大众报业集团联合人民网共建主流价值语料库(山东)

记者 夏侯凤超 济南报道

8月25日,人民网与大众报业集团、山东数字文化集团共建主流价值语料库(山东)、主流文化语料库合作签约暨项目推进会在济南举行。人民网与大众报业集团将共建主流价值语料库(山东),进一步巩固壮大主流思想舆论,扩大主流价值影响力,加强主流价值内容传播。

主流价值语料库由人民网依托党报党网长期建设发展形成的新闻、理论、评论、政策和科普等权威优质资源,以及党和国家重要文献资源,经科学采样、归集、清洗、标注、定制、风控等环节下足"笨功夫",精心打磨而成。

人民网在长期建设发展以来,形成了新闻、理论、评论、政策和科普等权威优质资源以及党和国家重要文献资源。下一步,将通过其在全国范围内的科学采样、归集、定制,对语料进行加工整理,保障意识形态安全,充实主流价值语料库(山东)的内容;大众报业集团将充分发挥在政策、人才、规划及产业资源等方面的优势,实现本土化深耕与产业多元化相结合,依托其相关专业技术和人员,助力主流价值语料库(山东)加速构建,推动形成高价值的区域性主流价值语料平台。

双方将利用各自优势资源,共同建设覆盖经济、政治、文化、教育、科技等领域的主流价值语料库(山东),从而完善AI模型训练、舆情分析、文化传播等应用,扩大主流价值影响力,加强主流价值内容传播。同时通过深度合作,提升语料库建设的智能化水平,积极探索"以AI治理AI"的新模式,实现"人工智能+",为人工智能的普及应用提供安全、向善、可信、可控的有力保障,实现"1+1>2"的优势。

全国首个主流文化语料库上线

山东数字文化集团与人民网共建,推动数字文化产业高质量发展

记者 夏侯凤超 济南报道

8月25日,人民网与大众报业集团、山东数字文化集团共建主流价值语料库(山东)、主流文化语料库合作签约暨项目推进会在济南举行。山东数字文化集团与人民网正式签约,共建主流文化语料库。

随着生成式人工智能技术的迅猛发展,高质量、安全可信的语料库作为其关键支撑资源,对行业大模型训练和应用具有至关重要的作用。2023年12月31日,国家数据局等17部门联合印发的《"数据要素×"三年行动计划(2024—2026年)》明确指出,"完善数据资源体系,在科研、文化、交通运输等领域,推动科研机构、龙头企业等开展行业共性数据资源库建设,打造高质量人工智能大模型训练数据集。"2025年1月,国务院办公厅印发的《关于推动文化高质量发展的若干经济政策》提出,"建设文化领域人工智能高质量数据集,支持文化领域大模型建设";2025年6月,《山东省支持文化和科技深度融合协同创新的政策措施》明确,"支持文化大模型开源利用类融合创新项目,集聚版本资源、文献资源和算力资源等,建设文化领域人工智能高质量数据集和语料库,形成一批文化大模型产品和服务。"

主流文化语料库是由山东数字文化集团依托人民网、大众报业集团等党报党端党网长期建设发展形成的新闻、理论、评论、政策等权威媒体资源,以及省内文化单位、高校多年来积累的优质私域文化资源,经数据采集、清洗、预标注、标注、增强、审校等环节,以"AI+人工"的方式,精心打磨而成。该语料库具有标准统一、结构完整、权威准确、开放共享等特色,可有效解决当下AI大模型普遍存在的敏感领域语料欠缺、重要文化领域语料不足、核心语料质量不高等问题。

作为全国首个主流文化语料库,一期重点聚焦山东优秀文化,目前已上线问答语料5万对、基础语料2000万篇,正在打造孔子学术研究、孔子画像等多个高质量数据集,后续计划分期分批建设覆盖广泛、内容丰富的主流文化数据集,推动文化大模型的性能跃迁与我省数字文化产业高质量发展。

主流文化语料库建设离不开文化数据的智能标注,集团自主研发的山东文化数据标注平台,提供数据采集、清洗、预标注、标注、增强、审核等一站式全链路服务,支持问答、图片、视频、音频、文件、图谱等多类型数据标注,标注后的语料支持一键发布到大模型或智能体中,实现数据从采集到使用的全流程闭环。

山东文化数据标注平台将面向全社会免费开放,助力各文化单位、高校、企业打造自己的高质量数据集,共建主流文化语料库。下一步,山东数字文化集团还将推出山东文化数据交易平台,提供文化数据集交易服务,推动数据要素流通利用与数据资产变现,全力打造全国文化和科技融合的新高地,为文化强省、数字强省建设贡献力量。

延伸阅读

主流文化语料库将为数文产业发展带来什么

记者 夏侯凤超 济南报道

什么是主流文化语料库?它的建设对数字文化产业发展有哪些意义?

必要性
语料库是大模型能力涌现的基础

语料库作为人工智能模型训练的核心资源备受瞩目。

人工智能大模型有"三驾马车":数据、算法和算力。而随着大模型技术的迅猛发展,在算法趋同、算力普惠的背景下,高质量数据集就成了构建与训练大模型的基础性关键资源。

高质量数据集是指用于训练、验证和优化大模型而收集、整理、标注形成的覆盖行业核心专业知识和生产经营活动的数据资源集合。如果没有一个语料库来训练AI大模型,大模型就无法学习;语料库越丰富,AI大模型就会变得越熟练、越智能。因此,规模庞大、内容准确的语料库,是大模型能力涌现的基础。

高质量数据集作为人工智能核心资源的地位不断凸显。2025年2月,高质量数据集建设工作启动会在京召开,会议落实"人工智能+"行动,推动高质量数据集建设,高效赋能行业发展。2025年3月24日,国家数据局局长刘烈宏在中国发展高层论坛2025年年会上表示,"国家数据局将充分调动社会各方力量,积极推动高质量数据集建设,持续增加数据供给。"

一方面是政策方针的支持,一方面是人工智能领域的核心竞争力,因此,语料库的建设势在必行,对于助力区域及垂类产业数字经济的高质量发展具有重要意义。

是什么
为主流文化传承提供"燃料"

主流文化语料库通过标准化的语料分类系统和专业的数据标注平台,解决了目前语料库普遍存在的格式不统一、质量不齐、标准差别等问题,助力各领域和垂类打造准确性、完整性、丰富性、一致性、时效性的高质量语料库,使语料库在场景应用上更加实用和便利。为解决通用大模型常因缺乏针对性语料而"水土不服"的问题,主流文化语料库可深入区域和垂直领域的具体场景,构建富含行业术语和场景化表达的精准语料资源,进一步增强AI的理解力,提升应用效能,加速AI技术与垂类领域的深度融合,驱动产业升级。

主流文化语料库一期重点聚焦山东优秀文化,目前已上线问答语料5万对、基础语料2000万篇,正在打造孔子学术研究、孔子画像等多个高质量数据集。后续计划分期分批建设覆盖广泛、内容丰富的主流文化数据集,推动文化大模型的性能跃迁与我省数字文化产业高质量发展。

山东数字文化集团党委书记、董事长魏传强表示,"主流文化语料库的建设,是山东数文集团贯彻落实国家文化数字化战略、山东文化强省建设的必然要求,是文化与科技融合的具体举措,也是加快发展新型文化业态,实现文化建设数字化赋能、信息化转型的重要内容。"

怎么用
从数据采集到使用一站式操作平台

语料库建设的关键在于数据标注。在日前山东出台的《关于加快释放数据价值加力推进数字经济高质量发展的实施意见》中明确,加快数据标注产业发展,2027年年底前,省内建设5个成效明显、特色鲜明的数据标注基地。

文化数据标识如同为数据绘制一幅精细的"画像",清晰注明其来源、特征和价值,从而成为我们在海量信息中快速定位的"导航地图"。山东数文集团充分发挥自有资源优势,依托在国家文化大数据体系标本库、基因库、素材库建设领域积累的深厚经验,以及在对多模态数据(视频、音频、图像、文本)进行处理、标识与关联整合的成熟技术成果,进一步推动文化大数据标识基地建设。

"山东数文集团自主研发打造山东文化数据标注平台,提供数据采集、清洗、预标注、标注、增强、审核等一站式全链路服务,支持问答、图片、视频、音频、文件、图谱等多类型数据标注。"山东数文集团技术总监宋耀介绍。标注过程遵循标准化流程。用户提交原始素材入库至统一数据源,完成数据采集与归集;后台剔除重复样本及低质量内容后,完成数据清洗与筛选;核心文本数据自动标注并解析文本语义,生成结构化问答对,通过AI增强功能自动实现问题泛化和答案多样性;在问答对生成后,人工进行精校和审核,以确保数据的完备性和准确性。

优势性
数据处理全流程闭环更加高效和高质

"山东文化数据标注平台构建了高效、无缝衔接的数据处理闭环,各操作流程兼顾用户导向与智能驱动,提供高度适配、便捷高效的体验。"宋耀介绍。

在上传素材、标注和校对等过程中,平台支持多维协作功能。比如上传过程中,同一数据集可邀请多人共同上传,标注过程也可实现协作处理。为了避免重复劳动,上传后的素材进行智能清洗,保证数据不重复;标注好的数据也会注明,不会重复生成,以确保数据源唯一性及标注成果准确性。

数据校对过程中,平台在图谱类别专门打造了一套AI识别映射关系体系,区别于目前大多数平台模糊、广泛的关系呈现,这套体系可实现关系梳理一目了然。

数据在处理完毕后,平台还支持一键发布至目标大模型,发布后的大模型可立即基于该数据集进行微调或推理,实现模型能力即时更新与增强,赋能模型快速适配应用场景。

"山东文化数据标注平台面向全社会免费开放,为大模型开发中数据收集、清洗、标注和使用提供工具,也为AI算法提供必需的语料资源。我们希望通过这个平台,开源共建主流文化语料库,形成包容、开放、有序、共享的AI语料新生态,达到1+1>2的效果。"魏传强表示,"下一步,山东数字文化集团还将推出山东文化数据交易平台,提供文化数据集交易服务,推动数据要素流通利用与数据资产变现。我们将持续贯彻落实国家文化数字化战略,坚决扛牢服务文化强省建设的使命担当,为深入推进中华优秀传统文化创造性转化、创新性发展贡献力量。"

编辑:彭传刚 组版:侯波