责任编辑：严粒粒 版式：潘泓璇 联系电话：0571-85312335 邮箱：youfenglai2014@sina.com

数据是“未来的石油”，怎样更快实现资源化价值化——

杭州语料库：给AI大模型“喂好料”

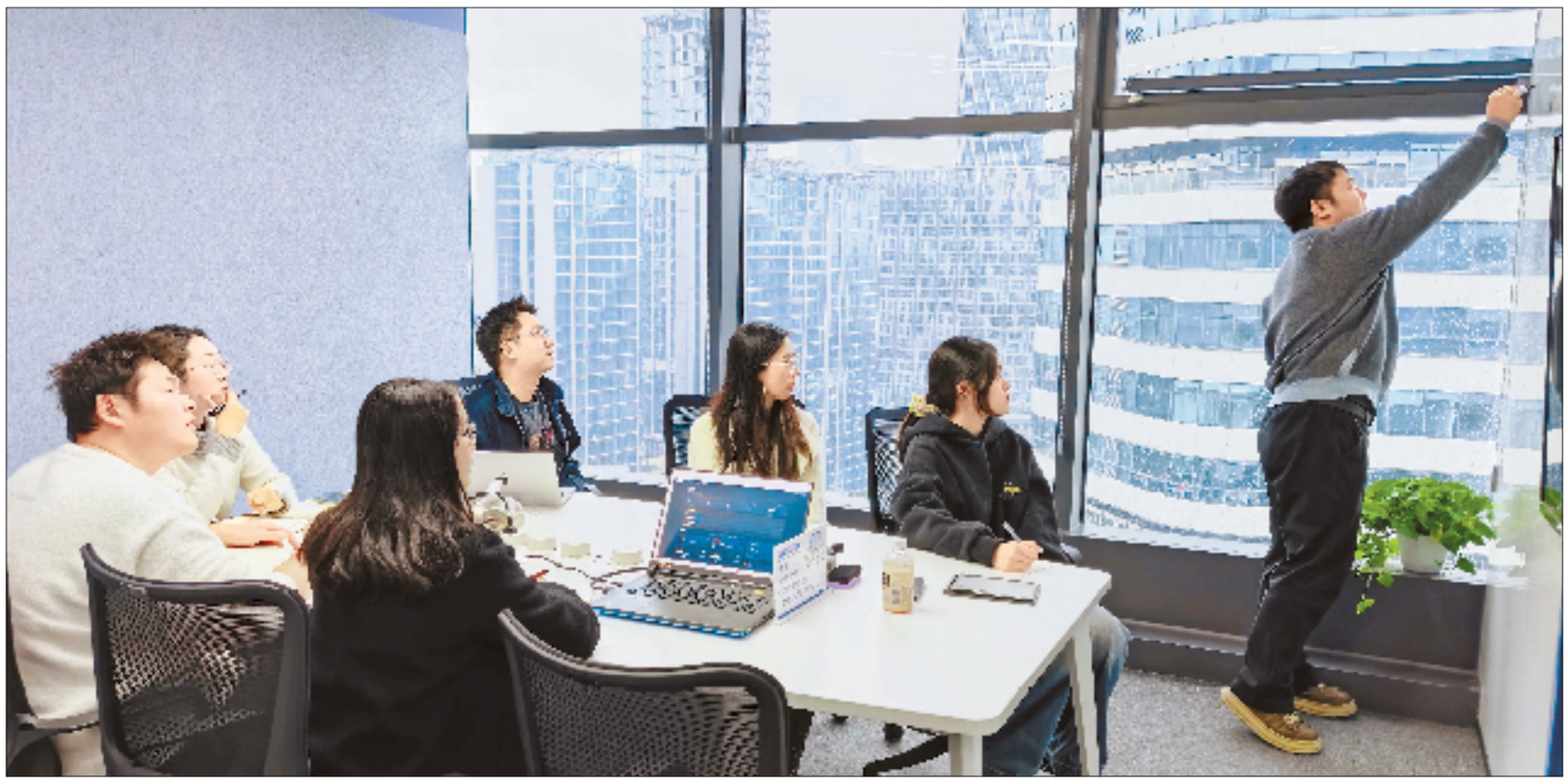
■ 本报记者 金春华

从春节到元宵，越来越多的人在这个新春选择用豆包、DeepSeek等大模型写祝福、生成拜年视频等。大家发现，大模型变得更好用、更“懂人心”了。

这背后，离不开大型高质量语料库的支持。

语料库被视为决定大模型能力上限的核心生产要素，此前以企业自建自用为主，少量对外开放。但由于各平台语料库良莠不齐，不少模型出现幻觉，“一本正经地胡说八道”。业内担忧，高质量语料库的稀缺，将制约产业长远发展。

2025年，我国启动布局新型国家语料库建设。当年底，杭州正式公布了杭州语料库建设图景和阶段性成效。其建设重点，是探索打造多元高效的数据供给、流通和应用体系，催生更多新技术、新产品、新业态，推动数据这一“未来的石油”实现资源化、价值化。



杭州景联文科技有限公司研发团队在讨论SolarSense语料工程平台开发实施细节。 受访者供图

从92号油到98号油

如果把AI大模型比作汽车，语料就是让它跑起来的汽油。

近几年，这辆车不断改造升级，已不再满足于“92号油”，而是需要适配“98号油”了。

浙江大学软件学院教授、人工智能专家张微向记者科普了几个核心概念：数据，是所有能被计算机系统存储、记录的信息。语料，全称语言材料，也即我们日常说的话，在大模型领域可以理解为AI的“学习资料”，包括文字、语音、视频等。语料经清洗、标注、结构化处理，就是语料库，也有人称之为数据集。

以浙产大模型代表之一的DeepSeek为例，其V3版训练的语料，据悉来自互联网、书籍和学术期刊等，数量达到约15万亿词元（token）。词元是大模型处理语料的基本单位，在不同大模型中，1个词元对应约0.5～2个汉字，或是3～4个英文字母。

目前，全球头部开源大模型训练的语料库规模，在10万亿～20万亿词元之间。以常见的86万字版《西游记》为参考，DeepSeekV3训练的语料约等于3000万本《西游记》，普通人不吃不喝、24小时不间断阅读，大概需要16万年。这是早期智人开始崭露头角直到现代的时间跨度。

大模型读的还不只一两本书，而是大型图书馆的藏书。

但随着大模型飞速进化，一个全球性难题出现了：语料库建设跟不上了。

据国家数据局披露，2024年初，我国日均词元的消耗量为1000亿。截至去年9月底，这一数字已突破40万亿，1年多时间增长了400多倍。

人工智能研究机构Epoch此前一项预测更显严峻：全球范围内，能训练出更优性能的高质量语言数据可能在2026年耗尽。

“人类语言一直在变，大模型想要变得跟人一样，也得及时升级语料库。”北京语言大学信息科学学院副教授柯登峰说。

作为语音识别专家，他参与过传统语料库建设，发现其与AI语料库有很大差异，“传统语料库一般只记录有代表性的说法，比如播音员的语音，但AI的语料库要尽可能覆盖人的各种说话方式和内容，最好不同年龄层、不同职业、不同受教育程度的都有涉及。”

他举了个例子：有方言专家用大半辈子收集一种方言的词汇，为1.5万条，但他们团队两个月内收集的该方言语料就有两万余条，包含了大量新词语，以及更多灵活的口头表达。

采访中，有业内人士甚至担心，大模型若缺乏高质量语料，最终只能产出“数字垃圾”。

近日，杭州互联网法院公开了一起关于AI幻觉的网络侵权纠纷案的审判情况。一位高考生的哥哥梁某在查询高校信息时，发现某AI平台生成的信息有误，且该平台在受到质疑后，仍底气十足地表示若内容有误将赔偿10万元。梁某一气之下将平台的研发公司告上法庭。一审判决驳回了诉讼请求。但此事进一步引发了人们对AI幻觉的重视。

“AI出现幻觉的原因有很多，数据缺陷、数据不足、知识更新滞后等语料方面的问题尤为致命。”迪安诊断首席科学家王宇说。前几年，大模型所用的数据主要来自网络，质量很难保障。“想要让大模型做专业的事，就得提供垂直领域的专业语料库。比如想让它给人看片子，就得让它像医生一样，不断阅读医学影像等专业文献。”

在迪安诊断实验室，工程师与检验专家正持续优化该公司的医检语料库。该语料库基于迪安诊断20余年积累的数据建立，涵盖已完成数据清洗、匿名化等处理的多组学、多模态数据，包括基因组学、微生物组学和影像组学等内容。仅病理切片数据就有约1500万份。

基于该语料库，迪安诊断去年底向杭州一家科技企业交付了一款用于疾病辅助诊断的高质量临床数据集。这也是杭州城市可信数据空间在医疗领域的首单数据集交易。

迪安诊断基于医学生物语料库研发的革兰氏染色涂片智能识别系统，可将一张涂片镜检的平均时间从15分钟缩短至约1分钟。 受访者供图

“油田”变为“炼化基地”

2025年11月公布首批数据合伙人；12月发布首批50个高质量数据集建设先行先试“揭榜挂帅”任务名单；近期又发布第二批任务……当下，杭州语料库建设脚步加快。

进入新一年，杭州数据交易所就上架了首笔具身智能数据集、首个卫星定位导航领域公共数据集等不少语料库方面的新产品。“杭州正在冲刺‘全国人工智能创新发展第一城’。在这场关乎未来产业主导权的竞速中，数据不再是附属资源，而是核心生产要素。”杭州数据交易所董事长、总经理应琦说。

从业者表示，这是杭州、上海、深圳等数据“大油田”的使命。

作为“数字经济第一城”，杭州一直在探索、鼓励数据资源化、价值化转化。2024年，“中国数谷”入选国家数字经济创新发展试验区建设案例。“中国数谷”就是一个涵盖杭州全市的数据产业集聚区。多位创业者表示，在杭州从事语料相关产业，有政策、有补贴，有技术、有市场，氛围也很好。

去年9月，杭甬温等全国10个地区获批国家要素市场化配置综合改革试点。两个多月后，随着首批高质量数据集建设先行先试“揭榜挂帅”名单等成果的发布，杭州语料库就率先与世人见面。

杭州市数据局相关负责人表示，杭州语料库的建设，主要就是为大模型训练提供易获取、高质量、规模化、低成本的语料资源，变“大油田”为“超级石油炼化基地”。

杭州有个“小目标”，争取在年底前建成100个具有一定规模的高质量数据集，服务人工智能模型训练10个以上。

首批50个高质量数据集，涉及具身智能工业场景、交通基础设施安全、医疗健康可视化等具体场景，“揭榜”的大多数是企业，横跨医疗健康、工业制造、具身智能等14个新兴领域。它们背后，是杭州扎实的、不断提供语料并生产语料库的数字经济产业。

记者发现，参与语料库建设的单位，犹如一个个同时拥有“油井”和炼油生产线的小基地。

杭州景联文科技有限公司就是其中一家。它承接了“教育大模型英文知识数据集”建设任务。“这一语料库包含了经标准化处理的英语听说读写等各类数据5600多万条，并进行了产权确认，可供其他教育科技公司、出版集团用于智能教育的研发训练。”景联文CEO刘云涛说。

记者近距离观察了其核心生产环节：

首先是“原油开采”，即多源合规语料采集，往往锁定权威英语教学素材、正规书籍期刊等优质“富油矿”；再经“原油除杂净化”，也即语料清洗筛选，如把PDF、网页、Word等不同载体统一起来，剔除劣质、杂质——错误、违规、低俗等表达；随后通过“分馏分级”，即精细化语料标注，完成难度、知识点、应用场景的精准划分，炼制成适配不同需求的“专用油品”；最终经质检封装后上架流通……

开采、提炼过程中，景联文还与浙江师范大学等专业机构合作，以保证语料质量。据悉，现在已经有教育企业来咨询该产品。

锁定上游核心资源

不少专家和从业者直言，未来的大模型之争，核心是语料库之争。

“大模型主要拼的是算法、算力和数据。算力靠芯片等硬件，算法则与数据息息相关。”柯登峰介绍，大模型算法主要分预训练、微调、人类偏好对齐、外部数据检索增强四大类，其中如预训练技术，是给大模型完成“通识基础教育”，要用百科全书式的语料；微调技术，相当于让大模型“专业定向深造”，需要医疗、教育、金融、法律等垂直领域的高质量语料库……

随着算法升级，语料库建设的专业性在不断加码。

以数据标注这一语料库建设的核心环节为例，不久前，各地曾火过一阵数据标注产业，吸引了不少人力的投入。但如今的新算法，已能让AI自主完成基础内容的标注。柯登峰打了个比方：“如果说此前的数据标注是中小学生水平，比如在图片上标注什么是树、什么是路；现在却要达到本科、硕博士水平，比如一道物理难题是如何一步步解题的，甚至还要从业多年的专家水平，比如如何判断病理。”

这些专业化的市场需求，又推动着语料库产业高速发展。

在国际上，Meta、OpenAI、谷歌等头部企业早已重金布局高质量语料库赛道，以锁定上游核心资源。国内上海等地也在大力推进语料库建设。

这场未来之争中，杭州已深度参与。

“我们在拓展专业用户，加快产品落地。”王宇介绍了同行们在努力的一个方向：培育更多市场需求。

去年，迪安诊断发布了一款订阅版科研文献智搜智能体——Repilot。它是基于海量医疗文献语料库建立的一个AI智能体。在以前，医生做课题，光文献调研，可能就要一个月，但现在输入关键词，几秒钟就能出来完整框架，还能自动匹配最新研究。

多位浙江三甲医院的医生评估，Repilot可以让他们省去约80%的低水平重复工作。

“大模型应用的重心，正由通用对话转向高价值垂直场景。融合领域知识与工作流的智能体，已成为AI商业化的关键突破口。这也可以让专业语料库建设形成资源化、价值化的闭环。”王宇说。

2月24日，迪安诊断发布了革兰氏染色涂片智能识别系统2.0版。革兰氏染色检测是识别细菌感染、指导抗生素治疗的关键手段。临床要求该检验能又快又准，但经验丰富的检验师完成一张涂片镜检，平均也要约15分钟。该系统基于迪安诊断的医学生物语料库研发，平均用时可缩短至约1分钟。

在语料库建设中，不少企业还从“卖石油”变成了兼“卖工具”，延长了产业链条。

去年，景联文发布了SolarSense语料工程平台、QApex专家众包平台。前者是统一的数据标注平台，可以把原来分散在不同团队、不同工具里的数据采集、标注和质检等流程统一起来，进而缩短交付周期、降低返工率；后者可以把专业数据传递到相关行业领域专家那里进行标注。

“语料库建设，已不再是以前的‘一次性买卖’，而是一项长期工程。”刘云涛说，这两个平台的建设，是在探索一种“平台+基地+行业”的新生产模式，以集合更多力量，实现语料库建设的规模化、可持续产出。

目前，杭州数据交易所上架的产品中，数据工具已和数据产品、数据服务成为三大主类。“智能化的数据工具能大大降低语料库建设的技术门槛，吸引更多参与者。”杭州市数据集团数据产业事业部副总经理张凯说。

面对这场未来之争，杭州还在持续培育生态，比如引进更多数据标注企业，建设语料库智能化标注基地、组建产业联盟等。

有业内人士指出，过去的标注产业偏劳动密集型，杭州的人力成本相对较高，并不占优。如今，标注产业的核心竞争力已转向AI赋能、专业知识支撑与产业生态加持，杭州在成本一效率比方面的优势就凸显了。

“我们希望更多主体参与进来。”杭州市数据局相关负责人介绍，杭州已推出系列扶持政策，比如设置“中国数谷”专项资金，在语料库开发、数据服务等方面给予资金扶持，对多模态语料库最高可给予200万元补助。

杭州语料库，值得期待。

杭州数据交易所上架的语料库部分相关产品。 受访者供图

链接

部分国内外知名AI语料库

当前，国内外机构已发布多款具备行业广泛影响力的AI语料库，覆盖通用、医疗、政务、方言、代码等多个核心赛道，以下为代表性产品的详细介绍：

1.上海市规划资源领域专项语料库，由上海市规划和自然资源局牵头研制，2025年7月正式亮相。该语料库覆盖自然资源全领域，包含规划编制、测绘地理、用地管理、不动产登记等核心业务板块，整合了学科教材、政策法规、技术标准、审批成果、城建档案等多类型数据，其中含1200余份技术标准、5.7万项覆盖近20年的城建档案成果，具备多模态、体量大、质量高、覆盖全等核心特点。目前已完成归集数据资产40TB，相当于10万部高清电影，远期规划容量可达200TB。

2.AI-Dim Sum粤语语料库平台，由广州市社会科学重点实验室——粤语语料库建设与大模型评测重点实验室研发，2025年12月正式发布。该语料库是国内规模领先的粤语多模态AI专项语料库，核心内容涵盖四大板块：文本语料，涵盖超1亿字规范处理文本，覆盖新闻、文学、社交媒体等多领域；语音音视频语料，涵盖完成3000小时高保真语音标注，整合1TB以上音视频资料，含《哪吒之魔童降世》《西游记之大圣归来》《花木兰》等动画作品，以及《外来媳妇本地郎》《溏心风暴》等经典粤语影视剧的字幕与标注语料，配套超1万句多用途粤语生活场景音文对照语料；多模态素材，含岭南文化相关图像素材10000张；安全与评测体系，构建了含6669条权威词条、30000条扩展词条的粤语安全语料库，配套粤语大模型内容安全多模态评测题超20万道。

3.The Pile，由美国Eleuther AI开源社区发起，整合了来自全球的22个高质量子数据集，于2020年12月正式发布，旨在为开源大模型提供训练数据。核心规模为825GiB未压缩高质量文本，对应约2000亿词元，覆盖学术文献、图书出版物、新闻资讯、代码、网络对话等22个细分领域。它是全球开源大模型领域的标杆通用语料库，被业内称为“开源大模型的通用语料基准”，完全开源免费。

4.MIMIC系列医疗语料库，由麻省理工学院（MIT）计算生理学实验室、美国贝斯以色列女执事医疗中心联合研发，核心版本MIMIC-CXR于2019年发布、MIMIC-IV于2020年发布，研发机构总部位于美国波士顿。其中MIMIC-IV版本覆盖超50万例住院患者的全周期电子病历，MIMIC-CXR版本包含超65万张胸部X光影像及对应临床文本报告。它是全球医疗AI领域最权威的临床级多模态语料库，所有数据均来自真实临床场景，完成了全流程合规脱敏与标准化标注，是全球医疗大模型训练、验证的黄金标准语料，被超1万篇医疗AI顶会论文引用。

（本报记者 金春华 整理）