

习近平文化思想引领下 主流价值语料库构建策略

摘要：建强主流价值语料库对于筑牢意识形态安全、实现人工智能大模型价值对齐、赋能媒体深度融合和支撑国际传播至关重要。目前，面临着顶层设计欠缺、标准规范众多、版权归属不明及中文语料全球占比过低等挑战。本文提出加强国家统筹规划、推动标准统一、创新产权机制及提升中文语料全球影响力等加强主流价值语料库建设的对策建议。

关键词：习近平文化思想 人工智能 主流价值语料库

◎ 田 丽 张 翀

在习近平文化思想引领下加强主流价值语料库建设，是巩固壮大奋进新时代主流思想舆论阵地的重大工程，对于筑牢意识形态安全、强化主流价值引领意义深远。今年4月25日，习近平总书记在主持二十届中共中央政治局第二十次集体学习时强调：“面对新一代人工智能技术快速演进的新形势，要充分发挥新型举国体制优势，坚持自立自强，突出应用导向，推动我国人工智能朝着有益、安全、公平方向健康有序发展。”语料库作为人工智能发展的核心要素，具有鲜明的意识形态属性，直接影响着人工智能的政治立场、价值导向和伦理边界。习近平文化思想关于“两个结合”“构建中国话语和中国叙事体系”等内容，为主流价值语料库建设厚植新时代思想精髓、彰显深厚文化自信、反映中国式现代化伟大进程等提供了理论遵循。

建强主流价值语料库的战略意义

建强主流价值语料库，在当前人工智能时代，既

是筑牢意识形态安全堤坝的核心支撑，也是赋能人工智能产业健康发展、驱动媒体深度融合、提升国际话语权的关键基础，对于服务国家战略全局至关重要。

筑牢安全堤坝，守护意识形态主阵地。生成式人工智能已成为意识形态交锋的关键阵地。建强主流价值语料库，本质上是打造强大而丰富的体现主流价值观的内容资源池。这些基于党的创新理论、中国式现代化发展成果、中华优秀传统文化以及中华民族精神的权威话语体系，将在人工智能时代的海量信息传播中筑起一道无形而坚固的堤坝。通过给人工智能大模型平台提供充足、精准的核心内容，将极大地提升主流话语在数字空间的抵达率与渗透力，持续强化正确价值导向，有效对冲历史虚无主义等错误思潮及恶意诋毁言论等“噪音”和“杂音”，维护国家意识形态安全，营造风清气正的网络空间。

实现价值对齐，夯实人工智能安全发展根基。作为人工智能大模型“核心燃料”的语料库，其内在价值观直接决定着输出内容的立场和观点。建设蕴含

社会主义核心价值观、体现中国文化立场的主流价值语料库，是国内人工智能大模型实现价值对齐的关键基础。通过为人工智能大模型训练提供内容丰富、导向正确的主流价值语料，从源头驱动大模型深刻理解并精准遵循我国的发展目标、社会规范与文化伦理，从而在价值层面与我国的国情民心“同频共振”。这不仅能保障大模型输出的安全性、可控性与可靠性，更能从人工智能发展的产业链源头规避潜在价值偏移风险，防范因输入“不良数据”导致文化侵蚀。主流价值语料库为我国人工智能自主创新和生态安全构筑起坚实底座，是推动产业健康、可持续发展的压舱石。

赋能深度融合，推进主流媒体系统性变革。在媒体融合向纵深发展的当下，主流价值语料库不仅是数据基础要素，更是推进主流媒体系统性变革的引擎。富含权威信息与深度知识图谱的语料库，可为智能采编系统、内容分发平台等应用注入优质的“智力源泉”，使其具备从海量信息中准确抓取热点、深刻洞察舆情、科学辅助决策并实现精准化个性化推送的能力。这使得主流媒体能以前所未有的效率优化生产流程、创新传播形态、增强用户黏性与舆论引导效能，真正实现从“融媒”向“智媒”的高质量发展。

支撑对外叙事，提升国际传播话语效能。建强主流价值语料库，是提升国际传播话语效能的底层支撑。一个体系完整、结构清晰、融通中外话语的主流价值语料库，能系统总结中国道路、理论、制度、文化的核心表述与标识性概念。基于此，我们可为国际受众定制更加符合其接受习惯、更具情感共鸣和文化认同的多语种、多媒体的“中国故事”；为智能翻译、多模态国际传播产品的生产提供准确的语义支撑，最大限度地突破语言壁垒、减少文化折扣，以国际社会易于理解的高质量叙事方式，立体、真实、全面地展现可信、可爱、可敬的中国形象，显著增强中国话语在国际舆论场的传播力与影响力。

建设主流价值语料库面临的挑战

当前主流价值语料库建设虽取得初步成效，但仍

面临着系统性不足、标准不一、权属模糊与全球影响力薄弱等突出挑战。这些挑战制约着主流价值语料库的进一步建强，需有效破解。

顶层设计欠缺，资源统筹与安全防线亟待加强。

目前，国内专注做主流价值语料库的单位不多，人民日报社主管的传播内容认知全国重点实验室在该领域发挥了引领作用，现已建成包含 3000 多万篇基础语料、30 多万对问答语料等 7 大板块的主流价值语料库。其他单位和机构大多做的是通用语料和行业专业语料，且这些语料分散于不同机构和平台，部分因数据安全性与竞争壁垒等限制而难以获取，这些语料数据缺乏顶层设计和系统化建设。同时，语料标注工作体量庞大，后续的更新与维护也需要大量人力物力的持续投入，各自重复建设势必导致资源浪费。

更为严峻的是，低质网络文本甚至带有价值偏差的内容混杂其中，导致人工智能大模型的错误信息反向污染语料库，亟须通过顶层设计加强意识形态的安全审核；而现有政策文件对语料使用的内容要求等细则尚未明确，企业创新面临合规风险，比如，数据安全法、个人信息保护法在语料领域的实施细则缺位，导致企业在数据采集时陷入“灰色地带”，限制了高质量语料库建设。

标准规范众多，“数据孤岛”制约资源整合效能。

标准规范对于语料库的系统化建设具有重要意义。我国相关主管部门和行业协会持续颁布相关标准和规范，用于指导语料库的建设。2023 年 7 月，国家网信办等 7 部门联合发布《生成式人工智能服务管理暂行办法》，提出“推动公共训练数据资源平台建设和公共数据分类分级有序开放，扩展高质量的公共训练数据资源”。2025 年，《网络安全技术生成式人工智能数据标注安全规范》发布，对人工智能训练数据的标注平台、标注工具、标注规则安全要求、标注人员要求、数据标注核验要求、数据标注安全评价等，作出明确规定。在行业和地方层面，也出台了一系列标准和规范的指导文件，比如，上海市人工智能行业协会在 2024 年 7 月发布《语料库建设导则》，用于指导语料库的建设和管理。

在语料数据的采集和标注过程中，国内很多数据服务商、主流大模型厂商并没有严格遵守这些标准和规范，取而代之的是，他们制订自己的数据采标规范和流程；多重标准并存，导致“数据孤岛”效应加剧。据统计，除基础编码规则差异外，数据标注的分词规范现存 10 余种主流方案，语义标注体系更是衍生出 20 余种自定义框架。语料数据的异构导致跨机构语料融合成本激增，国内的数据服务商和大模型厂商在使用外部数据时，又不得不花费相当比例的生产成本用于语料数据的格式转换，由此严重挤压企业自身在语料库建设上的有效投入。

版权归属不明，多方主体权益界定模糊。语料库所包含的海量数据，不可避免地涉及大量版权内容。这些版权内容大多以开源形式发布，其模糊的版权归属容易引发知识产权争议和法律风险；许多作品的作者可能既未授权共享，也未被知会其作品被收录。比如，2023 年《纽约时报》起诉 OpenAI 和微软公司，指控这些公司未经许可而使用其数百万篇文章训练 AI 大模型。AI 生成内容是否构成著作权法意义上的“作品”、版权归属如何界定等问题，现行法律规定尚不明确。由于版权问题对人工智能内容生成产业链影响巨大，目前尚存在较大争议，这使得发布者和使用者均易卷入侵权纠纷。此外，语料库建设过程中的数据清洗、标注、整理等，包含着语料库建设厂商的智慧和劳动，如何平衡数据原创者和数据处理者的权益也是有待解决的问题。

究其原因，这是技术发展和法律滞后产生的矛盾。生成式人工智能的生成内容是否构成合理使用，在司法实践中存在巨大争议：技术方主张大模型输出属于知识重构而非内容复制，而版权方则认为未经授权的商业数据训练侵害作品版权。我国现有法律既未明确“文本与数据挖掘”的免责范围，也未对衍生数据产权作出界定。尤为突出的是，当多源版权作品经过清洗、标注后形成新数据集时，原始权利人、数据加工方与模型开发者之间的权益分配缺乏法律依据，进一步削弱语料库建设的可持续性。

中文语料全球占比很小，国际话语权受到挤压。

根据阿里研究院 2024 年 5 月发布的《大模型训练数据白皮书》显示，中文语料在全球语料供给中的占比仅为 1.3%。虽然我们的主流媒体文稿、典藏文献等资源丰富，这些资源承载着主流价值观，但还存在开放程度不足、数字转化率不高的问题，大量非结构化数据无法直接转化为大模型可读的语义单元。这一“中文语料洼地”情况，迫使不少国内大模型厂商转而寻求英文数据进行模型训练，导致国产大模型在内容输出时会产生价值观的偏差。

这种数据失衡不但制约我国人工智能产业的发展，还会削弱我们的国际传播力，特别是涉及当代中国发展的语料更新滞后，如“中国式现代化”等核心概念缺乏跨语言跨文化阐释载体，阻碍了主流价值内容的对外传播。在国际舆论场中，算法认知战进一步放大风险——英文语料主导的推荐系统将中国议题标签化，而深度伪造技术配合西方价值语料库，极易产生污名化中国叙事。若不突破语料困局，我国将难以在人工智能时代提升国际话语权，中华文化的深层价值可能被压缩为浅层符号，甚至面临“他者化”重构。

建强主流价值语料库的实践路径

面对建强主流价值语料库的一系列挑战，我国需要坚持系统性思维，强化顶层设计，聚焦关键环节，推进改革创新，在习近平文化思想引领下，着力形成协同高效、安全规范、富有活力的新格局。

加强统筹规划，破解资源分散与安全风险困局。

当前主流价值语料库建设存在资源分散、标准不一、安全风险等问题，亟须以系统性思维加强国家层面的统筹规划。习近平总书记强调“坚持把马克思主义基本原理同中国具体实际相结合、同中华优秀传统文化相结合”，这就要求语料库建设必须立足主流意识形态阵地，通过顶层设计明确发展方向、资源整合机制和安全防线，将数据治理纳入国家文化数字化战略整体框架，统筹国家跨部门、跨机构协同机制，以主流价值为引领构建“语料供给侧结构性改革”新格局，从源头上保障语料库的政治性、规

范性和安全性。

可考虑从以下三个方面进行统一：一是统一规划体系，从国家层面制订主流价值语料库建设中长期规划，明确语料采集、标注、使用的流程规范；二是统一资源调度，可依托传播内容认知全国重点实验室等平台，有机汇集国内已有语料库，实现多源数据集成，解决“数据孤岛”问题，促进开放共享；三是统一安全标准，完善数据安全法在语料领域的实施细则，构建覆盖数据采集、存储、使用的全链条安全审计体系；此外，可通过设立国家语料建设专项基金，鼓励商业公司合规创新，避免重复投入，为主流价值语料库的建设提供坚实数据支撑。

推动标准统一，构建贯通全产业链的规范体系。国家层面、地方和行业层面已出台一系列语料库建设的相关标准和规范，但标准规范众多、多头分散管理，导致“数据孤岛”现象凸显，严重制约着语料资源的整合效能。习近平总书记提出“建构中国自主的知识体系”，这就要求我国以标准化建设为基石，既遵循国际技术规则，又立足国内实际，推动贯通语料数据建设全产业链的标准化体系，建构起语料领域的自主知识体系。


可考虑以已发布的《生成式人工智能服务管理暂行办法》为基准，将现有国家标准、地方规范整合升级为国家主流价值语料库建设导则，重点统一语料编码、分词、语义标注等基础规则。可考虑在国家标准化管理委员会下设立语料技术分委会，组织相关行业协会、科研院所联合攻关，推动中文语料标准框架从“20余种自定义”到“国家级通用方案”的跃升；同时，建立起语料流通“一码通”机制，通过元数据标签系统实现跨平台语料兼容，破除企业数据转换成本桎梏，为自主知识体系的构建提供标准化数据基座。

创新产权机制，明晰多方主体责任边界。语料版权纠纷折射出人工智能产业发展与传统著作权制度的深层次矛盾。习近平总书记提出“推动中华优秀传统文化创造性转化、创新性发展”，在数字语境下要求我们重新思考知识生产关系的调整路径。技术发展和

法律滞后之间的矛盾，需要通过制度创新实现原创权益保护与数据价值开发的平衡，这也是“两个结合”在法治领域的具体实践。

建议可构建分层治理体系来破解版权困局。在立法层面，完善著作权法配套制度，参考国际经验设立“文本与数据挖掘合理使用”条款，为人工智能的数据训练提供明确法律空间。在实践操作层面，探索设立国家级语料版权托管平台，通过集体授权管理模式集中协调海量作品使用权，既保障作品原创权益，又降低企业合规使用成本。此外，同步制订人工智能生成内容的权属认定规则，为人工智能技术发展筑牢法治根基。

提升中文语料全球影响力，增强国际话语主动权。中文语料在全球占比仅1.3%，不仅导致国产大模型的文化“失语”，更危及国家话语权构建。习近平总书记指出“要加快构建中国话语和中国叙事体系”，而中文语料的占比失衡正加剧“有理说不出”的传播困境。要走出这一困境，必须以“中国式现代化”的核心价值为引领，将增强中文语料的全球影响力上升为国家文化安全战略工程，打通中华文化数字资源的转化通道，使文化自信具象化为可计算、可传播的语料数字资产。

在传统文化资源转化方面，可考虑建立国家级文献数字化平台，系统整合古籍典藏、红色文献、主流媒体文稿等资源，运用智能技术实现非结构化文本的语义化改造，深度挖掘中华文明和中华优秀传统文化的时代价值。在当代话语体系构建方面，可考虑设立专项工程采集“中国式现代化”实践案例，形成涵盖政策阐释、发展成就、社会治理等多模态语料集群。同时，通过多语种的平行语料库，开发多语种对照的主流价值语料库，采取精准化、分众化的跨文化传播策略，在世界范围内各类语料数据平台、论坛及社区上推广，打破西方语料数据的话语垄断，让中国方案以可理解、可共鸣的方式参与全球文明对话。

（作者田丽系人民日报社研究部主任，张翀系人民日报社研究部主任编辑）

责任编辑：杨芳秀