模块并定义接口契约；支持单模块回滚；

b） 构建主路径+辅路径（轻量模型/规则）并动态融合；

c） 按输入复杂度自适应调整深度/宽度/分辨率；关键层加入参数校验与修复；扩展多模态表征与对齐；

d） 部署模型健康度监测与异常告警，联动降级/切备。

9.9 冗余容错设计

可采用CNN/Transformer/GNN、KL正则、梯度正交化、分布式推理等方法利用差异化集成、促进多样性与并行推理，降低相关失效与尾时延。包括：

a） 选择多类不同架构并差异化初始化；

b） 引入多样性促进损失；

c） 部署并行推理、异步管道与缓存。

9.10 动态适应设计

可采用漂移检测、门控切换、在线微调、场景识别、弹性伸缩HPA等方法使智能系统可随数据与环境变化自动调整模型与资源，保持SLO稳定。包括：

a） 监测输入漂移并触发在线微调/特征重映射/门控切换；

b） 识别场景变量并动态配置推理参数；配合调度系统实现计算弹性；

c） 对新数据执行增量学习并控制遗忘。

10 智能系统可靠性训练技术

10.1 对抗训练

在存在对抗风险的任务中，可采用FGSM、PGD、AutoAttack、对抗混合训练、温度缩放、知识蒸馏等方法在混合对抗样本训练中提升鲁棒性。包括：

a） 明确攻击口径与扰动预算（范数、步数、步长），固定随机种子与日志口径；

b） 生成对抗样本并与干净样本按比例混合训练，记录每轮对抗步数；

c） 训练中动态调整攻击强度与步数，在验证集同时评估鲁棒性能；

d） 以多攻击均值作为最终验收口径，必要时使用温度缩放/蒸馏平衡鲁棒性与校准。

10.2 噪声注入训练

在传感器或标注噪声显著任务中，采用高斯噪声、标签噪声模拟、对称交叉熵、梯度裁剪、渐进式增强等方法通过位置/强度可控的噪声注入，增强抗噪能力与训练稳定性。包括：

a） 确定噪声注入位置（输入/中间特征/标签）与强度区间，固定随机种子；

b） 将含噪样本与干净样本混合训练，分布与比例可分阶段提升（渐进式）；

c） 在“无噪/有噪”验证集分别评估，持续监控损失收敛与梯度震荡；必要时采用稳健损失与梯度裁剪。

10.3 数据增强策略

在数据不足或环境敏感任务中，可采用AutoAugment、RandAugment、CutMix/Mixup、GAN生成等方法系统化增强并约束分布与语义一致性，提升泛化并量化收益。包括：

a） 建立“任务-增强强度”映射与强度上限，固化随机策略；

b） 应用领域增强（几何、光照、遮挡等）与时序增强（抖动、缩放）；

c） 通过A/B测试评估增强收益与分布一致性。

10.4 不确定性感知训练

可采用MC Dropout、深度集成（Ensemble）、变分推断（VI）、温度缩放、共形预测等方法为安全关键或分布波动任务提供可信不确定性输出，完成校准与阈值规则落地。包括：

a） 选择概率建模路线：MC Dropout、深度集成或变分推断（VI）等；

b） 若选择VI，采用重参数化与KL权重调度；必要时采用两阶段流程（标准训练→校准/概率微调）；

c） 评估NLL、ECE、区间覆盖率等指标，输出认知/偶然不确定性分解；

d） 制定阈值与告警/人审规则；

e） 使用温度缩放或深度集成提升校准稳定性，并复核跨场景/漂移下的稳健性。

10.5 增量学习

在数据持续到达或类别扩展的长期系统中，可采用EWC、知识蒸馏（KD）、记忆回放（ER）、原型记忆、特征对齐损失等方法持续学习新知识同时保持历史任务性能。包括：

a） 选择至少两类增量方法组合，定义内存预算与样本替换策略；

b） 构建代表性样本缓冲，设定新旧数据采样比，建立基线模型与验收指标；

c） 执行回归测试并记录漂移检测窗口与触发条件。

10.6 自监督学习

在标注稀缺或跨域自适应场景下，可采用SimCLR、BYOL、MAE、自动增强搜索（RL/遗传）、线性探测、少样本微调等方法通过自监督预训练获取鲁棒可迁移表征。包括：

a） 设计领域自监督任务，构建对比学习正负样本对；

b） 可进行自动增强策略搜索，并设置安全约束；

c） 评估表征质量，记录增广口径、优化器与学习率计划，并与质量控制对齐。

11 智能系统可靠性测试技术

11.1 边界值测试

可采用边界值分析、等价类划分、数字孪生/仿真回放等方法覆盖输入域上下限与越界行为，验证防御性策略与安全裕量。包括：

a） 列举关键输入维度，给出工程上下限与安全缓冲区；
b） 生成样本：下限、上限、紧贴下/上限、越界小幅、越界大幅，并设计多变量同时逼近；
c） 在代表性场景组合运行并记录异常，输出最小覆盖表；
d） 定义越界输入的防御行为（拒绝/降级/回退），验证触发准确性与安全性；绘制安全余度与性能降幅曲线；
e） 用仿真/数字孪生扩展长尾极端边界。

11.2 对抗样本测试

可采用FGSM、PGD、AutoAttack、场景因子叠加等方法在已知威胁模型下量化鲁棒性能与业务风险。包括：

a） 明确威胁模型、范数口径与预算、步数；
b） 在验证集上用标准攻击生成场景化对抗样本（叠加亮度/视角/遮挡等）；
c） 评估鲁棒准确率、误检/漏检变化，并输出相对降幅与多攻击均值；
d） 关联业务损失，生成“鲁棒性能—风险成本”对照表。

11.3 长尾分布测试

可采用重采样/重加权、等权类评测、应力测试等方法确保罕见高风险场景下的可控性能与处置预案。包括：

a） 识别长尾事件类型与频率（历史/仿真/专家标注）；
b） 抽取/合成样本集，同时报告总体性能与长尾子集性能，计算差异；对高风险类单独阈值化；
c） 使用重加权/重采样形成“等权类”评测；
d） 对多异常并发做应力测试并形成预案。

11.4 基于失效模式的测试

可采用FMEA/FMECA、故障注入、FTA/最小割集、共因失效（CCF）测试等方法以失效模式为驱动构建可复现用例，量化检出能力并覆盖顶事件路径。包括：

a） 基于 FMEA 列出高 RPN 失效模式，形成测试需求列表；
b） 将失效模式映射为测试条件/数据/故障注入；对“难以触发”的模式采用注入（通信中断、传感器漂移）；
c） 依据 FTA 顶事件路径生成最小割集场景并执行测试；
d） 对每个高风险模式提供至少一个可复现用例与判据；记录检出率与误报率；对共因路径做相关性测试。

11.5 置信度校准测试

可采用温度缩放、保序回归、深度集成、共形预测等方法生成可信的概率输出与阈值/告警规则，满足部署口径与合规。包括：

a） 选择校准度量（ECE、Brier、NLL等）与分箱策略，保证数据与部署环境一致；
b） 评估原始模型、执行温度缩放/后验近似、复评；对子群分层评估，防止信息泄漏；
c） 形成阈值与告警规则，上线后持续监控与漂移联动；提供可视化与置信区间。

11.6 分布偏移测试

在上线分布变化时，可采用预测熵监控、门控切换/降级等方法及时发现并触发自适应/降级。包括：

a） 选择偏移检测方法并设定滑动窗口与告警阈值；离线回放调参，得到灵敏度/特异度；
b） 同步监测输入特征/嵌入与输出侧（置信度/熵）漂移；

c） 联动动态适应/降级策略，做闭环验证；在仿真中注入控制量量化触发边界。

11.7 快速梯度下降法测试

可采用FGSM、R+FGSM等方法快速评估一阶敏感性与基础鲁棒下限。包括：

a） 指定范数与扰动预算；计算单步梯度生成对抗样本；

b） 评估鲁棒准确率并可视化扰动；与随机启动单步法对照评估绕过风险。

11.8 迁移攻击测试

可采用替代模型集成、多架构对照、查询预算控制等方法在无梯度接口场景评估迁移性与查询效率。包括：

a） 训练替代模型（任务/分布一致），白盒生成对抗样本；

b） 迁移攻击目标系统，统计迁移成功率；

c） 控制查询预算并统计单位查询成功率。

11.9 边界攻击测试

可采用Boundary Attack、自适应步长、正交投影等方法仅用标签反馈逐步逼近最小扰动解。包括：

a） 从误分类起点出发，在决策边界附近随机游走并控制步长；

b） 使用拒绝采样保持误分类；

c） 记录查询次数与步长；

d） 自适应步长与正交投影提高效率。

11.10 对抗补丁测试

可采用物理可实现补丁、可微渲染、跨介质复现等方法验证可打印/可投影补丁对视觉系统的真实威胁。包括：

a） 定义物理约束；

b） 在渲染环路加入几何/光照变换；

c） 优化误分类/置信度目标；

d） 报告跨视角/光照/距离的成功率；

e） 现场重放验证跨介质一致性；

f） 设计安全阈值与告警。

11.11 传感器欺骗测试

可采用多模态融合抗性评估等方法评估多传感器在开放环境下的抗欺骗能力与安全边界。包括：

a） 在仿真或封闭场地测试，遵守安全/合规；

b） 设计注入/照射参数（脉宽、频率、强度、角度）；

c） 评估误检/漏检与定位偏差；

d） 记录攻击参数与防护响应；

e） 评估多模态融合抗欺骗能力。

11.12 数据变异测试

可采用位/字节变异、边界替换、格式错配、异常检测联动等方法用位/字节/边界/格式错配等变异考察鲁棒性与防护拦截。包括：

a） 设计变异算子：位级、字节级、边界值替换、格式错配；

b） 设置概率与强度批量生成输入；

c） 执行并记录触发事件与输出差异；

d） 安全关键字段使用白/灰名单控制破坏范围；

e） 统计拦截与漏检。

11.13 数据生成测试

面向有语法/协议/格式的输入，可采用NF/PEG、协议状态机、种子库回流等方法生成合规并且受控语义破坏样本。包括：

a） 用形式化语法（BNF/PEG）生成合规输入；
b） 引入语义层变异（键值错配、单位/时序不一致）；
c） 对协议状态机按上下文敏感调整并执行；
d） 按状态机覆盖率统计进展；
e） 失败样本回流种子库。

11.14 神经模糊测试

可采用GAN/VAE、神经元覆盖、感知损失（LPIPS等）、范数约束等方法在高维感知任务中探索“冷门”激活与边界行为。包括：

a） 采用GAN/VAE在隐空间生成并扰动，导向覆盖稀疏神经元或决策边界；
b） 使用覆盖导向（激活稀疏单元、最大化中间层差异）优化候选；并行施加感知损失与范数约束以控制伪影与语义偏移；
c） 统计触发新行为样本比例与层级激活分布变化；按目标行为（罕见类别、置信区间、边界样本）提升发现率。

11.15 层删除变异测试

可采用层删除算子、恒等映射适配、稳定性监控等方法模拟结构缺陷，评估对性能与稳定性的影响。包括：

a） 随机/按规则选择可删层（如 Dropout/BN/LN/轻量残差），保证张量形状可匹配；
b） 生成缺陷模型，运行全量测试；
c） 记录性能与训练/推理稳定性（NaN、梯度爆炸/消失、收敛失败）；
d） 避免删除关键维度变换层或用适配层；优先删除正则化/辅助分支模拟欠/过拟合失衡。

11.16 连接变异测试

可采用残差/跳连扰动、注意力头重排、梯度流分析等方法扰动残差/跳连/注意力连接，诊断关键通路脆弱度。包括：

a） 随机选择连接进行扰动；生成缺陷模型并在验证集运行；
b） 覆盖故障模式：特征传递中断、注意力头失衡/漂移；分层统计浅/深层敏感性与梯度流变化；支持单连接开关分析关键边。

11.17 权重扰动测试

可采用高斯权重噪声、剪枝模拟、通道重要性/注意力分布分析等方法注入权重噪声与剪枝模拟，评估容量与响应变化。包括：

a） 注入高斯噪声；显著权重置零；
b） 明确扰动分布与尺度保持策略；分别评估偏置与归一化参数；
c） 报告容量变化估计与关键特征响应变化（注意力分布/通道重要性）。

11.18 超参数变异测试

可采用LR 突变/重启、批量大小跳变、动量/权重衰减/梯度裁剪、收敛性分类器等方法验证训练过程对关键超参数扰动的稳定性与收敛鲁棒区间。包括：

a） 对学习率进行突变试验，对批量大小进行跳变，观察统计量稳定性；
b） 全程记录损失曲线、梯度范数、更新率、学习率调度状态、收敛时间与最终性能；
c） 分类运行结果为“收敛/停滞/发散”，设定阈值口径；叠动量、权重衰减、梯度裁剪等变异，分析耦合敏感性；
d） 输出稳定区间建议（学习率范围、批量大小区间与调度策略）。

12 智能系统可靠性验证技术

12.1 模型属性验证

可采用半代数几何、MILP/线性松弛、UPPAAL、NuSMV、PRISM等方法对分类/控制与时序决策模型给出可证明的安全与稳定边界或反例。包括：

a） 明确属性类型：决策边界（连通/有界/最小间隔）、稳定性（Lipschitz上界）、时序安全（LTL/CTL的“永不/最终/直到”等路径性质）；

b） 选择工具链：半代数几何/区间约束、Lipschitz上界估计、模型检查器（UPPAAL、NuSMV）；

c） 构建可验证抽象：线性/分段线性近似、离散化状态机/抽象转移系统；

d） 执行验证，收集“可证书”或“反例路径”，并回灌设计。

12.2 代码级形式化验证

可采用Frama-C、CBMC、Infer、KLEE、TSan/ASan、CUDA-MEMCHECK等方法对对关键流程与数值/并发敏感模块进行规格化证明与反例定位。包括：

a） 明确规格：前置/后置条件、不变式、域约束（溢出、边界、并发）；

b） 执行静态分析（控制流/数据流）、符号执行（路径覆盖）、不变式证明（如Frama-C、CBMC）；

c） 对底层算子/内核插入断言与边界检查（索引界、内存一致性、精度保护），收集证明或反例。

12.3 对抗鲁棒性验证

可采用DeepPoly/CROWN-IBP、CAP/Randomized Smoothing、AutoAttack/PGD等方法在已定义威胁模型下证实局部稳健半径与全局鲁棒曲线。包括：

a） 定义威胁模型（白/灰/黑盒）、范数与预算；

b） 局部稳健性：认证测试点在邻域内无对抗样本（输出“认证半径”或反例）；

c） 全局评估：生成对抗样本集，统计鲁棒准确率与平均扰动；绘制预算—鲁棒曲线；

d） 联合业务代价函数，生成“鲁棒性—风险成本”评估。

12.4 环境扰动验证

可采用CARLA/AirSim/Gazebo、HiL、NetEm、多模态遮挡/漂移脚本等方法在仿真/硬在环注入可控环境扰动，形成退化与恢复边界。包括：

a） 列出扰动清单：传感器噪声、光照/天气、通信延迟/抖动/丢包、多模态失效；

b） 设计可控注入，按强度阶梯注入；记录SNR、延迟、丢包、失效组合等参数；

c） 采集性能退化曲线与恢复能力指标，输出“失效组合—安全余度”矩阵。

12.5 不确定性量化验证

可采用温度缩放、深度集成、MC Dropout、共形预测、分位数回归/证据深度学习等方法验证概率输出的校准性、覆盖性与不确定性分解，支撑阈值/人审。包括：

a） 设定校准集/测试集，包含分布内/外样本；

b） 计算校准指标ECE/NLL/Brier与覆盖性（区间/分位数/共形覆盖）；

c） 分解认知/偶然不确定性，验证偏移场景下认知不确定性上升；

d） 形成阈值与人工复核触发规则，并给出置信带一致性曲线。

12.6 状态覆盖验证

可采用抽象解释、状态网格化、反例引导、MC/DC、模型检查（PRISM/UPPAAL）等方法量化序列决策的状态/路径覆盖，证明关键可达性并补齐盲区。包括：

a） 定义状态划分与关键决策点，将连续状态离散为网格/抽象状态；

b） 设定覆盖准则：状态/分支/路径覆盖（可采用MC/DC类最小路径覆盖目标）；

c） 生成测试轨迹并统计覆盖比例；对未覆盖状态定向采样或生成反例；
d） 结合模型检查证明关键目标状态可达/不可达。

13 智能系统可靠性评估技术

13.1 模型可验证性评估

可采用半代数几何、区间约束传播、可达集、NuSMV、UPPAAL、PRISM等方法评估模型在安全/稳定边界与时序性质上的“可验证性”，形成满足/反例/不确定结论与证据链。包括：

a） 明确形式化属性集合：决策逻辑完备性、状态空间可达性、时序安全（LTL/CTL：G/F/U）；
b） 选择工具与抽象：半代数/区间约束、可达集计算、模型检查器（NuSMV/UPPAAL）；构建近似或离散化抽象；
c） 执行验证与证据收集，逐项输出“满足/反例/不确定”，并记录假设、边界条件与抽象误差上界。

13.2 数学性质评估

可采用Lyapunov证据、谱范数/功率迭代、IBP/CROWN/DeepPoly、可达集（Zonotope/Polytope）等方法形成稳定性/收敛性/鲁棒边界的数学结论与上界估计，并在主要工况下落地。包括：

a） 选定性质：Lyapunov稳定性、收敛性（优化/训练）、鲁棒边界（Lipschitz/可达集半径）；
b） 建立证明或上界估计流程：Lyapunov函数候选、谱范数上界、区间传播/IBP、凸/线性松弛；
c） 针对主要运行工况与负载区间，输出结论并校验紧性。

13.3 对抗鲁棒性评估

在统一威胁模型与预算下，可采用FGSM/PGD/AA/CW、随机平滑RS、迁移/查询受限攻击、对抗检测ROC、风险成本映射等方法评估鲁棒准确率、攻击成功率、平均扰动与认证半径/差距。包括：

a） 确定威胁模型（白/灰/黑盒）与预算（范数、步数、查询上限）；
b） 执行攻击基准（FGSM/PGD/AutoAttack/CW等），统一参数口径；
c） 计算鲁棒准确率、攻击成功率、平均扰动；
d） 评估对抗样本检测率与误报率；分别给出黑盒迁移与查询受限成功率；
e） 结合业务代价函数，输出“鲁棒性—风险期望”对照表。

13.4 环境扰动评估

可采用CARLA/AirSim/Gazebo、硬件在环HiL、NetEm、多模态遮挡/漂移脚本等方法量化典型环境扰动对性能与安全余度的影响与边界。包括：

a） 制定扰动档位：传感器噪声、光照/天气、通信延迟/丢包、多模态失效；
b） 仿真/实测注入并记录性能退化曲线与恢复能力；
c） 汇总可运行边界与安全余度，形成“扰动—性能—余度”热图。

13.5 实时性能评估

可采用OpenTelemetry、Prometheus/Grafana、APM/Tracing、RCA等方法评估线上服务性能与资源效率，定位瓶颈，验证SLA达标。包括：

a） 建立指标体系：延迟/吞吐/可用性/错误率/超时率；资源（CPU/GPU/显存/带宽/能耗）；异常/告警；
b） 设置信号采样与阈值，运行观察与追踪；进行根因归因与容量评估；
c） 汇总SLA达标情况，输出优化与灰度回滚评估。

13.6 自愈能力评估

可采用混沌工程（Chaos Mesh/Litmus）、健康探针等方法在故障注入下量化恢复速度、成功率与服务保持能力。包括：

a） 设计故障注入：组件崩溃、超时、资源枯竭、网络中断；
b） 执行自愈流程：重启、容灾切换、模型回滚、降级模式；
c） 统计恢复时间、成功率与重试分布，覆盖峰值与多故障并发边界。

13.7 安全防护评估

可采用对抗检测/输入过滤、模型水印/指纹、速率限制/签名校验、API网关策略等方法验证攻防面对防护策略的有效性与成本，输出整改优先级。包括：

a） 明确攻击面：对抗输入、模型逆向/窃取、数据注入、接口滥用；
b） 开展攻防演练：对抗攻击、查询窃取、模型反编译/提取、输入验证绕过；
c） 度量防护效果与开销，形整改清单与优先级。

13.8 隐私保护评估

可采用RDP/MA会计、成员/属性推断基准等方法在威胁模型下评估隐私攻击效果与防护的隐私-效能权衡。包括：

a） 选择威胁模型：成员推断、属性推断、重识别；
b） 评估保护：匿名化、随机扰动、差分隐私（DP-SGD/查询噪声）；
c） 计算隐私指标并与任务效能联立评估。

13.9 可解释性评估

可采用SHAP/IG/LIME、ProtoPNet、规则蒸馏、反事实（DiCE）等方法量化解释的一致性、稳定性与效用，支撑人审与合规。包括：

a） 选择解释方法：特征重要性/可视化/规则抽取/反事实；
b） 组织人评或代理指标评测（正确性、精简度、覆盖率）；
c） 汇总一致性、稳定性与代价-有效性比，产出“解释可信度评分”。

13.10 人因可靠性评估

可采用NASA-TLX、Signal Detection Theory、HUDF/HRA等方法评估人在环的失效类型、风险增量与应急有效性，优化流程与界面。包括：

a） 识别关键人机接点与任务（告警确认、二次复核、接管）；
b） 设计人因实验：SOP、负荷、时间压力、信息呈现；
c） 统计错误类型与恢复策略，绘制“人因可靠度曲线（负荷—性能）”。

13.11 知识保持评估

可采用EWC、知识蒸馏、经验回放、原型记忆等方法衡量增量/持续学习中的旧知识保持、新任务增益与资源开销权衡。包括：

a） 设计任务序列与基线（初始性能），执行增量学习；
b） 周期性回归测试旧/新任务：输出保持率、遗忘度、后向/前向迁移；
c） 记录记忆缓存、蒸馏温度、EWC权重与算存开销，形成“性能—资源”帕累托。

13.12 在线自适应能力评估

可采用Confidence-based self-training、Temperature scaling/Platt scaling、经验回放（ER/Reservoir）、AutoML策略选等方法度量在线适应的速度、稳定性与风险暴露时间，比较多种自适应策略的帕累托最优解。包括：

a） 设置在线数据流与窗口策略：滑窗/指数衰减，定义适应触发条件（漂移置信度阈值、性能跌落阈值、时间/批次数间隔）；
b） 执行在线更新：参数微调（小步长/层冻结）、学习率热启动、伪标签自训练、增量校准（温度/阈值调整）、经验回放；
c） 统计适应速度与稳定性：收敛轮次、震荡幅度、回退次数；记录适应期间的最小性能与方差，评估风险暴露时间；

d） 策略对比：在“速度—稳健—成本”三轴上比较“仅校准/微调/回放/混合”，选择帕累托前沿方案；联动自动阈值与人审触发，记录误触/漏触比例。

14 智能系统不确定性量化技术

14.1 随机不确定性

随机不确定性（偶然不确定性）来源于测量噪声、环境扰动与标注差异，术语不可彻底消除的变异。可采用采用异方差建模、高斯过程回归与证据型回归等方法刻画观测噪声与环境变异，为预测提供置信区间并支持风险降级与标定优化。包括：

a） 需求定义：明确目标置信度、允许区间宽度、误报/漏报成本与关键切片（长尾、困难场景）；

b） 模型设计：回归采用异方差头输出，以高斯似然训练；小样本或需闭式后验采用GPR，选择核函数与噪声，最大化边际似然；

c） 规模化与稳定性：使用稀疏/结构化核（SKI/KISS-GP）或诱导点近似；启用Cholesky分解稳定求解并监测条件数；

d） 区间构建：依据后验或近似正态假设给出预测区间，明确置信度口径；

e） 校准与验证：报告 PICP/MPIW/NLL 与PIT/可靠图；按场景/频次/难度切片验证覆盖偏差与区间效率；

f） 部署与监控：上线监测覆盖率偏差、区间宽度漂移与触发日志，联动降级/复核策略并设阈值上下限。

14.2 认知不确定性

认知不确定性（模型不确定性）反映训练数据量不足或数据结构/参数认知不足导致的输出不确定。可采用采用贝叶斯神经网络、MC Dropout、深度集成与SWAG等后验近似技术量化模型认知不足，支持错误风险排序、OOD预警与稳健决策，在维持实时约束的同时将ECE/NLL显著降低并实现OOD不确定性分离。包括：

a） 后验近似选型：结合算存与延迟约束，在BNN（变分推断）、MC Dropout、深度集成、SWAG中择优或组合；

b） 训练与采样：BNN定义先验与变分族并用重参数化优化ELBO；MC Dropout在训练与推理均启用Dropout并设定采样次数T与层范围；集成训练M个差异化模型（初始化/数据打乱/轻架构扰动）；

c） 概率校准：在验证集进行温度缩放（全局/分组/向量），以NLL或ECE为目标；必要时对少样本切片单独校准；

d） 评估与门限：报告ECE/NLL/Brier、ID-OOD分离（AUROC）、不确定性-错误相关、延迟放大比与能耗；制定实时门限。

14.3 综合不确定性

综合不确定性同时考虑数据噪声（随机）与模型认知不足（认知）。通过方差分解与合成、证据深度学习及“温度缩放+集成/MC”组合实现对数据噪声与模型不确定的统一表征，用于动态阈值与策略联动以降低严重误判，在相同覆盖目标下取得更小PICP偏差或更窄MPIW，并在漂移/OOD场景中提升分离AUROC。包括：

a） 模块组合：回归合成总方差，据此给出置信区间；分类采用证据DL输出证据量，或先温度缩放再做集成聚合；

b） 一致性校验：验证不确定度与错误/难度单调关系，错误样本平均不确定度显著更高；

c） 策略与阈值：基于期望风险设计不确定度阈值与分流策略（人工复核/降级/拒识）；

d） 在线监控：监测PICP偏差、ID-OOD分离度、阈值触发率与暴露时间；设回退规则与人审。

15 智能系统全生命周期过程与活动

15.1 概念阶段

应基于系统目标与需求明确定义任务边界、使用场景与约束条件，以减少后续阶段偏差。其中主要活动包括：

a） 需求分析与任务定义；
b） 可行性研究（技术、数据、合规）；
c） 初始风险评估（含安全、安全性、隐私与伦理）；
d） 算法路线与资源评估与论证。

15.2 设计与开发阶段

应完成数据准备、特征工程、模型设计与训练开发，并建立相应风险控制措施与配置管理机制。其主要活动包括：

a） 数据采集、标注与质量控制；
b） 特征工程与数据处理流程设计；
c） 模型架构设计与超参数方案；
d） 训练策略设计与实现（含早停、正则化、校准与不确定性量化等）；
e） 开发与实验环境设置（含可复现性、依赖与镜像管理）；
f） 风险控制与安全机制设计。

15.3 验证与确认阶段

应依据V&V计划对模型与系统进行验证与确认，确保满足功能、性能与可靠性要求。其主要活动包括：

a） 离线模型评估（性能、鲁棒性、校准性与泛化能力）；
b） 系统集成测试（接口一致性、资源约束、端到端性能）；
c） 用户验收测试（在代表性与边界场景下进行）；
d） 压力与边界测试（连续运行、极限负载、容错与恢复）；
e） 合规性与安全性检查（数据合规、隐私保护、授权与访问控制）。

15.4 部署阶段

应将通过V&V的模型部署到目标平台，完成环境适配与上线验证，确保兼容性与可运维性。其主要活动包括：

a） 模型转换与优化（量化、裁剪、编译/加速）；
b） 目标环境适配（云/边/端形态一致性与资源约束）；
c） 服务接口开发与契约校验（API/SDK、协议与容错）；
d） 部署验证测试（功能、性能、回滚与灰度）。

15.5 运行监督阶段

应对系统运行状态进行持续监测，确保在可接受风险水平下稳定运行。其主要活动包括：

a） 实时性能与可靠性监控（吞吐、时延、错误率、可用度）；
b） 异常检测与告警（数据异常、性能退化、分布偏移/OOD）；
c） 日志与事件管理（含可追溯性与取证支持）；
d） 资源动态调度与扩缩容；
e） 安全监测（对抗性输入检测、访问与越权监测）。

15.6 重新评估阶段

应在出现数据漂移或概念漂移、性能退化或场景变化时，开展周期性或事件触发的重新评估。其主要活动包括：

a） 数据分布分析（漂移检测、代表性与覆盖性复核）；
b） 模型性能诊断（误差归因、鲁棒性复核、校准复核）；
c） 增量/迁移/持续学习实施（含回退与对比验证）；
d） 模型版本迭代与变更管理。

15.7 报废阶段

应在生命周期结束时对模型及相关数据进行合规报废或归档，防止泄露与不当使用。其主要活动包括：

a） 资产清单与依赖审计（模型、数据、密钥、证书、镜像与配置）；
b） 安全销毁或匿名化预归档（依标准或法律要求执行）；
c） 访问撤销与账户回收；
d） 报废记录与审计留存。

参 考 文 献

[1] GB/T 40647-2021 智能制造 系统架构
[2] GB/T 42755-2023 人工智能 面向机器学习的数据标注规程
[3] GB/T 43782-2024 人工智能 机器学习系统技术要求
[4] GB/T 45079-2024 人工智能 深度学习框架多硬件平台适配技术规范
[5] GB/T 45225-2025 人工智能 深度学习算法评估
[6] ISO/IEC 23053:2022 Framework for AI systems using machine learning
[7] ISO/IEC 24668 信息技术 人工智能 大数据分析的过程管理框架
[8] ISO/IEC TR 24030-2024 信息技术 人工智能 使用案例
[9] ISO/IEC TS 4213:2022 信息技术 人工智能 机器学习分类性能的评估
[10] AIOSS-01-2018 人工智能 深度学习算法评估规范