

消防语料库建设导则

(征求意见稿)

编制说明

2025年09月

《消防语料库建设导则》

团体标准征求意见稿编制说明

1. 工作简况，包括任务来源、协作单位、主要工作过程、团体标准主要起草人及其所做的工作等

1.1. 任务来源

《消防语料库建设导则》由上海市人工智能行业协会提出。

1.2. 协作单位

本标准的起草单位有上海库帕思科技有限公司、应急管理部上海消防研究所、上海脉策数据科技有限公司、中国联合网络通信集团有限公司等。

1.3. 主要工作过程

标准编制期间，其主要工作过程如下：

1) 2025年7月，标准起草单位调研了行业标准现状，根据目前消防行业语料库建设标准缺失的现状，提出了标准起草需求；

2) 2025年7月，标准起草单位结合消防行业语料的特点，起草了标准草案；

3) 2025年8月11日，由上海市人工智能行业协会进行了立项评审，会后根据专家意见对草案进行了修改并提交评审资料，2025年8月，下达标准计划；

4) 2025年8月, 标准牵头单位对外公开征集, 组建了标准起草工作组;

5) 2025年8月29日, 标准牵头单位收集标准草案意见并召开了第一次标准研制会, 讨论了标准草案, 要集中在语料分类、消防知识体系等讨论, 共收集到9条书面意见, 会后标准牵头单位根据收集的意见及现场意见对标准草案进行了标准框架优化与内容修改;

6) 2025年9月8日, 标准牵头单位组织召开第二次标准研制会, 讨论标准草案, 并根据讨论情况对标准进行了修改, 共收集到5条意见;

7) 2025年9月, 标准牵头单位根据研制会的建议对标准进行了修改, 形成征求意见稿。

1.4. 标准主要起草人及其所做的工作

上海库帕思科技有限公司负责了本标准的架构制定工作, 并对标准内容进行初审; 上海库帕思科技有限公司主要负责本标准的全文编制工作并主导了标准的会议讨论和标准的整体审核工作; 应急管理部上海消防研究所、上海脉策数据科技有限公司、中国联合网络通信集团有限公司参加了标准工作组会议, 提出标准内容的具体修改意见。

2. 标准编制原则和确定标准主要内容(如技术指标、参数、公式、性能要求、试验方法、检验规则等)的论据(包括试验、统计数据), 修订标准时, 应增列新旧标准水平的对比;

2.1. 标准制定的原则

标准编制遵循“统一性、适用性、一致性、规范性”的原则，注重标准的可操作性。本标准编写是执行 GB/T1.1-2020《标准化工作导则第1部分：标准化文件的结构和起草规则》。

2.2. 标准的主要内容（增加体现自创新部分的说明）

本文件规定了消防语料库建设的总体要求、语料分类、数据要求、语料生产与管理要求及治理与安全要求。

除了标准的规范性要素（范围、规范性引用文件、术语和定义等）外，本标准的正文部分主要内容包括：

a) 第5章语料库分类：按照业务驱动的特点，将语料库分为管理类、实务类、案例类和学科类；按照模型应用途经，分为预训练语料库、监督微调（SFT）语料库、思维链（CoT）语料库和知识库语料库；

b) 第6章数据要求：规定了语料库入库数据的基本信息、规模、多样性、密级和质量等属性要求，以及数据在提供过程、传输和存储等方面的技术要求；

c) 第7章语料要求：规范了从原始数据到合格语料的完整生产路径，包括采集、清洗、自动化与人工标注、存储、测试、应用和更新等全生命周期的技术规程和管理要求；

d) 第8章治理与安全要求：提出了语料库建设的管理体系、全生命周期治理和可持续运营等治理要求，并从接入安全、传输安全和应用安全等方面对数据全链路的保密性与完整性进行规范；

e) 第9章价值要求：规定语料产品应具备正确的价值导向，符合科技伦理与社会道德规范，保障数据处理符合消防行业的内部管理规定；

f) 附录A: 消防行业学科知识体系: 根据消防行业的学科知识体系, 规范了消防行业的学科知识体系;

g) 附录B: 消防行业业务知识体系: 根据消防行业的业务体系, 规范了消防行业的业务知识体系;

h) 附录C: 消防语料表达示例: 提供消防语料加工后的表达示例。

3. 主要试验（或验证）的分析、综述报告，技术经济论证，预期的经济效果；

3.1. 试验验证分析

无；

3.2. 综述报告

当前智慧消防系统内部数据缺乏统一标准，数据质量参差不齐、表达形式不一，难以满足构建消防语料，实现人工智能训练与智能化应用的需要。具体包括以下几个必要性：

一是填补行业语料标准体系空白。目前消防行业虽已有部分数据元及信息代码标准，但尚缺专门面向 AI 应用和语料资源建设的系统性规范，尤其缺乏对多模态数据（文本、语音、图像、空间数据等）协同表达和结构化标注的要求。

二是应对“AI 原生智慧消防”发展的迫切需求。随着 AI 原生智慧消防平台的推进，行业对高质量、多源异构语料的统一表达和集中管理提出了更高要求，缺乏标准将严重制约模型训练效果与智能体场景的适配能力。

三是构建行业语料共享与安全机制的基础支撑。语料作为关键的战略资源，其规范化建设不仅是行业内部协同的基础，也为跨部门相关数据共享、平台化语料服务和产业化智能应用提供了安全与合规保障。

四是对接国家战略与产业发展趋势。在“人工智能+”上升为国家战略的背景下，消防行业需通过标准化手段，提升语料生产、组织与交付的专业化水平，增强我国在智慧应急和城市公共安全领域的技术话语权。

3.3. 技术论证

在国际上，一些发达国家已建成较为成熟的国家级通用语料库，如英国国家语料库（BNC）和美国国家语料库（ANC），并广泛应用于语言理解、政策研判等领域。部分国家对消防行业的特定语言数据，进行了采集、分析和应用。如葡萄牙的 FireLoc 项目通过分析市民发布的关于森林火灾的社交媒体文本，自动提取火灾地点、严重程度等关键信息，实践自然语言处理（NLP）技术应用应急响应领域。由欧盟的 TEMA 项目，通过人工智能和多源数据融合，提升对灾害演变情况的精确评估。

在我国，通用型语言资源建设已取得阶段性成果，但在面向消防这类高专业性、强实战性的特殊行业，语料资源建设仍处于起步阶段，普遍存在标准缺失、分类不清、结构不一、覆盖不全等问题。近年来，随着“数字中国”、“人工智能+”等战略的推进，我国在消防数据平台和标准体系建设方面积累了一定基础，但从支撑大模型训练和高级 AI 应用的角度看，当前仍缺乏一个覆盖全业务流程、融合多数据模态、满足实战需求的行业统一语料标准。

3.4. 预期的经济效果

本文件以规范提升语料数据的共享性与互操作性。通过统一数据格式、命名规则、元数据结构和质量验收标准，推进跨平台、跨场景的协同应用。

通过本标准的编制，支撑“人工智能+消防”创新发展。为智能指挥、风险研判、战术推演、执法辅助、装备保障等典型场景提供结构完备、质量可控、可供机器学习的语料基础，加快人工智能技术在消防行业的规模化落地。

4. 采用国际标准和国外先进标准的程度，以及与国际、国外同类标准水平的对比情况，或与测试的国外样品、样机的有关数据对比情况；

与现有标准相比，本文件的突出贡献在于首次系统性地提出了消防行业语料体系建设导则，促进消防垂类大模型训练数据规范化、提升模型的理解能力和专业业务能力具有重要意义，填补了标准领域的空白。

5. 与有关的现行法律、法规和强制性国家标准的关系；

与现行相关法律、法规、规章及相关标准无矛盾。

6. 重大分歧意见的处理经过和依据；

无。

7. 标准的建议；

建议作为团体标准实施，建议标准发布后立即实施。

8. 贯彻标准的要求和措施建议（包括组织措施、技术措施、过渡办法等内容）；

本标准发布后，标准起草组将作为标准应用推广的主体，组织科研院所、相关企业、第三方检测认证机构等进行标准的宣贯。同时，起草组对标准的核心内容进行解读，方便后续的应用。

9. 废止现行有关标准的建议；

无，本标准为首次制定标准。

10. 其他应予说明的事项。

无。

起草工作组
2025年09月