**《****具身智能语料库建设导则》**

**（征求意见稿）编制说明**

《具身智能语料库建设导则》标准编制组

2025年3月24日

# 一、标准编制背景

制定《具身智能语料库建设导则》至关重要，一是能够为具身智能领域提供一个统一的标准和规范，确保语料库的建设和使用过程中的数据质量、一致性和可比性；二是有助于提高语料库的可用性和可靠性，减少因标准不一而导致的研究误差和资源浪费，为学术研究和工业应用提供坚实的基础；三是通过提供高质量的语料库，可以加速算法的迭代和优化，促进新技术的诞生和应用，从而推动整个行业的进步；四是有助于不同领域专家之间的合作，促进知识的共享和交流。对该导则的修订重点结合当前具身智能行业的新发展，充分考虑具身智能领域数据采集的多样性和数据质量的有效性，特别是对模型训练的可验证性等，制定具身智能语料数据标准。

本项目首次发布并实施于2024年11月16日。然而，随着行业的快速发展，实施过程中发现部分指标和内容存在缺失，已无法完全满足当前行业发展的需求。因此，于2025年3月初提出了标准修订申请，并获得归口单位的批准立项。

# 二、 编制过程

本标准的修订主要包括以下几个阶段：

**（一）立项准备阶段（2025年2月）**

牵头起草单位组织专家组对具身智能语料库的建设现状进行国内外调研，包括相关政策制度、技术路径、市场现状进行深入研究，梳理团体标准相关法律法规在标准审查中的符合性条文，开展具身智能语料库建设的技术要求研究，为后续标准制定提供理论基础和政策指导。

**（二）立项申请阶段（2025年3月）**

面向社会征集标准编写参与单位，组成标准工作组。工作组成员包括但不限于：上海库帕思科技有限公司，国家地方共建人形机器人创新中心，上海电器科学研究所（集团）有限公司，上海傅利叶智能科技有限公司，上海华为技术有限公司，上海穹彻智能科技有限公司，上海智元新创技术有限公司，松应科技有限公司，星海图（苏州）人工智能科技有限公司，中国电子科技集团公司第二十一研究所等。

**（三）标准起草（2025年3月）**

对具身智能语料库建设需求进行资料收集、查阅、分析和整理，对团体标准政策制度进行梳理，查找标准技术审查关键问题，确定标准主要技术内容和标准框架结构。依据GB/T 1.1-2020起草规范编写标准草案。

**（四）征求意见阶段（2025年3-4月）**

项目组组织了多轮内部讨论会，以专家学者、业务骨干座谈会以及企业调研等形式进行深入调研和讨论，对标准文本进行修改完善，处理内部意见 100 余条，形成《具身智能语料库建设》团体标准征求意见稿及编制说明。

# 三、 编制原则

本标准编制符合以下原则：

**（一）规范性原则**

本文件符合国家和本市现行法律、法规和规范性文件；并符合GB/T 1.1-2020 的起草要求。

**（二）前瞻性原则**

一是在着眼于可信AI未来的发展趋势和创新方向。预见人工智能技术变化，及时调整标准内容和要求，保证标准的长期有效性和适应性。二是在充分研究学术前沿的顶级论文、参考国内外相关标准和文献报告的基础上，提炼总结，进行了鲁棒性的测试指标设定，量化了具身智能语料库建设导则。

**（三）可操作性原则**

本标准提出的具身智能语料库建设立足实际，考虑具身智能数据内容、数据采集方式、数据生产方法和数据安全等方面技术性指导，具有较强的实用性和可落地性。

# 四、 范围

本文件提供了建设具身智能模型训练数据内容、数据采集、语料生产（质量评估、数据有效性评估）和数据安全方面的技术指导方法。本文件适用于具身智能语料库的研究、开发、维护、应用、评估等工作。其它与具身智能语料库建设相关的工作也可参照使用。

**（一）术语和定义**

定义了具身智能、本体、数据资源、数据片段、数据时刻、模拟仿真平台、合成数据相关定义和说明。

**（二）具身智能数据**

本章规定了具身智能语料库与作为语料库素材的数据资源的规范性说明。这些说明将帮助这些数据的使用方（本文件中的用户）了解数据的特性、质量和可用性，以对数据的应用价值作出初步评估。

**（三）具身智能数据采集**

本章规定了对语料库素材的数据资源的采集的规范性说明。内容包括真实数据采集和合成数据采集。为数据采集方的采集介质、采集行为等提供指导性意见，以对数据进行统一。

**（四）具身智能语料生产**

本章提供了具身智能语料生产前对采集数据进行完整性检查、数据资源清洗、数据资源标注等技术性指导。

**（五）数据安全**

语料数据的建设方宜在组织、制度、人员、平台等方面实施安全策略、执行覆盖资源数据和语料产品全生命周期的安全控制措施。

**（六）附录A 具身智能大模型方案**

给出了具身智能大模型方案的示例，供实际执行时参考。

**（七）附录B 具身智能语料库数据资源类型**

给出了具身智能各类型数据资源的推荐指标、语义分类、语义特征等，供实际执行时参考。

**（八）附录C 具身智能遥操作模式真实数据采集本体参考性能参数**

给出了具身智能遥操作式数据采集设备供参考的性能参数、适配标准及技术参数，供实际执行时参考。

**（九）附录D “生产伴随”式数据采集设备参考性能参数**

给出了具身智能生产伴随式数据采集设备供参考的性能参数、适配标准及技术参数，供实际执行时参考。

**（十）附录E 具身智能语料库数据有效性评估数据集要求**

给出了开展具身智能数据有效性评估时，所需数据集的选型准备和推荐质量要求，供实际执行时参考。

# 五、 重大分歧意见的处理结果及理由

本标准在修订过程中无重大分歧意见。

# 六、 标准作为强制性或推荐性标准发布的意见

推荐性。

# 七、 推动标准实施的措施建议

本标准发布后将尽快组织宣贯，加大贯彻实施力度。第一，在适用主体中推广应用该标准，开展测试评估，形成经验；第二，根据试点经验，推动标准更新实践，复制推广试点经验和标准适用主体范围扩大；第三，广泛收集意见和建议，及时归纳和总结，并不断完善标准，必要时提出标准修订。

# 八、 其他应予以说明的事项

无。