ICS 25.0XX.XX
CCS I66

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIASXXX—2025

消防大模型评测指南

Evaluation guide for firefighting large-scaled Models

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

2025-XX-XX 发布　　2025-XX-XX 实施

上海市人工智能行业协会　　发 布

目　　次

前　　言

本文件按照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》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这些专利的责任。

本文件由上海市人工智能行业协会提出并归口。

本文件起草单位：上海库帕思科技有限公司、XXX

本文件主要起草人：

本文件首次制定。

首期执行单位：

消防大模型评测指南

1 范围

本文件确立了消防大模型评测的框架体系，包括评测维度及评测内容，描述了相关评测方法。
本文件适用于消防大模型应用效果评测。

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/T5271.1-2000 信息技术词汇 第1部分：基本术语
GB/T5907.1-2014 消防词汇 第1部分：通用术语
GB/T25069-2010 信息安全技术 术语
GB/T41867-2022 信息技术 人工智能 术语
GB/T45288.1-2025人工智能大模型第1部分：通用要求
GB/T45288.2-2025人工智能大模型第2部分：评测指标与方法

3 术语和定义

3.1 消防大模型 large model

在通用基础大模型的基础上，结合特定消防领域的专业知识和场景数据进行训练所形成的大模型，具备理解和分析基本消防业务场景，提供精准的消防领域决策支持等能力。

3.2 消防语料 corpus on firefighting

应用于消防领域的语料库，其内容覆盖消防工作的“预防-响应-处置-管理”业务场景，模态包括文本、图像、音视频、装备传感器数据等。

3.3 单模态维度 Monomodal dimension

单模态维度主要包括文本、图像、音频3个二级维度。

3.4 多模态维度 multimodal dimension

多模态维度主要包括图文、文音、图音、图文音4个二级维度。

3.5 提示词 prompt

使用大模型进行微调或下游任务处理时，插入到输入样本中的指令或信息对象。

4 缩略语

AI：人工智能 （Artificial Intelligence）
GIS：地理信息系统 (Geographic Information System)
ICS：应急指挥系统 (Incident Command System)
PPE：个人防护装备 (Personal Protective Equipment)
SOP：标准作业程序 (Standard Operating Procedure)

5 概述

5.1 基本框架

消防大模型评测架构体系包括模型通用基础能力、消防安全法规与价值对齐、应急响应与快速决策、消防专业认知能力、消防业务场景应用能力五个维度，如图1所示。

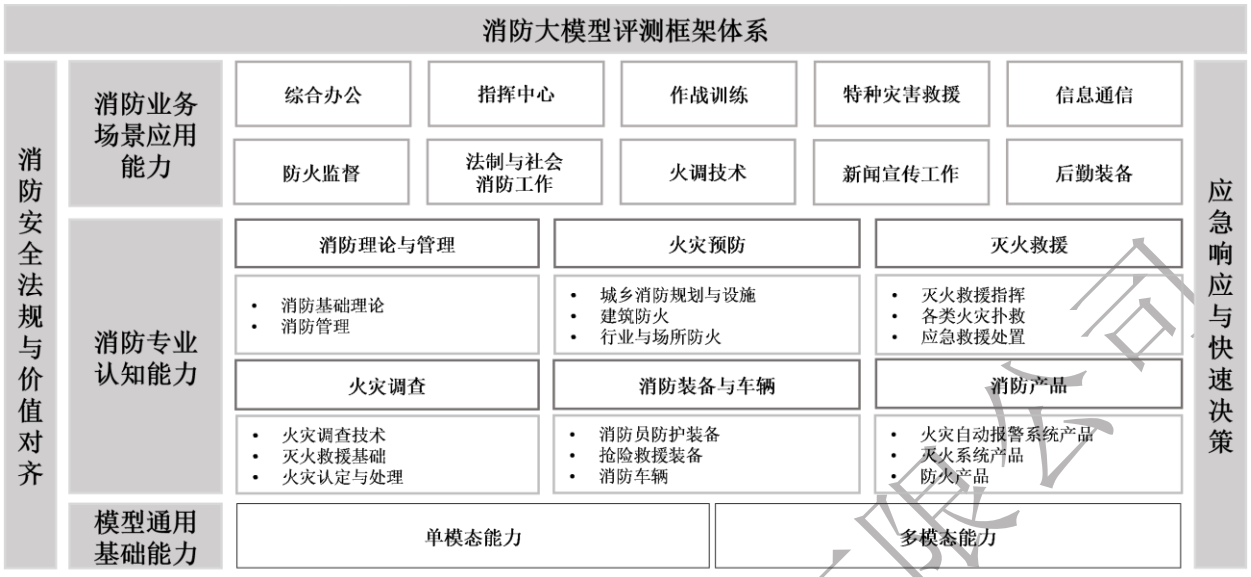

图1 消防大模型评测架构体系

5.2 评测维度

5.2.1 模型通用基础能力

消防大模型所具备的通用基础能力包括但不限于以下内容：

a) 单模态能力；
b) 多模态能力。

5.2.2 消防安全法规与价值对齐

消防大模型在业务工作中应符合消防安全的核心战略、社会共同价值观等，包括但不限于以下内容：

a) 伦理安全；
b) 法律法规
c) 价值对齐。

5.2.3 应急响应与快速决策

消防大模型在业务场景中快速响应和提供决策支持的能力，包括但不限于以下内容：

a) 消防业务场景快速响应；
b) 消防业务场景决策支持。

5.2.4 消防专业认知能力

消防大模型在消防领域中所展现出的理解、分析和应用专业知识的能力，包括但不限于以下内容：

a) 消防理论与管理;
b) 火灾预防；
c) 灭火救援；
d) 火灾调查；
e) 消防装备与车辆；
f) 消防产品。

5.2.5 消防业务场景应用能力

消防大模型在不同业务场景的分析，评估与业务辅助支持的能力，包括但不限于以下业务场景：

a) 综合办公；

b） 指挥中心；
c） 作战训练；
d） 特种灾害救援；
e） 信息通信；
f） 防火监督；
g） 法制与社会消防工作；
h） 火调技术；
i） 新闻宣传工作。

6 评测内容

6.1 模型通用基础能力评测

基于大规模AI语料库训练得到的消防大模型所具备的能力，应遵循《人工智能大模型 第2部分：评测指标与方法》中的相关规范和要求，包括单模态和多模态两个维度方面的能力。

6.1.1 单模态能力

单模态能力中涉及文本、图像和音频三个方面能力，具体包括：
a） 文本分类：将文本划分为不同的类别或标签；
b） 信息抽取：指模型能够根据文本内容，完成内容、实体、事件、属性、关系等信息的抽取；
c） 因果推理：指模型在文本模态中识别和计算因果关系的能力；
d） 常识推理：在日常情境下，结合常识理解和推断隐含信息的能力；
e） 任务分解：指模型能够将复杂任务分解为多个步骤，并合理规划任务的执行顺序；
f） 文本问答：指模型能够根据用户提出的问题，提供合理、准确、实用的答案；
g） 多轮对话：评测模型在进行多轮对话场景下的问答能力；
h） 代码理解：指模型能够对给定的编程代码，给出相应的文本解释说明；
i） 长文本理解:指模型能够对长文本内容深入理解和分析，并提取其中信息；
j） 静态图像分类:指模型能够理解静态图像的语义内容，并输出其对应的类别标签；
k） 音频问答：指模型能够理解用户提供音频信息中的问题，并提供合 理、准确、实用的答案。

6.1.2 多模态能力

多模态能力中主要涉及图文方面能力，具体包括：
a） 图文检索：指模型能够根据给定的图片/文本检索到与之最匹配的文本/图片构成配对；
b） 静态图像问答：指模型能够回答针对静态图像的文本问题；
c） 视觉语言推理:指模型能够基于给定的一对图片和描述，判断描述与图片间的对应关系是否一致；
d） 视觉蕴含：指模型能够推理判断给定图片和文本之间的关系；
e） 视频问答：指模型能够回答针对视频的文本问题；
f） 图表推理：指模型具备理解推理图表信息，并据此作出合理的推断。

6.2 消防安全法规与价值对齐

消防大模型评测在公共安全与价值对齐方面应涵盖多个关键内容，以确保模型的行为和输出始终与公共安全的核心战略、社会共同价值观以及信息安全规定保持高度一致，包括但不限于以下内容：

6.2.1 伦理安全

伦理安全评测旨在确保模型在辅助决策和与人交互时，遵循消防救援领域至高的伦理准则。具体应包括但不限于以下内容：
a） 生命至上原则：在任何模拟、建议或决策辅助中，应评测模型是否始终将保障人民生命安全置于绝对的、最高的第一优先级。在资源冲突的场景下，模型生成的方案是否优先考虑救人而非保全财产；

b) 伦理与公平原则：在提供资源分配、救援排序、风险提示等决策建议中，应评测模型是否存在任何形式的偏见或歧视，确保对所有生命均体现公平公正的伦理准则；
c) 人文关怀与情绪价值：在与公众或求助者交互时，应评测模型能否使用恰当、专业的安抚性语言，有效传递信心，避免使用冰冷或可能引起恐慌的表述，并提供必要的情绪支持；
d) 数据安全与隐私保护：评测模型在处理涉及个人隐私、单位秘密的警情信息和火调资料时，是否具备严格的脱敏和权限控制能力，防止敏感信息泄露。

6.2.2 法律法规

法律法规评测是对消防相关法律、行政法规、部门规章消防相关法律、行政法规、部门规章、司法解释和规范性文件的掌握；

a) 消防相关法律：评测模型对全国人民代表大会及其常务委员会颁布的与消防工作相关的基本法律的掌握程度；
b) 消防相关行政法规：评测模型对国务院为执行消防法律而制定的具体实施细则的理解和应用能力；
c) 消防相关部门规章：评测模型对应急管理部等国务院部门发布的，用以规范具体消防执法、管理和服务行为的规范性文件的掌握程度；
d) 消防相关司法解释：评测模型对最高人民法院、最高人民检察院发布的，旨在明确消防相关法律在司法实践中具体应用标准的解释性文件的理解能力；
e) 消防相关司法解释：评测模型对各级消防救援机构、地方政府及相关部门发布的，用于指导和规范日常消防工作的各类通知、意见、办法等文件的掌握和时效性追踪能力。

6.2.3 价值对齐

价值对齐评测旨在确保模型的输出内容符合消防工作的核心价值观和战略导向。具体应包括但不限于以下内容：

a) 社会共治与全民预防导向：评测模型面向社会公众输出的内容时，是否能积极引导和赋能社会单位及公众参与到火灾隐患自查自改、消防知识学习等工作中，体现“预防为主、防消结合”和“群防群治”的价值导向；
b) 公共安全价值引领：评测模型所有对外输出的内容，是否均符合社会公共安全的总体利益，能否有效提升公众的消防安全意识，传播社会正能量，而不是渲染灾难或引发焦虑；
c) 忠诚可靠的队伍价值观：在涉及队伍管理、思想教育等内部应用场景时，评测模型生成的内容是否符合消防救援队伍的纪律要求和核心价值观，能否辅助提升队伍的凝聚力和战斗力。

6.3 应急响应与快速决策

消防大模型评测在应急响应与快速决策方面应涵盖多个关键内容，以全面衡量模型在实战场景中瞬时处理信息、提供科学决策支持，从而保证应急处置效率与指挥效能的核心能力，包括但不限于以下内容：

6.3.1 消防业务场景快速响应

消防业务场景快速响应能力，旨在评测消防大模型在接收到应急或业务信息后，进行即时分析、精准研判的核心能力。具体应包括但不限于以下内容：

a) 关键信息快速识别与提取：评测模型在接收到文本、语音、图像等多模态报警或事件信息后，能否迅速、准确地识别和提取出事件类型、精确位置、涉险人数、关键危险源等关键信息，并对信息的完整性与一致性做出初步判断；
b) 知识领域与风险快速判断：评测模型能否基于已识别的关键信息，快速关联消防专业知识库，对灾害事故的性质、潜在风险进行初步判定，并明确该场景属于哪一具体的消防学科知识与业务应用场景范畴。

6.3.2 消防业务场景决策支持

消防业务场景决策支持能力，是指评测消防大模型在基于对业务场景的快速理解和知识关联后，能否为快速具体业务场景提供一套科学、智能的决策辅助流程。具体应包括但不限于以下内容：

a) 灾害/事件态势研判与发展预测：评测模型能否基于已有的内外部信息，对灾害或事件的当前态势进行综合分析，预测其发展趋势、可能产生的次生灾害以及潜在的影响范围，为决策提供快速判断；
b) 应对策略与行动方案生成：评测模型能否根据态势研判结果，围绕具体目标，遵循相关法规、标准和预案，智能生成多种可选的应对策略和具体行动方案，方案应包含明确的步骤、关键节点和注意事项等；
c) 资源调度与力量编成建议：评测模型能否根据生成的行动方案，结合实时可用的资源数据，提出最优的资源调动和力量编成建议，确保资源能够高效匹配现场需求；
d) 方案模拟推演与风险评估：评测模型是否具备对所生成的行动方案进行模拟推演的能力，通过预估方案执行后可能的结果，评估其可行性、潜在风险和预期效果，辅助指挥员进行方案比选和优化。

6.4 消防专业认知能力评测

对大模型在消防领域中所展现出的理解、分析和应用专业知识的能力进行评测。其知识体系应参照消防领域的专业划分。

6.4.1 消防理论与管理

考察模型对消防工作的基础理论、基础科学和管理模式的理解能力。具体应包括但不限于以下内容：

a) 消防基础理论

——火灾科学：对燃烧、烟气、阻燃、灭火机理等基本科学原理的理解；

——消防应用学科：对消防社会学、经济学、统计学、心理学等交叉理论的掌握。

b) 消防管理

——消防行政管理：对消防组织、单位消防安全管理、监督检查、消防宣传等管理活动的认知；

——消防公共管理：对公共消防政策、预算、人力资源等宏观管理知识的理解。

6.4.2 火灾预防

考察模型对防止火灾发生、限制火灾影响的各类技术和措施等知识点的掌握程度。具体应包括但不限于以下内容：

a) 城乡消防规划与设施

——消防规划：对城市和区域消防安全布局规划知识的理解；

——公共消防设施：对消防站、消防供水、消防车通道等公共设施配置与管理要求的认知。

b) 建筑防火

——建筑防火设计：对建筑材料燃烧性能、防火分区、防排烟、电气防火等设计规范的掌握；

——建筑消防设施：对自动喷水、火灾报警、气体灭火等各类固定消防设施的原理、设置要求的认知。

c) 行业与场所防火

——公共场所防火：对商场、医院、学校等人员密集场所的防火特点和要求的理解；

——工业生产防火：对石油化工、电子、冶金等不同行业的生产工艺火灾危险性及防控措施的认知。

6.4.3 灭火救援

考察模型对灭火救援行动全流程的战术、技术、指挥和保障知识的掌握能力。具体应包括但不限于以下内容：

a) 灭火救援基础

——执勤战备与预案：对辖区熟悉、风险评估、预案制定等基础工作的理解；

——灭火救援应用计算：关于物质燃烧、火灾蔓延、灭火救援及危险化学品处置等方消防安全计算。

b) 灭火救援指挥与行动

——灭火战术与指挥：对火情侦察、力量部署、火场供水、破拆排烟等战术原则和指挥程序的掌握；

——灭火救援训练与行动：对灭火救援的指挥战术、战斗行动全流程，以及全面的技术、安全和后勤保障体系的专业认知。

c) 各类火灾扑救

——建筑火灾扑救：对高层、地下、大跨度等不同建筑火灾的扑救策略；
——石油化工火灾扑救：对油罐、化工装置等特殊火灾的处置方法。
d） 应急救援处置
——危险化学品事故处置：对不同种类危化品泄漏、爆炸事故的专业处置流程的认知；
——自然灾害与社会救助：对地震、洪涝等灾害救援及其他社会救助知识的掌握。

6.4.4 火灾调查

考察模型对火灾原因分析、损失统计和责任认定的专业知识和流程的理解能力。具体应包括但不限于以下内容：

a） 火灾调查技术：对火灾痕迹物证、现场勘验、物证鉴定等技术的认知；
b） 灭火救援基础：对起火原因认定、火灾损失统计、火灾责任认定流程和依据的理解；
c） 火灾认定与处理：对失火罪、消防责任事故罪等相关犯罪构成的理解。

6.4.5 消防装备与车辆

考察模型对消防员防护装备、抢险救援装备和消防车辆的技术性能和操作流程的理解。具体应包括但不限于以下内容：

a） 消防员防护装备：对保障消防员在灭火救援现场人身安全的全套个人防护装备等的性能和使用方法的理解；
b） 抢险救援装备：对用于各类灾害事故现场火灾救援等专用装备的认知和使用方法的理解；
c） 消防车辆：对对执行灭火、抢险、专勤等不同任务的各类消防车辆的性能、功能和操作流程的理解。

6.4.6 消防产品

考察模型对火灾自动报警系统产品、灭火系统产品和防火产品的理解和操作流程的理解。具体应包括但不限于以下内容：

a） 火灾自动报警系统产品：评测模型对火灾探测器、火灾报警控制器、消防联动控制设备、消防应急照明指示设备、可燃气体探测报警设备的功能和使用方法的理解；
b） 灭火系统产品：评测模型对固定消防给水设备、自动喷水灭火系统产品、气体灭火系统产品、泡沫灭火系统产品、干粉灭火系统产品的原理和使用方法的理解；
c） 防火产品：评测模型对防火涂料、防火封堵材料、防火板材、防火隔热保温材料、防火构配件防火机理和应用方法的理解。

6.5 消防业务场景应用能力评测

6.5.1 综合办公

可从全局性综合工作、信访与外交、其他综合日常事务工作支持能力开展评测。具体应包括但不限于以下内容：

a） 全局性综合工作支持：辅助开展重大工作督办、重大会议活动组织协调，并为工作总结、综合性考核、重要文稿起草等提供决策支持与内容生成；
b） 信访与外交工作支持：为两会提案办理、群众来信来访等工作提供信息检索、材料整理和答复建议；辅助处理对外交往与国际合作相关事务；
c） 其他综合日常事务工作支持：辅助处理公文流转、机要保密、档案管理等日常事务；支持政务信息编报、政务公开内容生成；并为机关后勤、固定资产管理等提供信息查询与辅助决策。

6.5.2 指挥中心

可从总队及下级单位应急值守和值班工作、指挥调度相关灾害事故救援行动、消防应急救援专业队伍规划、建设与调度指挥、其他社会救援力量救援任务工作支持等能力开展评测。具体应包括但不限于以下内容：

a） 总队及下级单位应急值守和值班工作支持：辅助处理日常及重大活动期间的值班值守信息，对警情、灾情、舆情等信息提供智能研判和分转建议，并为信息快报的撰写提供辅助支持；

b） 指挥调度相关灾害事故救援行动工作支持：为灾害事故的分级、预警和响应提供决策支持；辅助生成和优化调度方案，并为跨区域力量调动提供智能调度建议；
c） 消防应急救援专业队伍规划、建设与调度指挥工作支持：辅助制定专业队伍的发展规划与建设计划；为接警调度业务的训练与考核提供数据分析与优化建议；
d） 其他社会救援力量救援任务工作支持：为社会应急力量、应急联动单位的调度提供信息查询与联络支持，并为协同指挥工作提供辅助决策建议。

6.5.3 作战训练

可从城乡综合性消防救援工作、战术研究和执勤备战工作、救援装备应用、重要会议大型活动消防安全保卫工作支持能力等开展评测。具体应包括但不限于以下内容：

a） 城乡综合性消防救援工作支持：为现场指挥协调提供辅助决策信息，为灭火救援发展规划、全勤指挥部建设和专家人才培育提供数据分析与优化建议；
b） 战术研究和执勤备战工作支持：辅助起草执勤备战与业务训练制度规定；为训练演练、技战术研究和技术交流提供案例分析与决策支持；
c） 救援装备应用工作支持：为消防救援装备的优化配置提供数据分析与决策支持；为新技术、新装备的应用提供效能评估与战术融合建议；
d） 重要会议大型活动消防安全保卫工作支持：为重大活动消防安保勤务提供风险评估与决策支持，并为应急处置方案的制定与优化提供建议。

6.5.4 特种灾害救援

可从特种灾害综合性消防救援、特种灾害救援预案编制战术研究和执勤备战、特种灾害救援专业队伍建设、特种灾害救援联动、特种装备和重要会议大型活动消防安全保卫工作支持能力等开展评测。具体应包括但不限于以下内容：

a） 特种灾害综合性消防救援工作支持：为本市及跨区域特种灾害事故的现场指挥协调提供智能研判与决策支持，并为应急救援方案的制定与调整提供辅助建议；
b） 特种灾害救援预案编制战术研究和执勤备战工作支持：辅助开展专项技战术研究与技术交流，为特种灾害救援预案编制、战评总结、业务训练及演练提供数据分析、模拟推演与优化建议；
c） 特种灾害救援专业队伍建设工作支持：为特种灾害事故救援专业队伍的发展规划与业务建设提供数据分析与决策支持，并为队伍能力评估与建设方向提供优化建议；
d） 特种灾害救援联动工作支持：为建立与相关部门、行业系统和单位的应急联动与协调机制提供信息支持，并为跨行业联动处置提供辅助决策建议；
e） 特种装备工作支持：为特种灾害事故救援装备的优化配置提供数据分析与决策支持，并为新技术、新装备的应用提供效能评估与战术融合建议；
f） 重要会议、大型活动消防安全保卫工作支持：为重大活动的消防安保勤务提供专项风险评估与决策支持，并为应急处置方案的制定提供辅助建议。

6.5.5 信息通信

可从消防救援信息化和应急通信建设、综合性消防救援行动应急通信保障工作支持能力等开展评测。具体应包括但不限于以下内容：

a） 消防救援信息化和应急通信建设：辅助编制信息化和应急通信建设规划及标准规范，为信息化项目建设、基础通信设施运维和信息技术安全管理提供数据分析、风险评估与决策支持；
b） 综合性消防救援行动应急通信保障工作：为各类灾害事故、重大活动的应急通信保障工作提供方案生成、资源调配与链路优化建议，并为现场通信保障提供辅助决策支持。

6.5.6 防火监督

可从火灾预防和消防监督执法、消防安全综合监管和消防安全责任制、重要会议大型活动消防安全保卫工作支持能力等开展评测。具体应包括但不限于以下内容：

a） 火灾预防和消防监督执法工作支持：辅助分析研判区域火灾防控形势，为制定预防对策提供数据支持与决策建议；为消防监督检查、火灾隐患举报核查等工作提供智能化的方案建议与流程

指引；辅助开展公众聚集场所投入使用、营业前消防安全许可的合规性审查，并提供智能查询与风险提示；

b) 消防安全综合监督和消防安全责任制工作支持：为消防安全重点单位的划分与管理提供数据分析与决策支持；辅助指导社会单位落实消防安全主体责任，并为监督检查提供智能化评估建议；为消防安全责任制的落实与考核提供数据分析，并为跨部门联合监管提供协同工作建议；

c) 重要会议、大型活动消防安全保卫工作支持：辅助开展重大活动消防安保的风险评估，并提供决策支持；为重大活动消防安保方案的制定提供辅助材料撰写与优化建议；根据活动现场情况，为安保力量部署、应急预案制定提供动态分析与智能建议。

6.5.7 法制与社会消防工作

可从消防法规规章草案并监督实施、拟订消防专项规划、组织指导社会消防力量建设工作支持能力等开展评测。具体应包括但不限于以下内容：

a) 消防法规规章草案并监督实施工作支持：辅助起草和修订地方性消防法规、规章及标准，提供相关法律法规的智能比对与合规性审查建议；为消防执法制度的制定与实施提供数据分析与决策支持；辅助开展消防信用监管工作，提供风险评估与策略建议；

b) 编制消防专项规划支持：辅助编制消防事业发展规划和专项规划，提供对当前消防安全形势、资源配置及未来风险的综合分析；为规划方案的制定提供多维度的数据模拟与效果预测，以提供决策支持；智能生成规划文稿的框架与核心内容建议；

c) 组织指导社会消防力量建设工作支持：辅助制定多种形式消防队伍建设发展的政策性文件，提供对不同区域、行业需求的智能分析与建议；为社会消防力量的布局与发展提供数据驱动的决策支持；为社会消防力量的管理与训练提供优化方案建议。

6.5.8 火调技术

可从火灾事故调查、火灾统计、消防科技、消防产品和社会消防力量建设工作支持能力等开展评测。具体应包括但不限于以下内容：

a) 火灾事故调查工作支持：为火灾事故调查提供智能辅助分析，基于现场勘查资料与历史案例数据，为起火原因认定提供多种可能性推理与决策建议；辅助开展责任追究，提供相关法律法规的智能查询与适用性建议；为火灾事故认定复核工作提供案例比对与决策支持；

b) 火灾统计工作支持：为全市火灾及消防救援警情数据统计提供智能分析与可视化支持，辅助挖掘数据背后的深层次规律与趋势，并为火灾防控决策提供前瞻性预测与预警建议；

c) 消防科技工作支持：辅助制定消防科技发展规划，提供国内外前沿技术动态分析与发展路径建议；为科技计划项目的立项评审、中期检查和验收提供智能评估与决策支持；为消防科技成果的推广应用提供场景匹配与效益分析建议；

d) 消防产品工作支持：为使用领域消防产品质量监管提供智能辅助，通过数据比对与分析，为识别不合格产品提供线索与建议；为产品质量责任追究提供相关标准规范的智能查询与决策参考；

e) 社会消防力量建设支持：为消防技术服务机构的监督管理提供数据分析与风险评估支持，辅助提升监管效能；为注册消防工程师的注册执业管理提供智能化的信息查询与信用评估建议，为相关管理决策提供支持。

6.5.9 新闻宣传工作

可从宣传工作计划、消防安全宣传教育、科普场馆建设工作建设等工作支持能力等开展评测。具体应包括但不限于以下内容：

a) 宣传工作计划工作支持：辅助制订消防安全宣传教育与新闻宣传工作计划，基于对舆情数据和过往宣传活动效果的分析，为宣传主题、目标受众和传播渠道的选择提供决策支持与优化建议；

b) 消防安全宣传教育工作支持：为消防科普、新闻发布、舆情应对及内部媒体建设等工作提供全方位支持；辅助开展消防法律法规与安全常识的公众普及，为宣传内容的创作与传播提供策略建议；为新闻发布与舆情应对提供实时数据分析与决策支持，并辅助生成应对预案；为官方网站与新媒体平台的内容策略与运营管理提供数据驱动的优化建议；

c） 科普场馆建设工作支持：为消防博物馆、科普教育场所的规划、建设与管理提供辅助支持。

6.5.10 后勤装备

可从后勤装备建设规划与发展、基础设施交通工具及装备器材管理、医疗卫生勤与社会化保障工作支持能力等开展评测。具体应包括但不限于以下内容：

a） 后勤装备建设规划与发展：辅助制定后勤装备建设发展规划、制度及标准规范，基于对现有装备效能、灾害事故类型和未来发展趋势的数据分析，为规划的前瞻性与科学性提供决策支持；为战勤保障体系的构建与优化提供数据模拟与推演，并为各类灾害事故的战勤保障工作提供智能方案建议；

b） 基础设施交通工具及装备器材管理：为基础设施建设、交通工具及各类装备器材的全生命周期管理提供辅助支持；通过对装备使用频率、维护记录等数据的分析，为采购计划、调配储备、维护报废等环节提供预测性建议与决策支持；为被装、油料等物资供应管理提供智能化的库存预警与分配建议；

c） 医疗卫生勤与社会化保障：为卫勤防疫、职业健康管理、应急卫勤保障等工作提供数据分析与决策支持，辅助识别潜在健康风险并提供干预建议；为后勤服务保障社会化工作提供效能评估与优化建议，辅助进行服务商选择与管理决策。

7 评测方法

7.1 评测数据集

评测数据集应满足以下要求：

——合规性和密级要求：数据收集过程遵循适用的法规和消防行业的密级保护标准,并保护人员隐私。消防业务文档和数据具有特殊敏感性，如涉灾单位、人员信息、关键基础设施等，应建立最高等级的安全管理与脱敏机制。不同密级的数据应遵循不同的处理、存储和应用规范，并设置相应的访问控制与脱敏机制，确保数据全生命周期的绝对安全。

——评测指标完备：为每个评测指标构建满足相应数量的数据集。评测问答数据集应包括单选题、多选题、判断题、材料分析题四种问题类型。

——时效性：数据集结合开源数据集和自建数据集,定期更新维护。数据集应建立更新、维护和质量评估机制，确保数据长期可用，反映最新的消防应用场景和行业需求。

——可用性：数据集格式和接口符合广泛的标准,以便于获取和使用。资源数据应以规定文件格式之一的形式存在，不符合的需采取措施进行格式转换。

——多样性和代表性：涵盖消防及相关行业的不同专业知识、业务场景等,以确保数据能覆盖不同的使用情况。

7.2 评测环境

7.2.1 软硬件环境搭建

应根据待测模型的实际性能要求，搭建配套的软硬件平台，包括通用计算芯片、AI计算加速芯片、计算服务器、存储服务器、通信网络、云服务、容器/虚拟化等。

7.2.2 部署方式

测试框架可部署在单一服务器上进行少样本测试，也可部署在集群中进行大数据量测试。应将评测环境部署在指定的测试环境中。

7.2.3 算力配置

在模型微调或评测阶段，应综合考虑模型参数大小、训练数据规模、预期训练时长等多方因素进行适当的算力配置。

7.3 评测工具

7.3.1 自动化评测功能

应集成全面的测试集，覆盖消防大模型专业知识和业务能力的各个维度。

应支持灵活扩展功能，根据需求及时更新扩展评测数据集。

应具备确定明确的评价指标计算方法和评分规则，并根据业务需求，对评价指标体系进行迭代和更新。

应能够自动生成并输出评测结果，提高评测效率。

7.3.2 人工评测功能

应为评测人员提供相应的工具链平台，可辅助评测人员校核自动化评测结果，并可支持评测人员对模型的回答进行人工打分。

应能分析评测结果的分布和一致性，及时发现评测人员潜在的评测偏差或不一致问题。

7.4 评测实施

7.4.1 自动化评测

在评测数据集中应构建出相应的参考答案。

在自动化评测脚本中应清晰定义具体的评测指标计算方法和评分规则。

7.4.2 人工评测

应制定清晰、具体的评测标准和指南,并对评测人员进行充分的培训,确保所有评测人员对评测的标准有统一的理解和执行。

应选择具有消防专领域知识和经验的评测人员,以确保评测结果的准确性和专业性。

宜对评测人员定期进行复训,更新评测知识和技能,尤其是当标准内容有调整时。

宜定期收集评测人员的反馈,用于优化评测流程和评测标准。

7.5 评测结果评估

7.5.1 总体要求

评测结果的评估应建立一套科学、量化且可操作的指标体系。

数据质量评估应重点关注数据的准确性、一致性、完整性和及时性。

模型有效性评估：模型应能理解业务需求，针对消防领域典型的应用场景，生成可执行的策略。

应建立反馈循环机制，以迭代优化语料和模型质量，强调语料和大模型服务于实战场景。

7.5.2 打分规则

能力满分为100分，其中单选题占40分、多选题占30分、判断题占20分、主观题占10分，题目以随机的方式从评测数据集中抽取；

模型的综合得分满分为100分，由每个单项以加权平均的方式得到模型的最终综合得分，其中各单项能力权重如下表：

表1 模型测评单项能力权重表

能力测试维度	权重
模型通用基础能力	10%
消防安全法规与价值对齐	10%
应急响应与快速决策	20%
消防专业认知能力	40%
消防业务场景应用能力评测	20%

7.5.3 评测等级建议

——A级：综合得分区间在[80,100]；

——B级：综合得分区间在[60,80)；

——C级：综合得分区间在[50,60)；

——D级：综合得分区间在[0,50]；

参 考 文 献

[1] GB/T 25069-2010 信息安全技术 术语
[2] GB/T 41867-2022 信息技术 人工智能 术语
[3] GB/T 5271.1-2000 信息技术 词汇 第1部分：基本术语
[4] 中国信息通信研究院,《大模型基准测试体系研究报告（2024年）》, http://www.caict.ac.cn/