ICS 35. xxx
CCS Lxx

团　　体　　标　　准

T/ISC XXX—XXXX

具身智能实景应用能力要求及评价方法

Requirements and evaluation methods for real-scenario application capability of embodied intelligent systems

（征求意讲稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中 国 互 联 网 协 会　　发布

目　　次

前　　言 3
具身智能实景应用能力要求及评价方法 4
1 范围 4
2 规范性引用文件 4
3 术语和定义 4
3.1 具身智能系统 embodied intelligent system 4
3.2 场景任务 scenario task 4
3.3 场景挑战度 scenario challenge level 4
3.4 应用成熟度 application maturity level 4
3.5 测试实例 test instance 4
3.6 有效测试实例 valid test instance 4
3.7 等效测试场景 equivalent test scenario 4
3.8 人工介入 human intervention 5
3.9 人工接管 human takeover 5
3.10 安全否决事件 safety veto event 5
4 符号和缩略语 5
5 总体要求 5
5.1 评价原则 5
5.2 C-L 场景应用评价模型 5
5.3 评价边界 6
6 场景任务与场景挑战度 6
6.1 场景任务确定 6
6.2 场景任务特征 6
6.3 场景挑战度分级 7
7 实景应用能力要求 7
7.1 总则 7
7.2 任务执行能力 8
7.3 自主作业能力 8
7.4 稳定可靠能力 8
7.5 安全协同能力 9
7.6 应用效能与经济性 9
8 实景应用成熟度评价方法 9
8.1 总则 9
8.2 评价准备与数据获取 10
8.3 任务执行能力评价方法 11
8.4 自主作业能力评价方法 13
8.5 稳定可靠能力评价方法 14

8.6 安全协同能力评价方法 15
8.7 应用效能与经济性评价方法 16
8.8 评价数据有效性与统计处理 17
9 评价与结果表示 18
9.1 L 等级判定规则 18
9.2 C-L 评价结果表示 20
9.3 综合评分 20
9.4 评价报告 20
附 录 A （规范性） 场景挑战度确定方法 21
A.1 挑战因素 21
A.2 C 等级确定 21
附 录 B （资料性） 典型应用场景参考 23
附 录 C （资料性） 典型评价方案参考 24
C.1 通用方案结构 24
C.2 工业制造：零件上料与装配 25
C.3 仓储物流：混合货品拣选与搬运 26
C.4 商业服务：动态公共区域递送 26
附 录 D （资料性） 综合评分计算规则 28
D.1 总则 28
D.2 指标与权重 28
D.3 指标得分换算 28
D.4 综合评分计算 29
D.5 安全否决与结果表示 30
D.6 计算示例 30
附 录 E （资料性） 安全否决条件参考 31
E.1 总则 31
E.2 典型安全否决条件 31
E.3 使用说明 31

前　言

本文件按照中国互联网协会团体标准管理有关规定组织制定。

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。

本文件由中国互联网协会提出并归口。

本文件起草单位：中国信息通信研究院、北京石景山产业发展有限公司、首钢基金冬奥广场人形机器人产业基地、中石油（北京）数智研究院、北京交通大学、湖北大数据集团、杭州航迹中试验证科技有限公司

本文件主要起草人：洪逸伦、王玮、王扬、黄伟、李晓蓉、吴海楠、赵方艾、任基泽、霍向、金攀、郑献明

具身智能实景应用能力要求及评价方法

1 范围

本文件规定了具身智能系统实景应用能力的总体要求、场景任务与场景挑战度、实景应用能力要求、应用成熟度评价方法以及评价与结果表示。

本文件适用于具身智能系统在工业制造、仓储物流、商业服务、医疗康养、家庭服务、特种作业等领域的真实场景或等效场景中的应用能力测试与评价，可用于应用验证、产品选型、项目验收、实景实训中的系统应用能力验证及第三方评价。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1 具身智能系统 embodied intelligent system

具备环境感知、任务理解、决策规划和物理执行能力，并能够通过与物理环境交互完成规定任务的智能系统。

3.2 场景任务 scenario task

具身智能系统在规定环境、作业对象及约束条件下，为实现特定作业目标所执行的一组连续或离散作业活动。

3.3 场景挑战度 scenario challenge level

由作业环境、作业对象、动态交互、任务过程以及异常与扰动等因素共同形成的，对具身智能系统完成规定场景任务所提出的综合挑战程度。

注：场景挑战度采用 C 表示。

3.4 应用成熟度 application maturity level

具身智能系统在规定场景任务条件下，自主、稳定、安全、有效完成实际作业任务所达到的应用成熟程度。

注：应用成熟度采用 L 表示。

3.5 测试实例 test instance

按照评价方案规定的场景变量、初始条件、任务目标和判定规则组织实施的一次完整测试活动。

3.6 有效测试实例 valid test instance

测试条件、实施过程和数据记录符合评价方案要求，能够用于指标计算和等级判定的测试实例。

3.7 等效测试场景 equivalent test scenario

通过仿真、数字孪生、替代对象、限能运行或者其他方式构建，并经核验证明能够在规定边界内表征目标真实场景关键特征的测试场景。

3.8 人工介入 human intervention

任务执行过程中，人员通过提示、参数调整、人工重置、遥操作或者其他方式影响系统自主执行过程的行为。

3.9 人工接管 human takeover

任务执行过程中，人员取得系统控制权，并直接完成任务动作或者实施安全处置的行为。

3.10安全否决事件 safety veto event

评价过程中发生的、依据评价方案不能由其他指标得分或者综合评分补偿，并导致评价结果判为不通过或者未定级的安全事件。

4 符号和缩略语

下列符号和缩略语适用于本文件。

C——场景挑战度；

L——应用成熟度；

MTBF——平均无故障运行时间（mean time between failures）；

P90——第90百分位数。

5 总体要求

5.1 评价原则

本文件的评价对象为完成规定场景任务的完整具身智能系统，包括机器人本体、感知系统、决策与控制系统、具身智能模型、末端执行机构及完成任务所需的配套软硬件。

本文件不限定具身智能系统的本体形态。

本文件所形成的评价结果仅用于表征规定场景任务和评价边界内的应用能力，不替代适用的产品安全、系统安全、功能安全、医疗器械注册检验、特种设备监督检验及其他行业准入要求的符合性评价，也不作为行政许可的直接依据。涉及医疗器械功能、人体接触或者其他特定行业安全要求的，应符合相应法律法规和专项标准。

a）场景任务导向。应以具身智能系统完成实际场景任务的能力和效果为主要评价对象。

b）系统整体评价。应对完成规定任务的完整系统进行评价，宜优先采用能够直接反映实际作业结果的指标。

c）客观可验证。评价指标应具有明确的数据来源和计算方法，测试过程应可重复，评价结果应可追溯。

d）分类适配。应根据场景任务特点和评价目的选择适用指标，并确定相应测试条件。

e）边界明确。评价结果应与被评价系统版本、场景任务、场景挑战度和测试条件绑定，不应超出评价边界使用。

5.2 C-L 场景应用评价模型

具身智能系统实景应用能力采用C-L场景应用评价模型。

a）场景挑战度C用于表征规定场景和任务条件对具身智能系统形成的挑战程度；

b）应用成熟度L用于表征具身智能系统在规定场景任务条件下达到的实际应用成熟程度。

C应在正式测试前确定，仅表征测试条件和评价边界，不直接表征被评价系统能力；L应依据测试结果及预先规定的等级门槛判定。不同C等级下均可形成相应的L评价结果。

基本评价结果采用C-L形式表示。根据产品选型、方案比较等需要，可在C-L结果后附综合评分。

5.3 评价边界

开展评价前应明确下列评价边界：

a）被评价系统组成、主要硬件配置、软件及模型版本；

b）应用场景、核心任务和任务成功判据；

c）任务起始状态、结束状态及允许的作业范围；

d）测试条件、场景变量，以及完成任务所依赖的本地、边缘和云端组件、通信条件、外部设备与服务；

e）允许的外部辅助条件、人工支持和人工介入方式，以及系统内组成、系统外必要条件及其责任边界；

f）评价结论的适用范围和限制条件

评价期间外部依赖、辅助条件或者人工支持发生变化时，应留存记录并分析其对评价结果的影响。

6 场景任务与场景挑战度

6.1 场景任务确定

具身智能系统场景任务应根据实际应用需求明确应用领域、典型场景和核心任务，至少应明确：

a）任务目标；

b）任务起始状态和结束状态；

c）作业对象及主要环境条件；

d）主要作业流程；

e）任务成功和失败判据。

复杂任务可按实际业务流程划分为若干作业单元。作业单元可包括识别、定位、移动、抓取、搬运、放置、装配、检测、设备操作和状态确认等活动。

6.2 场景任务特征

场景任务应从下列方面描述对任务执行具有显著影响的特征，并明确其固定状态、变化范围、离散取值或组合方式：

a）作业环境，包括空间布局、通道、地面、光照、温湿度及其他物理环境条件；

b）作业对象，包括对象种类、数量、尺寸、重量、材质、位置、姿态及其变化；

c）动态交互，包括人员、车辆、设备及其他动态主体的位置、运动状态和交互情况；

d）任务过程，包括任务顺序、目标位置、任务参数、任务路径和流程变化；

e）异常与扰动，包括遮挡、目标缺失、路径变化、临时障碍、通信波动和作业条件改变等情况。

6.3 场景挑战度分级

表 1 场景挑战度等级

等级	名称	描述
C1	固定条件级	场景和任务条件基本固定，环境、对象和任务过程具有较高确定性，无明显动态交互和异常扰动。
C2	受控变化级	场景总体稳定，部分变量在预先规定范围内变化，任务过程总体确定，存在少量可预期动态因素。
C3	典型实景级	场景能够反映实际业务中的典型变化，存在多种对象变化、环境变化或动态交互，系统需根据实际状态调整任务执行。
C4	复杂动态级	多个变量明显变化，任务过程中存在多主体动态交互或多类常见异常，系统需自主调整并处置异常。
C5	开放挑战级	场景具有较高开放性和不确定性，可能出现未完全预先定义的变化，系统需依据实时状态自主规划和调整。其中，开放性限定在预先声明的应用运行边界和安全边界内。

场景挑战度C应在正式测试前按照附录A确定，并在评价过程中核验实际测试条件未超出相应等级边界。

表2和表A.1中的“少量”“多种”“明显”“多主体”“频繁”等定性描述，应在评价方案中结合具体场景任务转换为可核验的变量数量、取值范围、组合方式、发生频次、变化速率或者时间尺度，并说明判定依据。C5所称开放性是指在预先声明的应用运行边界和安全边界内存在未完全预定义的变化，不表示无限开放环境。

7 实景应用能力要求

7.1 总则

具身智能系统实景应用能力主要包括任务执行能力、自主作业能力、稳定可靠能力、安全协同能力、应用效能与经济性。开展应用成熟度等级评价时，应对五类能力维度进行适用性分析，并根据拟判定等级和场景任务特点，将具体评价指标确定为必选、可选或不适用。

应用成熟度L用于表征具身智能系统在规定场景任务条件下达到的实际应用成熟程度，划分为L1至L5五个等级。

表 2 应用成熟度等级

等级	名称	描述
L1	演示验证级	具身智能系统具备完成核心任务的基本能力，能够形成可验证的任务结果。
L2	受控作业级	具身智能系统能够重复完成规定任务，主要任务性能达到基本应用要求。
L3	场景适用级	具身智能系统能够以规定自主程度完成任务，人工介入和人工接管处于规定范围内。
L4	稳定作业级	具身智能系统能够持续稳定执行规定任务，并能够处置规定的常见任务异常。
L5	规模部署级	具身智能系统能够满足常态化或规模化部署所需的任务性能、自主程度、稳定可靠性、安全协同能力及应用效能要求。

L等级描述规定评价目标，相应指标、样本量和门槛应结合具体场景任务在评价方案中量化。同一指标可针对不同L等级规定不同的样本量、持续运行要求和指标门槛。指标选择应优先反映完整任务的实际执行效果，避免以多个高度相关指标重复评价同一能力。

7.2 任务执行能力

任务执行能力用于评价系统完成规定任务目标并达到规定作业质量的能力。宜根据场景任务选择：

a）任务完成率，表征系统完整完成规定场景任务的能力；

b）作业单元完成率，表征系统完成复杂任务中各作业单元的能力；

c）作业质量达标率，表征任务结果满足规定质量要求的程度；

d）操作精度，表征实际操作结果与目标位置、姿态、力、轨迹或其他规定参数之间的偏差；

e）任务完成时间，表征系统从规定任务起始状态至达到结束状态所需的时间；

f）节拍达标率，表征任务或作业单元在规定节拍内完成的程度。

能够采用完整任务评价时，应优先采用任务完成率。作业质量已经纳入任务成功判据时，可不重复设置作业质量达标率。

7.3 自主作业能力

自主作业能力用于评价系统在减少人工参与条件下自主完成规定任务的能力。宜根据场景任务选择：

a）自主完成率，表征系统在无规定外人工介入条件下完成任务的能力；

b）人工接管率，表征任务执行过程中需要人员取得系统控制权，并直接完成任务动作或安全处置的的程度；

c）人工介入频度，表征任务执行对提示、参数调整、人工重置或其他方式影响系统自主执行过程的人工参与依赖程度；

d）异常自主处置成功率，表征系统自主识别、处置规定异常并恢复任务的能力。

7.4 稳定可靠能力

稳定可靠能力用于评价系统在规定运行时间或任务量条件下持续保持正常运行及任务性能的能力。宜根据场景任务选择：

a）系统可用率，表征计划运行时间内系统处于可执行规定任务状态的时间比例；

b）故障发生率，表征单位运行时间、任务量或作业量内系统故障发生的频度；

c）平均无故障运行时间，表征系统相邻两次故障之间的平均正常运行时间；

d）故障恢复时间，表征系统发生故障后恢复至规定任务能力所需的时间；
e）性能一致性，表征系统在重复或持续运行中主要任务性能的稳定程度。

7.5 安全协同能力

安全协同能力用于评价系统与人员、设备及环境交互过程中保障安全的能力。应根据场景风险选择：
a）人员或障碍物检测成功率，表征系统正确检测规定人员或障碍物的能力；
b）碰撞避免成功率，表征系统在规定交互条件下避免碰撞的能力；
c）安全停止或急停响应时间，表征触发安全停止或急停后系统完成规定响应所需的时间；
d）安全事件发生情况，表征危险动作、安全事件以及设备或作业对象损伤事件的发生情况；
e）安全状态保持能力，表征系统在规定故障或异常条件下进入并保持安全状态的能力。

不存在人员或动态对象交互的封闭、隔离场景，可根据实际风险确定适用指标。涉及产品本体安全、功能安全等已有专门标准规定的要求时，应按照相应标准执行。

评价方案应根据被评价系统形态、应用领域和风险评估结果，明确实际适用的法律法规、强制性标准和专项标准及其适用条款。未构成本文件统一评价方法直接依据的专项标准，可作为特定评价方案的符合性依据，不宜不加区分地用于全部具身智能系统。

7.6 应用效能与经济性

应用效能与经济性用于评价系统相对于规定基准作业方式形成的实际应用价值。宜根据评价目的选择：
a）效率提升率，表征系统相对于规定基准作业方式的作业效率提升程度；
b）作业质量提升率，表征系统相对于规定基准作业方式的作业质量提升程度；
c）人工投入降低率，表征系统相对于规定基准作业方式减少人工工时或人员数量的程度；
d）高风险作业暴露降低率，表征系统相对于规定基准作业方式降低人员高风险作业暴露的程度；
e）成本类指标，表征系统的单位任务成本、运行维护成本、投资回收期或全生命周期成本。

应用效能与经济性指标在L1至L4等级中一般作为可选指标；评价目的涉及效能目标、产品选型或项目验收时，可在评价方案中规定为相应等级的必选指标。L5应根据常态化或规模化部署目标选取适用指标，并作为必选指标。

8 实景应用成熟度评价方法

8.1 总则

具身智能系统实景应用成熟度评价应以规定场景任务下的实际作业结果为依据，通过测试获取评价数据，并按照本章规定的方法计算评价指标。

评价应按照第7章规定对五类能力维度进行适用性分析，并根据拟判定等级、场景任务特点和评价目的选择具体指标。评价指标应标明必选、可选或不适用，其属性和相应门槛应在评价实施前确定。

本文件规定的评价指标不能满足具体场景任务的评价需要时，委托方可参照本文件规定的指标结构，根据场景任务特点补充或调整评价指标。补充或调整的评价指标应具有明确的名称、定义、适用条件、数据来源、计算方法和判定门槛，并在评价实施前经评价实施机构技术评审后纳入评价方案。补充或调整的评价指标不应降低适用的安全要求，也不应改变已选指标的统计口径。

评价所采用的指标、计算方法、样本量、统计规则和指标门槛应在评价实施前确定，不应在获得评价数据后进行影响评价结论的调整。

各项指标的评价结果应与预先规定的指标门槛进行比较，作为第9章应用成熟度L等级判定的依据。

8.2 评价准备与数据获取

8.2.1 评价方案

正式评价前应编制评价方案。评价方案至少应包括：

a）评价目的和应用场景；

b）被评价系统的组成、主要配置、软件及模型版本；

c）场景任务、起始状态、结束状态、作业单元和任务成功判据；

d）场景挑战因素、场景挑战度C及其确定依据；

e）拟判定的应用成熟度等级，评价指标及其必选、可选或不适用属性，计算方法、指标门槛和安全否决条件；

f）测试实例、实例生成方式、最小有效样本量和持续运行要求；

g）允许的外部辅助条件以及人工介入、人工接管的分类和记录规则；

h）数据采集、有效性判定、统计处理和异常值处理方法；

i）安全风险、测试停止条件和应急处置措施。

评价方案还应规定测试暂停、恢复、补测和重测条件。不得通过替换失败实例或者反复追加实例直至达到门槛形成通过结论；采用序贯或者自适应样本方案时，应预先规定停止边界和风险控制方法。整改后实施的测试应与原版本测试结果区分，原失败记录应保留。

8.2.2 评价条件和测试实例

评价条件应与规定场景任务及场景挑战度C相一致。

评价方案应规定测试实例的初始物理状态和软件状态、预热与校准要求、电量或供能状态、任务顺序、会话与记忆状态、重复次数及实例间复位方法；涉及随机过程且能够获取随机性设置或者随机种子时，应予以记录。

测试实例应依据规定的场景变量生成，并覆盖对任务结果具有显著影响的主要变化因素。测试实例可采用随机、分层随机或规定组合方式生成，不宜仅采用单一固定条件。

用于评价系统适应能力时，可设置正式测试前未向被评价方明确具体取值或具体组合的测试实例，但场景变量类型、取值范围和安全边界应在评价方案中规定。采用等效场景开展评价时，应说明其与目标真实场景在作业环境、作业对象、动态交互、任务过程以及异常与扰动等方面的等效关系和适用边界。采用仿真、数字孪生等方式时，应根据评价目的核验关键物理参数、动力学特性、感知噪声和扰动模式，并提供必要的真机对照验证证据。未经等效性核验的数据不应直接用于应用成熟度等级判定。评价报告应区分真机测试、仿真测试和其他等效方式获得的数据，并说明核验依据和结论限制。样本量应根据指标类型、指标门槛、任务周期、结果波动程度和评价风险确定。评价方案应规定最小有效样本量。用于L4、L5等级判定的评价，应规定持续运行时间、累计任务数量或累计作业量。

正式测试过程中，不应针对具体测试实例调整可能实质影响评价结果的模型、策略或关键运行参数。因安全处置或系统恢复需要进行调整时，应记录调整原因、调整内容和影响范围，并按照评价方案判定相关测试实例的有效性。

正式测试前应锁定被评价系统的硬件配置、软件及模型版本和关键运行参数。固定配置评价中，正式测试实例及其判定信息不得用于该轮被评价对象的训练、调参或者规则优化；测试过程中不应进行可能影响本轮结果的在线学习。评价目的确需考察在线学习或现场适应能力时，应在评价方案中预先规定适应数据、学习阶段、独立验证实例和结果表示方式，并分别记录学习前后版本及其时间顺序。

8.2.3 测试实施与记录

任务执行能力和自主作业能力宜在完整场景任务中同步测试。稳定可靠能力应通过规定运行时间、任务数量或累计作业量的持续测试获取数据。安全协同能力应根据场景风险设置测试条件，不应以可能对人员、设备或环境造成不可接受风险的方式实施测试。

每个测试实例至少应记录：

a) 实例编号、场景变量和任务条件；

b) 任务开始时间、结束时间和持续时间；

c) 任务成功或失败结果及其判定依据；

d) 作业质量、操作精度及其他任务性能数据；

e) 人工提示、参数调整、遥操作、人工重置和人工接管事件；

f) 任务异常、系统自主处置及任务恢复结果；

g) 系统故障、停机和恢复情况；

h) 人员、设备、作业对象及环境相关的安全事件；

i) 原始数据、运行日志、影像记录或其他可追溯证据。

不同数据源之间应采用统一时钟或建立可追溯的时间对应关系。

8.3 任务执行能力评价方法

8.3.1 评价实施

任务执行能力应在规定的任务起始状态、结束状态、成功判据和质量要求下进行评价。完整任务能够直接反映实际作业结果时，应优先采用任务完成率；复杂任务需要分析中间作业过程时，可增加作业单元完成率。

作业质量已经纳入任务成功判据时，可不重复设置作业质量达标率，但应保留相应质量检测记录。

8.3.2 任务完成率

任务完成率按式（1）计算：

$$P_T = N_S / N \times 100\% \qquad (1)$$

式中：

P_T——任务完成率；

N_S——满足任务成功判据的测试实例数；

N——有效测试实例总数。

任务因被评价系统故障、超时、规定外人工接管或未达到任务成功判据而终止的，应计为任务失败。

8.3.3 作业单元完成率

作业单元完成率按式（2）计算：

$$P_U = N_US / N_U \times 100\% \qquad (2)$$

式中：

P_U——作业单元完成率；

N_US——成功完成的作业单元数；

N_U——实际执行的有效作业单元总数。

评价方案应规定作业单元的划分方法和每个作业单元的成功判据。

8.3.4 作业质量达标率

作业质量达标率按式（3）计算：

$$P_Q = N_QS / N_Q \times 100\% \quad (3)$$

式中：

P_Q——作业质量达标率；

N_QS——作业质量达到规定要求的次数；

N_Q——实施质量检测的有效次数。

质量检测方法、检测设备、允许偏差和达标门槛应在评价方案中规定。

8.3.5 操作精度

操作精度应根据任务特点，采用位置误差、姿态误差、力或力矩误差、轨迹偏差、数量误差等适用参数进行评价。单次操作误差可按式（4）计算：

$$e_i = |x_i - x_0| \quad (4)$$

式中：

e_i——第i次操作误差；

x_i——第i次操作的实测值；

x_0——目标值或基准值。

评价结果应给出测量参数、单位、样本量以及均值、中位数、最大值或规定分位数。涉及多维参数时，应在评价方案中规定误差合成方法。

式（4）适用于具有明确目标值的标量绝对误差。位置向量、姿态、轨迹及其他非标量结果应采用与任务特点相适应的距离、角度、轨迹偏差或者其他计算方法，并在评价方案中明确。

8.3.6 任务完成时间

任务完成时间表征系统完成单次规定任务所需的时间。单次任务完成时间按式（5）计算：

$$T_i = t_e,i - t_s,i \quad (5)$$

式中：

T_i——第i个成功完成任务的任务完成时间；

t_e,i——第i个任务达到规定结束状态的时刻；

t_s,i——第i个任务达到规定开始状态的时刻。

任务完成时间宜报告样本量、中位数和P90，必要时报告均值、最大值和离散程度。

任务完成时间仅对成功完成的任务进行统计。超时任务应按照任务成功判据计为任务失败，不应因未纳入完成时间统计而忽略。

8.3.7 节拍达标率

节拍达标率表征在规定节拍内完成任务的测试实例占有效测试实例总数的比例，按式（6）计算：

$$P_C = N_C / N \times 100\% \quad (6)$$

式中：

P_C——节拍达标率；

N_C——在规定节拍内完成任务的测试实例数；

N——有效测试实例总数。

8.4 自主作业能力评价方法

8.4.1 评价实施

自主作业能力评价应记录任务执行过程中的人工提示、参数调整、遥操作、人工重置和人工接管等事件，并记录事件发生时间、原因、持续时间及其对任务结果的影响。

评价方案应区分正常业务输入和人工介入。任务开始前按照规定方式下达任务指令，不宜计为人工介入；任务执行过程中由人员对系统行为进行提示、修正、控制或接管的，应计为人工介入。发生人工接管的任务不应计入自主完成任务。

8.4.2 自主完成率

自主完成率按式（7）计算：

$$P_A=N_A/N\times 100\% \qquad (7)$$

式中：

P_A——自主完成率；

N_A——除评价方案预先规定的正常业务输入外，未发生影响任务决策或执行的人工介入且成功完成任务的测试实例数；

N——有效测试实例总数。

8.4.3 人工接管率

人工接管率按式（8）计算：

$$P_H=N_H/N\times 100\% \qquad (8)$$

式中：

P_H——人工接管率；

N_H——发生人工接管的测试实例数；

N——有效测试实例总数。

同一测试实例发生多次人工接管时，在人工接管率计算中按一个发生接管的测试实例统计，同时应另行记录接管次数和累计接管时间。

8.4.4 人工介入频度

人工介入频度按式（9）计算：

$$F_I=N_I/Q \qquad (9)$$

式中：

F_I——人工介入频度；

N_I——人工介入事件总数；

Q——评价方案规定的任务数量、运行时间或作业量。

人工介入频度的统计单位应在评价方案中规定，例如次/任务、次/小时或次/百件。

8.4.5 异常自主处置成功率

异常自主处置成功率按式（10）计算：

$$P_E=N_ES/N_E\times 100\% \qquad (10)$$

式中：

P_E——异常自主处置成功率；

N_ES——系统自主识别、处置异常并恢复规定任务状态的次数；

N_E——评价方案规定且实际发生的异常次数。

仅应统计评价方案规定的异常类型。异常处置过程中发生规定外人工介入或人工接管的，不应计为自主处置成功。

8.5 稳定可靠能力评价方法

8.5.1 评价实施

稳定可靠能力应在规定运行时间、任务数量或累计作业量条件下持续评价。评价过程中应记录系统故障的发生时间、故障现象、影响范围、故障原因、恢复措施和恢复时间。由被评价系统自身故障导致的任务中断、停机或性能下降，不应作为无效测试剔除。

用于L4、L5等级判定或者形成稳定可靠性结论时，评价方案应根据待验证门槛、任务周期、故障风险和结果波动程度规定最低运行时间、任务数量或累计作业量。样本或观察时长不足以支持可靠性推断时，只报告实际观察结果，不应外推形成超出证据范围的可靠性等级结论；采用置信区间、可靠性下限或者其他统计推断方法时，应明确置信水平、计算方法和适用条件。

8.5.2 系统可用率

系统可用率按式（11）计算：

$$A = T_A / T_P \times 100\% \qquad (11)$$

式中：

A——系统可用率；

T_A——系统处于能够执行规定任务状态的时间；

T_P——计划运行时间。

由测试设施故障、外部供电中断等非被评价系统原因造成的停机时间，可按照评价方案规定从计划运行时间中扣除。由被评价系统故障、恢复、重启或维护造成的停机时间不应扣除。

8.5.3 故障发生率

故障发生率应按运行时间、任务数量或累计作业量进行计算，计算口径应在评价方案中规定。按运行时间计算时，故障发生率按式（12）计算：

$$\lambda = N_F / T_O \qquad (12)$$

式中：

λ——故障发生率；

N_F——评价期间发生的系统故障次数；

T_O——有效运行时间。

评价结果应同时给出故障总数、故障分类和统计单位。

8.5.4 平均无故障运行时间

平均无故障运行时间按式（13）计算：

$$T_MTBF = T_O / N_F \qquad (13)$$

式中：

T_MTBF——平均无故障运行时间；

T_O——有效运行时间；

N_F——评价期间发生的系统故障次数。

评价期间未发生故障时，不应将平均无故障运行时间表示为无穷大，应表述为“在规定观察时长内未发生故障”，并给出实际观察时长。

8.5.5 故障恢复时间

故障恢复时间应从故障发生或系统失去规定任务能力的时刻起计算，至系统恢复规定任务能力的时刻止。评价结果宜给出每次故障恢复时间，以及恢复时间的均值、中位数、最大值或P90。需要人工参与恢复时，应同时记录人工操作内容和人工投入时间。

8.5.6 性能一致性

性能一致性应采用能够反映主要任务性能波动的统计量进行评价，可采用标准差、极差、变异系数或分时段达标率。当采用变异系数时，按式（14）计算：

$$CV = s/\bar{x} \times 100\% \qquad (14)$$

式中：

CV——变异系数；

s——主要性能指标的样本标准差；

$\bar{x}$——主要性能指标的样本均值。

当样本均值为零或接近零时，不宜采用变异系数，应采用标准差、极差或其他适用统计量。

8.6 安全协同能力评价方法

8.6.1 评价实施

安全协同能力评价应根据安全风险评估结果，确定人员、设备、作业对象和环境相关的安全测试条件、评价指标和安全否决条件。

可能造成不可接受风险的测试，可采用仿真装置、替代对象、限能运行或其他等效方式实施。采用等效方式时，应说明其与目标风险条件的对应关系。涉及机器人产品安全、功能安全或特定行业安全的要求时，应按照相应标准执行。

8.6.2 人员或障碍物检测成功率

人员或障碍物检测成功率按式（15）计算：

$$P_D = N_D / N_O \times 100\% \qquad (15)$$

式中：

P_D——人员或障碍物检测成功率；

N_D——系统在规定时间内正确检测到人员、障碍物或危险状态的次数；

N_O——人员、障碍物或危险状态有效出现的次数。

评价时应同时记录漏检、误检以及误检对正常作业造成的影响。

8.6.3 碰撞避免成功率

碰撞避免成功率按式（16）计算：

$$P_CA = N_CA / N_CR \times 100\% \qquad (16)$$

式中：

P_CA——碰撞避免成功率；

N_CA——系统自主采取规定的避让、减速、停止或者其他安全措施，避免非预期危险接触且未产生次生危险的测试次数；

由安全员介入、人工接管或者外部保护措施直接避免事故的，不应计为系统自主避碰成功，应单独记录。必要的安全停止可判为安全处置成功，但不应自动计为任务完成或者自主恢复成功。

N_CR——评价方案规定的有效碰撞风险事件次数。

8.6.4 安全停止或急停响应时间

安全停止或急停响应时间应从规定安全触发条件满足的时刻起计算，至安全停止或急停动作实际生效的时刻止。评价方案应明确安全触发点、安全动作生效点、计时精度和响应时间门槛。评价结果宜给出最大值和P90。

8.6.5 安全事件发生情况

评价过程中应记录危险动作、碰撞、人员伤害、设备或作业对象损伤、超出安全边界运行以及安全功能失效等事件的次数、后果和处置结果。

8.6.6 安全状态保持能力

对于规定的系统故障或任务异常，应检查系统能否进入并保持规定的安全状态，必要时统计安全状态保持成功率。评价方案规定为安全否决条件的事件一旦发生，应按照第9章进行判定，不得通过其他指标得分或综合评分进行补偿。

8.7 应用效能与经济性评价方法

8.7.1 评价实施

应用效能与经济性应通过被评价系统与规定基准作业方式的对照数据进行评价。基准作业方式应具有明确来源，并与被评价系统保持相同或可比的任务目标、作业质量、作业范围、工作时间和统计周期。无法保持一致的，应说明差异及其对评价结果的影响。

评价边界应明确是否包括设备购置、部署改造、能源消耗、耗材、人员监控、人工介入、运行维护、停机损失和残值等因素。

8.7.2 效率提升率

效率提升率按式（17）计算：

$$R_E = (E_S - E_B) / E_B \times 100\% \quad (17)$$

式中：

R_E——效率提升率；

E_S——被评价系统的作业效率；

E_B——基准作业方式的作业效率。

作业效率宜采用单位时间内有效完成的任务量或作业量表示。

8.7.3 作业质量提升率

采用数值越大表示质量越高的指标时，作业质量提升率按式（18）计算：

$$R_Q = (Q_S - Q_B) / Q_B \times 100\% \quad (18)$$

式中：

R_Q——作业质量提升率；

Q_S——被评价系统的作业质量指标值；

Q_B——基准作业方式的作业质量指标值。

采用缺陷率、差错率等数值越小表示质量越高的指标时，应计算缺陷或差错降低率，并在评价结果中说明计算方向。

8.7.4 人工投入降低率

人工投入降低率按式（19）计算：

$$R_H=（H_B-H_S）/H_B\times100\% \quad （19）$$

式中：

R_H——人工投入降低率；

H_B——基准作业方式的人工投入；

H_S——采用被评价系统后的人工投入。

人工投入应包括直接操作、任务监控、人工介入、故障恢复和日常维护等相关投入，并采用一致的人员数量、工时或人时口径。

8.7.5 高风险作业暴露降低率

高风险作业暴露降低率按式（20）计算：

$$R_R=（R_B-R_S）/R_B\times100\% \quad （20）$$

式中：

R_R——高风险作业暴露降低率；

R_B——基准作业方式下的高风险暴露量；

R_S——采用被评价系统后的高风险暴露量。

高风险暴露量可采用人员进入次数、暴露时间、人时或其他适用参数表示，统计口径应保持一致。

8.7.6 成本指标

单位任务成本按式（21）计算：

$$C_U=C_T/N_S \quad （21）$$

式中：

C_U——单位任务成本；

C_T——规定统计周期和评价边界内的相关成本；

N_S——统计周期内有效完成的任务量或作业量。

运行维护成本、投资回收期和全生命周期成本应根据评价目的选择。计算时应明确统计周期、成本构成、折旧方式、设备利用率、人工成本和其他主要假设，不同系统进行比较时应采用一致的计算边界和参数口径。

8.8 评价数据有效性与统计处理

无效测试仅限于测试设施故障、外部供电或网络异常、操作人员违反评价方案或者其他非被评价系统原因造成且足以影响结果的情况。

无效测试的原因和证据应记录，并补充等量测试实例。因被评价系统任务失败、系统故障、性能不达标或者发生规定外人工介入、人工接管形成的测试结果，不得作为无效测试剔除。

评价数据的统计处理应符合下列要求：

a）比例类指标应给出有效样本数、符合条件的样本数和百分比；

b）连续量指标应给出样本量以及均值或中位数，宜同时给出标准差、极差或规定分位数；

c）持续运行类指标应给出实际运行时间、累计任务数量或累计作业量；

d）发生次数较少但后果严重的事件，应逐项记录，不宜仅以平均值表示；

e）异常值识别和处理规则应在评价实施前规定，异常值的保留、修正或剔除应具有可追溯依据；

f）对数值越大越好的指标，实测值不低于门槛时判定为达到要求；对数值越小越好的指标，实测值不高于门槛时判定为达到要求；对区间型指标，实测值应处于规定范围内；

g）评价数据、计算过程和判定依据应可追溯，并能够支持评价结果复核。

评价数据有效且达到最小样本量后，方可用于应用成熟度L等级判定。

9 评价与结果表示

9.1 L等级判定规则

L等级应依据第7章规定的能力要求、按照第8章获得的评价结果以及评价方案中预先确定的指标门槛判定。

L1至L5等级均应对任务执行能力、自主作业能力、稳定可靠能力、安全协同能力以及应用效能与经济性进行适用性分析。能力维度下的具体指标应按照下列规则确定属性：

a）必选指标用于L等级判定，达到相应门槛后方可判定满足该项要求；

b）可选指标用于补充反映系统能力、产品比较或综合评分，不直接决定L等级，不得补偿未达到门槛的必选指标；

c）不适用指标不纳入L等级判定或综合评分，但应说明不适用的理由，并在评价实施前由委托方与评价实施机构确认。

某一能力维度被确定为必选，不表示该维度下所有指标均为必选。评价方案应根据场景任务特点选择能够充分反映该维度要求的必选指标。安全协同能力中的适用安全要求不得整体排除，安全否决条件应适用于所有L等级。

表 3 L 等级五类能力维度判定要求

等级	等级定位	五类能力维度要求
L1	演示验证	任务执行：必选，核心任务形成有效结果。 自主作业：可选，记录人工介入和人工接管情况。 稳定可靠：可选，记录故障、中断和恢复情况。 安全协同：适用要求必选，不得发生安全否决事件。 应用效能与经济性：可选，宜明确评价边界并记录基础数据。
L2	受控作业	任务执行：必选，重复完成任务，主要任务性能达到基本门槛。 自主作业：可选；评价方案有自主程度要求时应列为必选。 稳定可靠：必选，重复运行的故障发生情况和主要性能一致性达到基本要求。 安全协同：适用要求必选，满足受控作业条件下的安全要求。 应用效能与经济性：可选，宜取得与基准作业方式比较所需的数据。
L3	场景适用	任务执行：必选，主要任务性能达到实际场景要求。 自主作业：必选，自主完成、人工介入和人工接管达到规定门槛。 稳定可靠：必选，在规定任务数量或运行时间内，故障发生情况和主要任务性能一致性达到基本门槛。 安全协同：适用要求必选，满足实际场景中的安全协同要求。 应用效能与经济性：可选；涉及效能目标、产品选型或项目验收时应按评价方案列为必选。
L4	稳定作业	任务执行：必选，持续达到规定任务性能要求。 自主作业：必选，在规定异常条件下保持自主作业或自主恢复。 稳定可靠：必选，持续运行、系统可用率、故障、恢复和性能一致性达到规定门槛。 安全协同：适用要求必选，持续运行过程中不得发生安全否决事件。 应用效能与经济性：可选；涉及运行效能或验收目标时应按评价方案列为必选。
L5	规模部署	任务执行：必选，满足常态化或规模化作业要求。 自主作业：必选，满足规模部署所需的自主程度和异常处置要求。 稳定可靠：必选，满足长周期或规模化稳定运行要求。 安全协同：适用要求必选，满足规模部署条件下的安全协同要求。 应用效能与经济性：适用指标必选，达到常态化或规模化部署目标。

L等级采用逐级累积判定。判定某一等级时，应同时满足该等级及较低等级的全部必选要求。

必选指标均达到门槛时方可判定相应等级。达到多个等级要求时，取满足条件的最高等级。未达到L1要求时，不给出L等级，可表述为“未定级”。

评价方案规定安全否决条件的，发生否决事件时应判为不通过或未定级；涉及法定或强制性安全要求的，不符合时不得给出通过性结论。

某项必选指标因测试条件不足无法获得时，不得直接判定依赖该指标的等级；确需采用替代证据时，应在评价报告中说明依据和适用限制。

L等级仅对评价报告所载被评价系统版本、场景任务、C等级和评价边界有效。系统或评价边界发生实质变化时，应重新评价。

9.2 C-L 评价结果表示

基本评价结果应表示为“C等级-L等级”。例如，C3-L4表示被评价系统在C3场景挑战度条件下达到L4应用成熟度。C和L不得脱离场景任务、系统版本及评价方案单独使用。

结果表示宜同时附场景任务名称、被评价系统及版本、评价方案编号、评价时间、关键指标摘要和评价边界。跨系统或跨批次比较时，场景任务、C等级、评价方案、指标门槛和数据处理规则应一致。

9.3 综合评分

需要在相同场景任务和C-L判定条件下进行产品选型或方案比较时，可计算综合评分S。综合评分不作为L等级判定依据。

综合评分的指标选择、权重设置、指标得分换算、汇总计算及结果表示应按照附录D执行。

9.4 评价报告

评价报告至少应包括：

a) 被评价系统组成、主要配置、软件及模型版本；

b) 评价目的、评价边界和评价方案标识；

c) 场景任务、测试条件、C等级及其确定依据；

d) 拟判定的L等级，评价指标及其必选、可选或不适用属性，指标门槛、安全否决条件和数据处理规则；

e) 有效样本量、持续运行量和原始结果摘要；

f) 关键指标结果、L等级判定和C-L评价结果；

g) 综合评分及评分方案编号（如适用）；

h) 无效测试、偏离项、替代证据和评价限制；

i) 评价结论的适用范围和复评触发条件；

j) 原始数据、日志、影像及其他证据的索引。

附 录 A
（规范性）
场景挑战度确定方法

A.1 挑战因素

场景挑战度应从作业环境E、作业对象O、动态交互I、任务过程T、异常与扰动R五类因素确定。各因素按照表A.1判定为1级至5级。与任务无关的因素可标记为“不适用”，并说明理由。

表 A.1 场景挑战因素分级判据

因素	C1 固定条件级	C2 受控变化级	C3 典型实景级	C4 复杂动态级	C5 开放挑战级
作业环境E	布局、地面、光照等条件固定、可控	单一环境变量在规定范围内变化	多种典型实景变量变化，需适应环境状态	多变量明显或动态变化，需实时调整	开放环境，存在未完全预定义的条件变化
作业对象O	对象种类、位置、姿态及属性固定	对象单项属性按已知集合或范围变化	类型、位置、姿态等多项典型变化	对象组合复杂，存在形变、遮挡或状态变化	对象类别或状态开放，无法完全预定义
动态交互I	无动态主体或完全隔离	少量主体按规定路径或规则活动	存在典型人员、车辆或设备动态交互	多主体并发，行为具有明显不确定性	开放人机环境，主体和行为不可完全预知
任务过程T	步骤、路径和参数固定	部分步骤或参数在规定集合内变化	需依据任务状态选择路径、顺序或参数	多任务并发或流程频繁变化，需动态规划	目标或流程仅部分定义，需在线分解和规划
异常与扰动R	不设置异常或扰动	少量可预期、易恢复的扰动	设置实际业务中常见异常，要求识别或处置	多类异常可能组合出现，要求自主恢复	存在未完全预定义异常，要求在线判断和调整

A.2 C等级确定

C等级应按照各挑战因素的分级结果确定，步骤如下：

a) 确定五类挑战因素的适用性，并说明不适用理由；

b) 根据表A.1分别判定各适用因素的等级。等级判定应以与规定场景任务直接相关的条件及其形成的实际挑战为依据；

c) 在全部适用因素中取最高因素等级，作为场景挑战度C；

d) 记录判定依据，并核验决定C等级的关键挑战条件得到有效测试覆盖，且实际测试条件符合评价边界。

$$C = \max(c_j),\ j \in J \qquad (A.1)$$

式中：

c_j——第j个适用因素的等级值；

J——适用于规定场景任务的因素集合。

注：某任务的因素等级为E2、O3、I2、T3、R4，五类因素均适用，则C＝max（2，3，2，3，4）＝4，场景挑战度为C4。若R不适用且有充分依据，则C＝max（2，3，2，3）＝3，场景挑战度为C3。

附 录 B
（资料性）
典型应用场景参考

表 B.1 典型应用场景与推荐指标

领域	典型任务	重点挑战因素	推荐指标
工业制造	上料、装配、分拣、检测、设备操作	对象姿态与公差、工位节拍、人机协作、异常料件	任务完成率、质量达标率、操作精度、节拍达标率、自主完成率、可用率、人工投入降低率
仓储物流	混合拣选、搬运、码垛、盘点、配送	货品种类与摆放、通道变化、动态车辆、订单变化	拣选成功率、搬运完成率、任务时间、接管率、碰撞避免成功率、吞吐量
商业服务	迎宾导引、递送、清洁、补货	人流变化、临时障碍、语音交互、任务插单	任务完成率、自主完成率、响应时间、安全事件、服务时长、人工介入频度
医疗康养	物资递送、送药辅助、巡房、助行辅助	人员安全、流程约束、对象识别、隐私与应急	正确交付率、任务完成时间、接管率、安全响应、可用率、异常处置成功率
家庭服务	清洁、整理、取放、家电操作	非结构化布局、多类别物品、家庭成员和宠物	任务完成率、物品损伤事件、自主完成率、介入频度、覆盖率、单位任务时间
特种作业	巡检、采样、远程操作、危险区域处置	恶劣环境、通信波动、高风险对象、未知障碍	任务完成率、覆盖率、通信中断恢复、安全状态保持、故障率、高风险暴露降低率

注：推荐指标仅供评价方案编制参考，实际选择应基于任务目标、场景风险和评价用途。

附 录 C
（资料性）
典型评价方案参考

C.1 通用方案结构

评价方案宜按照“定边界—定任务—定C—选指标—设实例—实施测试—计算指标—判定L—表示结果”的顺序编制。每个测试实例应使用唯一编号，并记录场景变量、任务结果、人工介入、异常、故障和安全事件。

表 C.1　评价方案要素

要素	应明确的内容
评价对象	系统组成、软硬件及模型版本、外部依赖
场景任务	目标、起止状态、作业单元、成功和失败判据
场景挑战度	E、O、I、T、R分级、判定证据和综合C等级
指标属性与门槛	拟判定的L等级，必选、可选和不适用指标，各等级指标门槛及安全否决条件
实例与样本	变量组合、生成方式、最小有效样本量、持续运行量
测试实施	操作顺序、数据采集、人工介入、停止和恢复规则
统计与判定	有效性、异常值、指标统计、L判定和评分方法
评价报告	C-L结果、关键指标、限制条件、偏离项和证据索引

C.2　工业制造：零件上料与装配

表 C.2　零件上料与装配评价方案示例

项目	示例内容
任务	从料框识别并抓取零件，放入工装并完成规定装配；装配质量满足图样公差。
场景条件及C等级	C3。零件类型、位置和姿态按分层随机变化；人员按规定路线经过；设置错料、轻度遮挡等常见异常。
实例与样本	20种变量组合，每种重复3次，共60次完整任务；另进行不少于4h或规定任务量的持续运行。
主要指标	任务完成率、质量达标率、操作精度、节拍达标率、自主完成率、接管率、异常自主处置成功率、可用率。

项目	示例内容
实施要点	测试前锁定版本；按随机顺序执行实例；同步采集控制日志、工位影像、质量检测和介入记录。
示例门槛	仅用于说明：任务完成率≥90%，质量达标率≥95%，自主完成率≥85%，无严重安全事件。实际门槛应由应用方确定。
结果	按照第9章判定并表示为“C3-Lx”；相同评价方案下比较时可附综合评分。

C.3 仓储物流：混合货品拣选与搬运

表 C.3 混合货品拣选与搬运评价方案示例

项目	示例内容
任务	按订单从货架或料箱拣取货品，放入目标容器并运送至指定工位。
场景条件及C等级	C3。覆盖不同包装、尺寸、位置和遮挡；通道存在人员或车辆动态活动；设置目标缺失和临时障碍。
实例与样本	不少于20个订单实例，覆盖不少于100件货品；持续运行量按目标班次或吞吐量要求确定。
主要指标	正确拣选率、完整任务完成率、任务时间、单位时间吞吐量、自主完成率、接管率、碰撞避免成功率、可用率。
实施要点	订单和货位组合在测试开始后公布；失败货品不得替换后继续计为成功；临时障碍按方案规定时点设置。
示例门槛	仅用于说明：正确拣选率≥97%，完整任务完成率≥90%，接管率≤5%，碰撞事件为0。实际门槛应由应用方确定。
结果	报告货品覆盖、订单覆盖、失败类型和C-L结果；评分时明确吞吐量与质量权重。

C.4 商业服务：动态公共区域递送

表 C.4 动态公共区域递送评价方案示例

项目	示例内容
任务	接收递送指令，规划路径，通过公共区域到达目标位置，确认接收并返回或继续下一任务。
场景条件及C等级	C3或C4。人流、临时障碍和任务插单按规定范围变化；C4可增加多主体并发和组合异常。
实例与样本	不少于30次完整递送，覆盖高低人流时段、不同起终点和至少3类规定异常。
主要指标	任务完成率、准时率、自主完成率、介入频度、异常自主处置成功率、安全停止响应时间、人员最小间距、可用率。
实施要点	设现场安全员但不参与正常决策；所有提示和接管均记录；安全测试与正常服务测试分开组织。
示例门槛	仅用于说明：任务完成率≥90%，自主完成率≥85%，规定异常处置成功率≥80%，安全否决事件为0。实际门槛应由应用方确定。
结果	分别报告正常实例和异常实例结果，给出系统版本、C等级、L等级及适用限制。

附 录 D
（资料性）
综合评分计算规则

D.1 总则

综合评分用于在场景任务、C等级、评价方案和L等级判定规则一致的条件下，对被评价系统的表现进行进一步量化比较。综合评分不代替L等级判定，也不得补偿未达到门槛的必选指标。

发生安全否决事件或缺少评分所必需的有效数据时，不应给出可用于比较的综合评分。

D.2 指标与权重

评分指标应从第7章规定的五类能力中选择，并与评价方案中的指标保持一致。未纳入评分的必选指标仍应用于L等级判定。

表 D.1 综合评分维度参考权重

能力维度	参考权重W_k	说明
任务执行能力	0.35	反映任务成功、质量、精度、时间或节拍
自主作业能力	0.20	反映自主完成、介入、接管及异常处置
稳定可靠能力	0.20	反映可用率、故障、恢复及性能一致性
安全协同能力	0.15	反映安全检测、响应和协同事件；否决项单独处理
应用效能与经济性	0.10	反映效率、人工投入和成本等应用价值

具体场景可调整维度权重，但应在评价前确定并形成评分方案编号。各维度权重之和应为1。维度内包含多个指标时，第i项指标的最终权重按式（D.1）确定：

$$w_i = W_k \times v_i \qquad (D.1)$$

式中：

W_k——第k个能力维度的权重；

v_i——指标i在所属维度内的权重，同一维度内 $\Sigma v_i = 1$；

w_i——指标i的最终权重， $\Sigma w_i = 1$。

D.3 指标得分换算

各指标应统一换算为0分至100分。0分锚点和100分锚点应依据应用要求、合同门槛、适用标准或稳定基准确定，不应以本批被评价系统的最小值和最大值临时确定。

表 D.2　指标得分换算规则

指标类型	换算规则	适用示例
正向指标	数值越大越好。按式（D.2）线性换算，并将结果限制在0分至100分。	任务完成率、自主完成率、可用率
负向指标	数值越小越好。按式（D.3）线性换算，并将结果限制在0分至100分。	人工接管率、故障发生率、任务时间、成本
目标区间型指标	处于目标区间内计100分；区间外按预先规定的分段函数递减，达到边界值计0分。	操作力、温度、距离
二值或事件型指标	满足要求计100分，不满足计0分；安全否决事件按照D.5处理。	安全状态保持、规定功能是否实现
扣分型指标	以100分为基准，按照预先规定的事件等级和次数扣分，最低为0分。	轻微服务差错、非否决性告警

正向指标按式（D.2）计算：

$$s_i = clip[100 \times (x_i - x_i,0)/(x_i,100 - x_i,0)] \quad (D.2)$$

负向指标按式（D.3）计算：

$$s_i = clip[100 \times (x_i,0 - x_i)/(x_i,0 - x_i,100)] \quad (D.3)$$

式中：

x_i——指标实测值；

x_i,0——0分锚点；

x_i,100——100分锚点；

clip（y）＝min[100，max（0，y）]。

正向指标应满足x_i,100＞x_i,0，负向指标应满足x_i,0＞x_i,100。

D.4　综合评分计算

综合评分按式（D.4）计算：

$$S = \Sigma\ (w_i \times s_i) \quad (D.4)$$

式中：

S——综合评分，取值为0分至100分；

w_i——指标i的最终权重；

s_i——指标i换算后的得分。

各指标加权得分的中间计算结果保留2位小数，综合评分保留1位小数，并符合下列要求：

a）某项指标不适用时，应在评价前调整评分指标和权重，使适用指标权重之和重新归一化为1；

b）某项指标因测试失败、数据缺失或样本量不足无法计算时，不得直接删除该指标或临时调整权重，应补充测试或不出具综合评分；

c) 同一指标存在多个统计量时，应在评分方案中明确用于计分的统计量，不得在获得结果后择优选取；

d) 不同场景任务、C等级、评分方案、系统版本或数据处理规则形成的分数不应直接比较。

D.5 安全否决与结果表示

安全否决项不参与加权补偿。发生评价方案规定的安全否决事件时，应停止综合评分的通过性解释，评价结果标注为“不通过”或“未定级”；确需保留过程分数时，应明确标注“仅供分析，不作为比较结论”。

综合评分结果应与C-L结果、评分方案编号和关键权重同时表示。例如：“C3-L4，综合评分91.2（评分方案P1）”。

D.6 计算示例

表D.3给出了综合评分的计算示例。指标得分均为按照D.3换算后的得分。

表 D.3 综合评分计算示例

能力维度	指标	最终权重w_i	指标得分s_i	加权得分w_i×s_i
任务执行能力	任务完成率	0.20	92	18.40
任务执行能力	作业质量达标率	0.15	88	13.20
自主作业能力	自主完成率	0.12	85	10.20
自主作业能力	人工接管率	0.08	95	7.60
稳定可靠能力	系统可用率	0.12	90	10.80
稳定可靠能力	性能一致性	0.08	100	8.00
安全协同能力	安全协同指标	0.15	100	15.00
应用效能与经济性	效率提升率	0.10	80	8.00
合计	—	1.00	—	91.2

注：本示例仅说明计算过程，不构成通用权重或合格门槛。

附 录 E
（资料性）
安全否决条件参考

E.1 总则

本附录给出具身智能系统实景应用评价中安全否决条件的参考情形。评价方案宜根据场景任务、作业风险及适用的安全要求，确定安全否决条件及其触发判据。

E.2 典型安全否决条件

表E.1给出了典型安全否决条件参考。

表 E.1 典型安全否决条件参考

类别	典型安全否决情形
人员安全	因系统原因造成人员严重伤害，或发生可能造成严重伤害且安全防护措施失效的危险事件
安全功能	急停、安全停止、防护联锁等规定的关键安全功能失效
危险行为	系统发生失控、非预期危险动作或超出规定安全边界运行，造成不可接受的安全风险
风险检测与避险	对规定的重大危险状态发生漏检，或未能采取必要的避让、停止等安全措施，造成不可接受的安全风险
故障安全	系统在规定故障或异常条件下，未能进入或保持要求的安全状态
设备与环境安全	因系统原因发生严重危险碰撞、重大设备损伤或环境安全事件
行业安全要求	不符合适用的法律法规及强制性安全要求，导致系统不满足规定作业安全条件
类别	典型安全否决情形
人员安全	因系统原因造成人员严重伤害，或发生可能造成严重伤害且安全防护措施失效的危险事件

E.3 使用说明

a) 安全否决条件宜根据场景风险选择，并在评价实施前明确触发判据、判定方法和记录要求；
b) 一般任务失败、可恢复故障、正常安全停止及规定的人工接管，不宜直接作为安全否决条件；

c）对于特定行业的高风险作业，可补充相应安全否决条件，涉及法定或强制性安全要求的，不得通过评价方案予以豁免；

d）发生安全否决事件时，按照第9章规定进行判定，不得通过其他指标或综合评分予以补偿。