ICS 35.240
CCS L 70

T/CS

团 体 标 准

T/CS XXX—2026

AI 智能应用系统通用技术要求与验收规范

General technical requirements and acceptance specifications for ai intelligent application systems

（征求意见稿）

2026-XX-XX 发布 2026-XX-XX 实施

中国商品学会 发 布

目　　次

前言 II
1 范围 1
2 规范性引用文件 1
3 术语和定义 1
4 缩略语 2
5 系统架构要求 2
6 功能要求 2
7 性能要求 4
8 安全要求 4
9 可靠性要求 5
10 兼容性要求 5
11 验收要求 6
12 验证方法 6

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。

本文件由喀喇沁旗公安局提出。

本文件由中国商品学会归口。

本文件起草单位：喀喇沁旗公安局、上海海事大学、济青高速铁路有限公司、重庆祥富机电技术服务有限公司、金算利（深圳）科技有限公司、华媒互动网络技术（北京）有限公司、黑龙江起时科技发展有限公司。

本文件主要起草人：李明飞、卢长利、高世超、曾祥富、李子晋、杨亚洲、时凤轩。

AI 智能应用系统通用技术要求与验收规范

1 范围

本文件规定了 AI 智能应用系统的系统架构、功能要求、性能要求、安全要求、可靠性要求、兼容性要求、验收要求及验证方法。

本文件适用于 AI 智能应用系统的设计、开发、测试、验收及运行维护。

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 22239 信息安全技术 网络安全等级保护基本要求

GB/T 25000.51 系统与软件工程 系统与软件质量要求和评价（SQuaRE） 第 51 部分：就绪可用软件产品（RUSP）的质量要求和测试细则

GB/T 25058 信息安全技术 网络安全等级保护实施指南

GB/T 28035 软件系统验收规范

GB/T 35273 信息安全技术 个人信息安全规范

GB/T 42018 信息技术 人工智能 平台计算资源规范

GB/T 45225 人工智能 深度学习算法评估

GB/T 45288.2 人工智能 大模型 第 2 部分：评测指标与方法

GB/T 45652 网络安全技术 生成式人工智能预训练和优化训练数据安全规范

GB/T 45654 网络安全技术 生成式人工智能服务安全基本要求

GB/T 45674 网络安全技术 生成式人工智能数据标注安全规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

AI智能应用系统 AI intelligent application system

以人工智能模型为核心推理组件，集成数据处理、模型推理、结果输出及运维管理功能，面向特定业务场景提供智能化服务的软件系统。

3.2

兼容性 compatibility

系统与不同硬件、操作系统、接口、数据库、消息中间件及深度学习框架等协同工作并保持规定功能的能力。

3.3

吞吐量 throughput

在单位时间内系统成功完成推理或数据处理任务的数量。

4 缩略语

下列缩略语适用于本文件。

API：应用程序编程接口（Application Programming Interface）

CPU：中央处理器（Central Processing Unit）

GPU：图形处理器（Graphics Processing Unit）

MTBF：平均无故障工作时间（Mean Time Between Failures）

5 系统架构要求

5.1 系统应采用分层架构设计，至少包含数据处理层、模型推理层和应用服务层。

5.2 数据处理层应支持结构化数据和非结构化数据的采集、清洗、标注和存储。

5.3 模型推理层应支持至少一种人工智能模型的加载、推理和版本管理。

5.4 应用服务层应提供标准化的 API 接口，接口描述应符合 GB/T 25000.51 中关于功能性的要求。

5.5 系统各层之间应实现逻辑解耦，任一层的版本更新不应导致其他层的非预期中断。

5.6 系统应支持容器化部署方式，容器编排应满足弹性伸缩需求。

6 功能要求

6.1 数据管理功能

6.1.1 系统应支持结构化数据、半结构化数据和非结构化数据的接入：

a）结构化数据接入应支持关系型数据库的直连读取；

b）非结构化数据接入应支持文本、图像、音频和视频四种基本格式；

c）数据接入应支持单表采集、多表采集、增量采集、批量采集和流式采集五种模式。

6.1.2 系统应提供数据清洗、数据转换和数据归集功能，数据处理与质量校验应满足以下要求：

a）数据清洗应支持空值处理、异常值检测和重复值去除；

b）数据预处理应支持数据拆分、聚合、过滤和排序操作；

c）数据质量校验应覆盖完整性、真实性、一致性、及时性和可追溯性五个维度。

6.1.3 系统应支持文本、图像、音频和视频四类数据的标注功能，数据标注应满足以下要求：

a）标注活动应符合 GB/T 45674 的规定；

b）系统应支持多人协同标注，并具备标注任务分配和进度管理能力；

c）标注数据应具备版本管理功能，支持不同版本之间的数据分布差异分析。

6.1.4 系统应支持对原始数据、中间数据和产出数据的增删改查操作。数据操作、权限与溯源应满足以下要求：

a）数据检索应支持按表粒度和字段粒度的权限控制；

b）系统应具备敏感数据溯源管理功能，记录数据的来源、流转路径和使用记录。

6.2 模型全生命周期管理功能

6.2.1 系统应覆盖模型创建、训练、评估、部署和退役五个阶段的全生命周期管理：

a）模型创建阶段应支持模型结构定义、算法选择和超参数配置；

b）模型训练阶段应支持训练数据输入、训练过程监控和训练中断恢复；

c）模型评估阶段应依据 GB/T 45225 执行，大模型应用场景下的评测指标应符合 GB/T 45288.2 规定；

d）模型部署阶段应支持部署方式选择、资源分配和上线发布；

e）模型退役阶段应支持模型下线、归档和资源回收。模型训练应支持二次训练和模型参数的多版本保存。

6.2.2 系统应具备模型的导入、导出、更新和版本管理功能：

a）模型导入应支持从本地存储和远程对象存储两种地址形式加载模型文件，并对导入模型的格式和完整性进行校验；

b）模型导出应支持将模型文件输出至本地存储或远程对象存储，导出文件应包含模型参数和模型配置；

c）模型更新应支持模型参数的增量更新和全量更新，更新操作应记录操作人和更新时间。

d）模型版本管理应支持多版本并存、版本切换和版本回滚，各版本应记录版本号、创建时间和变更说明；

e）系统应实施基于角色的模型访问权限控制，不同角色对模型的查看、调用和修改权限应明确划分并记录。

6.2.3 系统应提供模型封装和发布能力，通过统一接口对外提供模型推理服务。

6.3 推理服务功能

6.3.1 系统应支持同步推理和异步推理两种服务模式：

a）同步推理适用于实时性要求高的场景，单次推理响应时间不应大于 500 ms；

b）异步推理适用于批量处理场景，系统应支持批量请求的提交、排队和结果回传。

6.3.2 大模型推理场景下，首词元时延的 99%分位值不应大于 500 ms，推理延迟的 99%分位值不应大于 1000 ms。系统应支持多模型并行推理，不同模型之间的推理任务不应相互阻塞。

6.3.3 推理服务应具备熔断、隔离、限流和降级四种容错机制。

a）熔断机制应在推理服务连续失败次数达到设定阈值时自动切断对该服务的调用，并在等待设定时间后进入半开状态尝试恢复；

b）隔离机制应将不同模型或不同租户的推理任务分配至独立的资源池，单一任务异常不应影响其他任务的正常执行；

c）限流机制应限制单位时间内进入推理服务的请求数量，超出限额的请求应被拒绝或排队；

d）降级机制应在推理服务出现异常时自动切换至预设的备用逻辑，返回默认结果或友好提示，不应直接抛出未处理异常。

6.4 结果输出功能

6.4.1 系统应支持结构化数据输出和自然语言输出两种结果呈现方式：

a）结构化输出应包含预测结果、置信度或不确定性标识；

b）自然语言输出应语法通顺，不应包含明显的逻辑矛盾。

6.4.2 系统应记录每次推理的输入数据摘要、模型版本号和输出结果，形成完整的推理追溯链路。追溯记录的保存周期应与日志保存周期一致。

6.5 日志与审计功能

6.5.1 系统应记录操作日志、推理日志和异常日志三类日志：

a）操作日志应覆盖用户登录/退出、权限变更、模型部署与退役等关键操作；

b）推理日志应记录每次推理请求的时间戳、输入摘要、模型版本和输出摘要；

c）异常日志应记录系统运行过程中出现的错误和告警事件。

6.5.2 日志留存时间不应少于 180 天。日志存储应采用多副本机制，确保在存储周期内日志数据不丢失、可恢复。应对审计记录进行保护，防止未预期的删除、修改或覆盖。

6.5.3 日志内容应支持按时间范围、用户标识、操作类型和事件等级四个维度进行组合检索。检索响应时间在日志数据量不超过 1000 万条的条件下不应大于 5 s。

7 性能要求

7.1 计算资源要求

7.1.1 系统的物理计算资源配置应符合 GB/T 42018 的要求。人工智能加速设备应支持乘加运算加速功能。服务器 CPU 应配备 32 位或 64 位多核处理器，三级缓存容量不应低于 32 MB。
7.1.2 系统应具备虚拟化的 CPU 资源。系统应具备至少一种虚拟化的人工智能加速处理器。系统应能实时监控虚拟计算资源的运行状态。

7.2 吞吐量要求

7.2.1 在标准测试环境下，系统推理吞吐量的实测值不应低于设计标称值的 85%。吞吐量测试应采用公认的推理性能测试方法执行。测试应持续不少于 30 min，取稳定运行阶段的数据作为测试结果。
7.2.2 数据预处理吞吐量不应低于设计标称值的 80%。流式数据接入场景下，数据处理的端到端延迟不应大于 2 s。

7.3 并发能力要求

7.3.1 系统应支持不少于 50 路并发推理请求。在满并发条件下，推理准确率的下降幅度不应超过 5%。
7.3.2 并发测试应逐步增加并发数至设计上限并保持 10 min，记录各并发等级下的响应时间和准确率数据。
7.3.3 系统应实施并发请求的优先级调度机制。在相同并发条件下，高优先级请求的平均响应时间应显著低于低优先级请求，具体调度策略和性能指标由系统设计确定并在验收测试中记录。

7.4 资源利用率要求

7.4.1 在典型负载条件下，GPU 利用率不应低于 60%。系统应提供 GPU 利用率的实时监控和告警功能。
7.4.2 在典型负载条件下，CPU 利用率不应低于 40%。系统应提供 CPU 利用率的历史趋势分析和容量规划辅助功能。

8 安全要求

8.1 数据安全要求

8.1.1 训练数据的采集、存储和处理活动应符合 GB/T 45652 的规定。对采集的互联网网站数据应记录统一资源定位符，确保数据来源的可追溯性。涉及个人信息的数据处理活动应符合 GB/T 35273 的要求。
8.1.2 数据标注活动应符合 GB/T 45674 的规定。标注过程中涉及敏感信息的，应实施脱敏处理后再进行标注操作。
8.1.3 系统应制定数据备份策略，明确备份范围、备份周期和备份介质。备份数据应支持完整恢复和增量恢复两种模式。数据恢复时间目标不应超过 4 h。

8.2 服务安全要求

8.2.1 生成式 AI 服务的提供应符合 GB/T 45654 的规定。系统应具备输入内容过滤能力，对恶意输入进行拦截。输出内容不应包含暴力、恐怖、仇恨、歧视和色情等违法和不良信息。
8.2.2 系统应具备抵抗提示词注入和越狱攻击的能力。系统应定期开展对抗性测试，验证模型在异常输入条件下的输出稳定性。测试应采用白盒和黑盒两种对抗攻击方式。

8.3 网络安全要求

8.3.1 系统的网络安全等级保护应符合 GB/T 22239 中相应等级的要求。等级保护的实施应符合 GB/T 25058 的规定。
8.3.2 应在网络边界、重要网络节点、服务器和应用程序四个层面实施安全审计。审计范围应覆盖每个用户，对重要用户行为和重要安全事件进行审计。审计记录应包含以下内容：

a) 日期；
b) 时间；
c) 用户标识；
d) 事件类型；
e) 事件描述；
f) 事件结果。

8.4 访问控制要求

8.4.1 系统应实施基于角色的访问控制机制。管理员操作应支持双因素认证。系统应对远程访问的用户行为和访问互联网的用户行为单独进行行为审计。
8.4.2 系统应实施最小权限原则。用户权限的授予、变更和撤销应经过审批流程并记录操作日志。

9 可靠性要求

9.1 系统可用性不应低于 99.5%。系统应提供可用性监控和定期报告功能。
9.2 系统应具备故障自动恢复能力。单次故障的恢复时间不应超过 30 min。系统应支持服务级别的熔断、隔离、限流和降级，确保单点故障不扩散至整个系统。
9.3 系统 MTBF 不应低于 800 h。系统应支持在额定负载 50%以上的强度下连续运行不少于 72 h 而不出现非计划停机。
9.4 系统存储数据的持久性不应低于 99.9%。模型文件的存储应支持版本回滚，回滚操作的时间不应超过 10 min。

10 兼容性要求

10.1 硬件兼容性要求

10.1.1 系统应支持 x86 和 ARM 两种 CPU 架构。系统在两种架构上的功能可用性应保持一致。
10.1.2 系统应支持至少两种主流人工智能加速处理器。加速器的驱动程序应支持主流深度学习框架的调用。

10.2 操作系统兼容性要求

10.2.1 系统应支持至少两种主流操作系统。
10.2.2 系统在每种目标操作系统上的功能测试通过率应为 99%。

10.3 接口兼容性要求

10.3.1 系统提供的 API 接口应兼容 RESTful 风格。接口描述文档应完整、可机器解析。接口调用成功率在正常网络条件下的测试值不应低于 99.9%。
10.3.2 系统应支持与主流关系型数据库、NoSQL 数据库和消息中间件的对接。数据接口的调用一致性应符合 GB/T 25000.51 中关于兼容性的要求。

10.4 框架兼容性要求

10.4.1 系统应支持主流深度学习框架的模型导入和推理调用。
10.4.2 深度学习框架应支持多硬件平台适配。

11 验收要求

11.1 验收条件

系统完成开发并通过内部测试后，方可进入验收阶段。验收应依据 GB/T 28035 中的要求组织实施。

11.2 验收内容

验收内容应覆盖第 6 章至第 10 章规定的全部技术要求。

11.3 验收准则

11.3.1 功能验收：所有功能项应逐项通过测试，通过率应为 99%。
11.3.2 性能验收：第 7 章各项性能指标实测值不应低于规定值。
11.3.3 安全验收：第 8 章各项安全要求应全部满足，不得存在未修复的高危漏洞。
11.3.4 可靠性验收：系统连续运行测试不应少于 72 h，期间不应出现非计划停机。
11.3.5 兼容性验收：第 10 章各项兼容性要求应逐项验证通过。

11.4 验收结论

验收结论分为“通过”和“不通过”两种。全部验收项满足准则时判定为“通过”；任一验收项不满足准则时判定为“不通过”。验收不通过时，应出具整改意见，整改完成后重新组织验收。

12 验证方法

12.1 功能验证

12.1.1 功能验证采用黑盒测试方法。测试用例应覆盖第 6 章全部功能项，每个功能项测试用例数量不应少于 3 个。
12.1.2 测试结果记录应包含输入数据、预期输出和实际输出。

12.2 性能验证

12.2.1 性能测试应在系统达到稳定运行状态后进行。测试工具应具备压力生成和响应时间采集能力。
12.2.2 吞吐量测试应持续不少于 30 min，取稳定阶段数据作为测试结果。
12.2.3 并发测试应逐步增加并发数至设计上限并保持 10 min。

12.3 安全验证

12.3.1 数据安全验证应审查数据处理流程和相关配置，确认符合 GB/T 45652 的要求。
12.3.2 服务安全验证应依据 GB/T 45654 规定的证实方法执行。
12.3.3 网络安全验证应按 GB/T 22239 规定的测试方法执行。

12.4 可靠性验证

12.4.1 连续运行测试应连续执行不少于 72 h，测试期间以不低于额定负载 50%的强度持续运行。
12.4.2 故障恢复测试应采用模拟故障注入方式，记录从故障发生到系统恢复服务的时间。

12.5 兼容性验证

12.5.1 CPU 架构兼容性测试应在 x86 和 ARM 环境中分别部署系统并执行功能测试。

12.5.2 操作系统兼容性测试应在每种目标操作系统上完成完整的功能测试。

12.5.3 API 兼容性测试应验证接口的请求和响应格式符合接口描述文档。

12.6 验收测试报告

12.6.1 验证完成后应编制验收测试报告。

12.6.2 报告内容应包括：

a) 测试环境描述；

b) 测试用例清单；

c) 测试结果汇总；

d) 不符合项清单及整改建议。