



聚力高质量数据集 书写人工智能新篇

——贵州省人工智能语料供需对接会暨贵州数据集(语料)服务平台上线启动仪式举行

湛思宇



活动现场。

这是贵州人工智能发展迈出的坚实一步。

7月10日,贵州省人工智能语料供需对接会暨贵州数据集(语料)服务平台上线启动仪式在贵阳举行。在贵州省大数据发展管理局的指导下,由贵州大数据集团负责建设运营的贵州数据集(语料)服务平台正式上线。

当天,贵州大数据集团分别与4家数据需求方、8家数据供给方签署合作意向协议,一个规范、高效、安全的语料生态正加快构建。

拥抱新生事物,才能跟上时代前进的步伐。人工智能产业发展,数据是核心要素,高质量数据集更是人工智能创新发展的关键。

省政府副秘书长、省大数据局局长朱宗尧表示,当前,贵州智能算力优势明显,场景应用全面推进,数据生态加快构建。“十五五”时期,贵州在新一轮人工智能浪潮中的核心定位,是为全国人工智能“供算、供数、率先应用”,贵州数据集(语料)服务平台的上线运营,正是为人工智能发展提供高质量数据。

夏日的贵阳,生机盎然,欣欣向荣。“贵州数据集(语料)服务平台的上线,表明贵州正走在人工智能产业的前列。”本次活动上,北京交通大学教授张向宏不禁赞叹。

张向宏说,人工智能发展至今,正从通用大模型技术探索,转向行业大模型的产业深耕。数据是训练行业大模型的核心原料,因此建设高质量数据集乃发展所需,意义重大。

从政策层面看,近期国家数据局出台

“这是一个集‘语料超市、语料加工厂、语料生态枢纽’于一体的一站式服务平台。”活动宣介环节,贵州大数据集团党委副书记、总经理毛胤强这样介绍贵州数据集(语料)服务平台。

他说,平台上承政务与行业高质量数据资源,下接大模型训练及应用场景,可加速原始数据向高价值语料产品转化,切实打通从数据资源到AI应用落地的价值链条。

如何通过平台打通数据到AI应用的价值转化链条?贵州大数据集团给出的方案

贵州数据集(语料)服务平台上线后,将为人工智能产业带来哪些变革?

“平台要解决的核心问题,就是本地AI企业普遍面临的‘高质量语料从哪来’的痛点。”科大讯飞股份有限公司贵州大区总经理杨少明算了一笔账:过去本地中小AI团队要训练垂直场景模型,要么花高成本外购通用数据,要么自己从零开始标注,投入大、周期长,还很难保障数据合规性。

“现在通过平台盘活省内沉睡的政务与行业数据,能直接把本地企业的大模型训

的《关于推进行业高质量数据集建设行动的实施方案》,明确提出加快构建数据要素与人工智能协同演进的共生生态。

机会,稍纵即逝;发展,时不我待。贵州的高质量数据集建设规划,有迹可循。2025年底,省大数据局便审时度势,宣布谋划建设贵州语料中心。

作为全省数字经济领域的领军企业,贵州大数据集团勇担重任,历经半年多集中攻坚,顺利推出贵州数据集(语料)服务平台。

“该平台的正式上线,标志着我省在以公共数据为牵引、汇聚省内公共语料和特色产业语料、拓展省外语料资源方面迈出了关

勇立潮头开新篇

乘势而上启新程

是:聚数据、建平台、育生态。该集团依托掌握的政务标杆语料,推动贵人智办、城运小智等应用加快落地,初步形成公共数据“归集—加工—交易—AI应用”的完整闭环。同时,面向行业数据和市场需求扩大发力,对接上架百余个标准语料,覆盖十余类应用场景,为语料的价值转化奠

定基础。联合行业伙伴,该集团将贵州山地气象、酱香白酒等特色资源转化为高质量语料,让这些“贵州基因”成为训练行业大模型的硬核支撑。

资源有了,怎么用?平台搭建了统一的语料加工工具箱,

重塑产业新生态

练数据成本压下来。”杨少明说,还可以让贵州已建成的算力底座和刚上线的语料资源池真正联动起来,为全省AI产业筑牢最基础的“数据地基”。

“平台最大的价值,在于为各行业垂类大模型提供了统一的资源底座和流通渠道。”贵州中医药大学信息工程学院执行院长谢泉说,通用大模型解决基础认知问

题,真正落地到产业应用则要靠垂直领域模型,中医药正是典型的高壁垒垂类赛道。平台打通了数据供给方与需求方的对接链路,让专业数据能够快速转化为AI训练资源,支撑新药研发、智能诊疗、智慧农业等场景落地。

谢泉表示,学校将依托平台,逐步上线覆盖中医药全产业链的高质量数据集,

键一步。”贵州大数据集团党委书记、董事长李旭说,平台集语料资源汇聚、加工标注、质量评测、交易流通、模型训练等功能于一体,将成为贵州推动数据从“数据仓库”向“数据工厂”转型的关键抓手和建设国家数据要素综合试验区的重要阵地。

“平台上线很及时,直击人工智能行业痛点。”人民数据管理(北京)有限公司总经理杨森滢对平台充满信心,“人民数据已第一时间将核心语料资源接入平台,相信能为贵州语料生态建设注入‘源头活水’。”

通过有机整合智能数据治理、多模态标注与知识库管理,并引入智能预标注技术,大幅提升语料标注的效率,让数据加工能力得到较大提升。

此外,平台自规划之初,便坚持共建、共创、共享原则,着力打造全省语料产业生态枢纽,围绕数据供给、加工生产、质量评测、合规流通、场景落地等关键环节,构建分工清晰、能力互补、价值共享的协同体系。目前,平台已汇聚供数单位、需求企业、服务机构等240余家生态伙伴。

为贵州医疗健康领域的人工智能应用提供数据支撑。

“我们2025年启动的民族医药、民族语言两大特色数据集项目,今年将正式交付平台。其中苗药数据集依托自研词元技术,破解了苗侗医药知识碎片化、标准不统一的长期行业瓶颈,形成了其他地区难以复制的贵州特色优势。”大数据国家工程研究中心大数据应用技术实验室主任曹扬坚信,在全国语料产业同质化布局的背景下,平台依托贵州独有资源,可以走出差异化发展路径。

现场

现场签约12家供需企业 贵州数据集(语料)服务平台汇聚240余家生态伙伴

湛思宇

7月10日,在贵州省人工智能语料供需对接会暨贵州数据集(语料)服务平台上线启动仪式上,贵州大数据产业集团有限公司分别与4家数据需求方、8家数据供给方签署合作意向协议。目前,贵州数据集(语料)服务平台已汇聚240余家生态伙伴。

贵州数据集(语料)服务平台自规划之初便坚持共建、共创、共享原则,着力打造全省语料产业生态枢纽,为构建“供给充沛、需求活跃、合规流通、业态丰富”的贵州语料产业新格局奠定坚实基础。

“高质量数据集已成为人工智能产业发展的关键生产资料,但当前语料资源供需错位矛盾仍较为突出。资源方有数据但缺少规范流通渠道,需求方急需垂直领域语料却面临获取难、成本高、合规风险大等问题。”贵州大数据集团数据事业部负责人王仕品说,举办本次语料供需对接会,目的就是建立供需对接机制,打通线

上线下供需通道。

王仕品介绍,随着贵州数据集(语料)服务平台正式上线,一批语料资源也集中发布,并建立形成了通过“平台发布,供需对接”的机制,推动资源、需求、场景和服务精准匹配,降低企业获取语料、开发模型和合规使用的综合成本。眼下,供方、需方、加工方、服务方为一体的协作平台正加快构建,形成高质量语料的获取、标注、服务、应用生态,持续为我省数智产业高质量发展提供坚实支撑。

据了解,以本次活动为起点,贵州大数据产业集团将持续推进语料供需对接工作的体系化、常态化。重点围绕优质语料资源汇聚、数据要素合规流通、产业链协同发展、标准规范完善和应用场景落地等方面持续发力,不断打通语料流通堵点,补齐产业发展短板,打造具有区域辐射能力的语料产业发展高地,为人工智能产业持续健康发展提供长期支撑。

访谈

“大家在这个行业里,实实在在需要这样的平台。”7月10日,在贵州省人工智能语料供需对接会暨贵州数据集(语料)服务平台上线启动仪式上,人民数据管理(北京)有限公司总经理杨森滢由衷感慨。

当天,贵州数据集(语料)服务平台正式上线,人民数据作为首批签约供给方,将核心语料资源接入平台,为贵州语料生态建设注入“源头活水”。

人民数据是人民网旗下“党管数据”的理论与实践平台。近年来,人民网强化以人工智能为核心的内容科技应用,推动数据驱动的智能业务拓展与全体系智能化转型。其核心产品——人民网主流价值语料库规模超300T、基础语料超300亿字,覆盖党政、经济、文化等数十个领域,2025年入选国家数据局高质量数据集典型案例。

“此次签约入驻,我们上线了三类核心语料资源,涵盖人民网主流价值语料库、舆情数据语料库及产业链数据资源。”杨森滢介绍,主流价值语料库历经科学归集、清洗、标注、风控,可有效助力大模型完成主流价值对齐,破解大模型“幻觉”与价值偏差问题,筑牢AI内容安全底线。

杨森滢将此次合作定位为“双向互赢”。她表示,人民数据打造了面向全国的数据交易平台,一方面将优质权威语料输入贵州平台;另一方面,未来贵州的特色语料也可上架到人民数据平台,借助人民数据的资源优势、平台优势,面向全国打开数据交易的大市场。

“当本地特色语料被加工成标准化可计价的产品,文本、图片有了清晰的产权和流通属性,本地算力就能转化为AI生产力,真正实现数据变现的规模化闭环。”杨森滢表示。

谈及贵州大数据产业的发展,杨森滢给出了积极评价。她认为,贵州大数据集团在行业内发展得比较早,是风向标式的存在。而贵州本身的资源禀赋同样令人期待。

“少数民族的语言、服饰,这些语料数据都是大模型训练需要的,是贵州的特色。”她表示,贵州凭借起步早的行业优势,率先搭建省级语料综合服务平台,精准回应了行业供需对接不畅、标准不统一的痛点,对语料产业规范化、规模化发展具有重要推动作用。

据了解,贵州数据集(语料)服务平台提供语料汇聚、加工标注、质量评测、交易流通全链条服务。人民数据入驻后,将在资源归集、标准共建、场景落地等方面与平台深度协同,助力贵州构建完整的语料生态,为人工智能产业发展夯实数据底座。

7月10日,贵州省人工智能语料供需对接会暨贵州数据集(语料)服务平台上线仪式在贵阳举行,240余家生态伙伴入驻,多份供需合作协议现场签署。

在这场聚焦AI基础资源的产业盛会上,大数据国家工程研究中心大数据应用技术实验室主任曹扬表示:要用全链条质量管控,为全省AI产业的“数据原料”把好第一道关。

“我们的核心角色是第三方质量服务方,为平台流通的数据集提供质量、标准、合规三重认证。”曹扬这样明确团队定位。针对行业内“高质量语料”定义模糊的普遍痛点,他主导搭建起覆盖三大维度、17项细分指标的可量化评测体系,不仅校验数据本身的真实性、准确性与合规合理性,更开创性加入模型适应性评估模块。

“采购方在下单前,就能明确知道这份数据适配哪些场景、兼容哪些模型。”这套体系相当于给每一份流通数据附上了详尽的“品质说明书”,直接解决了本地AI企业长期面临的“不敢买、用不好”外部语料痛点,从根源上降低了企业的试错成本。

在全国语料产业同质化布局的背景下,曹扬团队锚定贵州独有资源,走出了差异化发展路径。2025年启动的民族医药、民族语言两大特色数据集项目,今年将正式交付平台。

如果说评测标准是丈量数据质量的标尺,自研工具就是让标准高效落地的核心支撑。曹扬团队打造的“数质Link”评测工具,采用“规则驱动+算法赋能”双模式,内置全部17项评测指标,可实现TB级数据集分钟级自动化评测,替代90%以上的人工抽检工作。

“产业生态的健康运转,必须建立在统一的技术标准之上。”曹扬认为,国家层面侧重产业顶层引导,而落地层面的具体技术细则,恰恰需要地方先行先试,这正是贵州的先发优势所在。依托大数据国家工程研究中心建设运营的贵州数据要素质量评测基地,以及同步发布的“数质Link V1.0”评测工具,正与贵州数据集(语料)服务平台深度协同,推进跨区域认证互认,让贵州产出的语料实现“出生即带证、流通即增值”。

曹扬把评测基地的角色凝练为“质量守门人、标准定义者”。他的目标十分明确:“让‘贵州认证’成为全国数据流通的通行证,让贵州产出的高质量数据集,在全国市场都能拥有通用的质量背书。”

(本版图片由贵州大数据集团提供)

本版责编:胡卡妮 熊燕 张元斌 杨露怡
版式设计:陈倩

——访人民数据管理(北京)有限公司总经理杨森滢

——访大数据国家工程研究中心大数据应用技术实验室主任曹扬

为AI产业『数据原料』把好关