赛迪展望2025

数字经济：呈现多极化、深层次创新发展格局

赛迪智库
数字经济形势分析课题组

2024年，我国数字经济统筹高质量发展和高水平安全，着眼数据协同和产业创新双轮驱动，支撑投资、消费、贸易增长新动能加速释放，助推经济发展稳中有进、新质生产力厚积薄发。展望2025年，随着“十五五”系列规划启动谋篇布局，数据要素、实数融合、产业集聚化等领域增量政策和创新制度将成谋划部署重点，多层次数据要素制度实践加速积累，“人工智能+制造业”引领产业高端化、智能化、绿色化、集聚化协同发展，国际数字治理对话和数字经济优势产能“走出去”并行推进。

我国数字经济持续稳定增长

（一）数字经济持续调优增长，新赛道新动能加快谋篇布局

2024年，我国数字经济持续稳定增长，为国民经济“稳增长”注入源头活水，高质量发展根基不断夯实。2024年1—10月份，规模以上电子信息制造业增加值同比增长12.6%，增速分别比同期工业、高技术制造业高6.8个和3.5个百分点；软件业务收入11.06万亿元，同比增长11.0%。电商平台成为消费品以旧换新重要渠道，拉动网络销售市场活力回升，1—10月份，实物商品网上零售额同比增长8.3%，增速快于社会消费品零售总额3.5个百分点。数字贸易对提升外贸效率和韧性形成重要的牵引作用，前三季度，跨境电商进出口1.88万亿美元，同比增长11.5%，在亚马逊销售额超过100万美元的中国卖家数量近两年增长近55%，对我国推进高水平对外开放的支撑作用不断增强。

展望2025年，“十五五”数字经济发展规划将进入谋划制定阶段，各领域数字化发展重大项目加紧部署，在新赛道拓展、新动能培育、新优势塑造等方面发挥重要牵引作用。一是围绕数据这一发展主线，央地协同工作机制将进一步健全，多层次、一体化数据基础设施建设及应用加快协同推进，有力支撑数据供给、流通和开发利用，推动数据要素市场化配置改革。二是立足布局谋划新质生产力，各地将更加注重数字创新生态建设，聚焦技术攻关、标准研制、适配验证和应用生态协同发展，引导更多资源要素向新一代信息技术、人工智能等新兴产业集聚，优化人形机器人、脑机接口、6G等未来产业中长期政策支持。三是城市将成为统筹数字经济新质生产力培育和适数化改革的综合试验场，数字经济创新发展、数据基础制度综合改革将依托城市纵深推进，制造业新型技术改造、中小企业数字化转型等试点城市将深入探索产业数字化转型有效路径，各地数字经济将呈现多极化、深层次创新发展格局，进一步推动经济社会结构全面重塑。

（二）数据要素市场供给活跃，制度和产业双需求激活价值

2024年，央地协同、纵向贯通的数据管理工作体系基本建立，数据要素市场化配置改革开局良好。数据基础制度体系加快完善，公共数据开发利用、数字经济高质量发展、城市全域数字化转型、可信数据空间发展等重要政策密集出台。数据资源供给持续丰富，截至2024年7月，已经有243个省级和城市的地方政府上线数据开放平台，开放的有效数据集超过37万个，最近8年增长44倍。城市治理、金融服务、绿色低碳等重点领域开放的有效数据集数位居前列。数据要素市场运营体系深入推进，数据企业数量超过19万家，24家数据交易机构签订互认互通协议，北数所、上数所、深数所等在行业数据专区、数据跨境清单等流通模式探索取得明显进展。数据要素开发利用热潮兴起，工业领域数据要素应用、“数据要素×”、数字经济创新发展试验区、数字中国等典型应用场景大量涌现，数据应用广度和深度持续拓展。

展望2025年，数据要素市场化配置改革将深入推进，更多领域、更多地方将积极探索数据基础制度落地实施方案，全面释放公共数据授权运营、可信数据空间建设、“数据要素×”场景创新等发展需求，进一步激发大数据产业活力，加速数据要素化、价值化进程。一方面，各地将采用数据基础制度综合改革实验田建设等方式，推动数据确权授权、流通交易、收益分配等创新制度先行先试，构建具有地方特色的数据高质量供给、高效率流通和多样化应用创新路径。另一方面，围绕全国一体化数据要素市场建设，各地将持续发展壮大大数据产业，试点开展数据企业、数据产品认定和培育，打造集数据标注、数据集构建、模型训练、多架构算力支持等于一体的数据工厂，实现数据供得出、流得动、用得好。

（三）数实融合聚力协同推进，“人工智能+”引领三化协同

2024年，各地以制造业数字化转型为抓手，围绕“点、线、面”协同发力，推动实体经济和数字经济深度融合。智能工厂建设步伐加快，截至9月，累计培育421家国家级智能制造示范工厂，13家中国企业新入选全球“灯塔工厂”，中国“灯塔工厂”总数已达72家，占全球42%。重点行业数字化转型走向深水区，工业互联网与重点产业链融合发展取得重要进展，钢铁、电子信息、工程机械等重点行业融合应用指南发布，围绕产品设计、计划调度、质量管控等15个环节的40个典型融合应用场景成为行业标杆。产业集群数字化转型成为新抓手，如广东省围绕战略性支柱产业集群和战略性新兴产业集群开展“链式改造”，针对行业集聚度高、链条比较完整的产业集群，从产业集群资源共享、协同制造等重点环节切入，推动东莞松山湖电子信息、广州花都狮岭箱包皮具、佛山顺德小家电等16个产业集群率先试点。

展望2025年，人工智能作为时代关键变量的作用将加速释放，以“人工智能+”为特点的数字化赋能推动数实融合走深向实，驱动产业走向高端化、智能化、绿色化协同发展。一是电子信息、装备制造、消费品、原材料等数字化基础较好的重点行业将进一步推动“人工智能+”应用，聚焦研发、生产、质检和供应链管理等关键环节，发展数字孪生设计、人机协同生产、质量智能检测、供应链精准协同等新模式新应用，引领生产方式和组织形态变革。二是行业骨干企业和国有企业将联合大模型企业，开展垂类大模型研发和行业数据集建设，运用“大模型+小模型”融合方式落地建设典型应用场景，逐步推动智能驱动的业务流程、运营模式和决策机制转变。三是电力、工业、交通、建筑等重点碳排放领域将积极探索“人工智能+”应用模式路径，发展“碳足迹”“碳标签”等智能应用，提升实数融合“含智量”“含绿量”。

（四）数字产业探索集聚化路径，数据和创新赋能生态竞争

2024年，各地积极布局新一代信息技术、数据要素、人工智能等产业发展，发挥重大项目牵引作用，加强企业梯队培育，运用集聚化方式实现快速成长壮大。无锡物联网、上海集成电路等数字领域的国家先进制造业集群加快建设，围绕大数据、软件、工业互联网等方向打造了一批国家新型工业化产业示范基地。浙江杭州、上海张江、河南郑州等地聚焦数据要素产业新赛道，以算力中心、数据标注、数据要素开发利用等为牵引，打造数商集聚优势，积极探索数据要素驱动、数字平台支撑的产业集聚化发展路径。数字产业加快开放发展，速卖通、TikTok、希音、Temu等跨境电商带动平台上的中小企业组团“走出去”，亚马逊全球站点成为国内中小企业海外业务拓展的集聚地，亚马逊平台全球前20大商户基地中有13家中国城市，占比超过60%，深圳、广州、莆田的卖家数量占13家城市卖家总数的一半以上。

展望2025年，各地将加快推动数字产业高质量发展，以创新引领和数据驱动为主线，打通产业数据链，提升创新策源和大规模产业化能力，构建链群数字化协同发展生态，积极探索新要素驱动的新发展路径。一是聚焦数据驱动，依托数据基础设施统筹建设，加快推动区域内外关联产业的数据汇聚、流通和开发利用，提升全要素生产率，拓展数字新应用、新业态、新价值。二是聚焦创新引领，更加注重发挥企业主导作用，将建立中长期创新投入机制，运用生态思维构建数字技术攻关、标准研制、适配验证和应用生态协同发展的新型创新体系，着力实现数字技术革命性突破、掌握技术生态话语权。三是聚焦跨域协同，积极探索基于平台的“跨越地理边界”发展模式，利用关键技术能力、生产制造基地、市场订单规模等优势，推动跨域产业链供应链实现基础设施联通、重要数据连接、规则标准贯通、合作利益分享，促进域外资源向本地集聚和流动配置，为本土产业发展注入新动力、拓展新空间。

（五）数字贸易规则竞争加剧，数字市场开放成为合作焦点

2024年，数字贸易规则的构建与发展成为世界各国竞相角逐的焦点，围绕全球数字经济市场和全球数字经济治理话语权的争夺愈演愈烈。一方面，美国、欧盟等主要经济体正在推动制定新领域规则、签订区域贸易协定等，加快部署体现西方价值观和利益诉求的数字经贸规则。我国作为数字贸易大国，积极推进加入《全面与进步跨太平洋伙伴关系协定》（CPTPP）和《数字经济伙伴关系协定》（DEPA）的进程，发布《全球数据跨境流动合作倡议》，“丝路电商”伙伴国增加到33个，数字领域的开放合作不断扩大。

展望2025年，随着地缘政治局势紧张以及数字主权意识提升，各国政府将越来越注重对本国数字产业的自主决策权，全球数字贸易规则竞争将日趋激烈。一方面，数字贸易规则碎片化与数字贸易全球化的矛盾将更加激烈，全球数字贸易继续维持增长态势，不同国家也将结合自身情况制定多样性规则。如何平衡规则的一致性和灵活性，将成为我国发展数字贸易、融入全球数字治理体系的重要关卡。另一方面，围绕落实《全球数据跨境流动合作倡议》《数字经济和绿色发展国际经贸合作框架倡议》，我国将积极参与数字贸易、数字平台、数据跨境流通、人工智能治理等领域国际规则制定，推动多双边数字治理合作平台和机制建设，在开放合作中打造数字经济发展“命运共同体”。

不断推进数据要素市场建设

（一）加强数据协同治理，增强产业数据供给能力

一是引导各地数据管理国家标准（DCMM）贯标，鼓励有条件的地方和企业积极探索首席数据官制度建设，为企业构建数据治理体系提供技术指引和人才支持。二是持续推进数据要素市场建设，面向北京、上海等地数据交易机构搭建全国交流平台，鼓励各类市场主体参与数据要素市场建设，探索数据集、数据产品、数据服务等多种形式的数据交易模式，推动数据要素价值转化。三是引导龙头企业建设高质量行业数据集，探索建立同行业企业数据集共享机制，满足不同行业发展人工智能的数据需求。

（二）改革创新投入机制，推动价值链中高端开放

一是提升数字技术态势研判能力，推动建立数字技术瞭望平台，编制发布数字经济新兴领域、重点领域发展态势报告和技术路线图，加强与国内外一流行业研究机构、技术研究机构合作交流，全面跟踪发展热点、革新亮点、技术路径、壁垒难点等，支持投资机构投早、投小。二是加快技术攻关和应用生态协同发展，推动建立企业家、技术投资机构、重要行业用户共同参与的协同创新机制，在创新需求凝练、创新项目攻关、创新技术适配验证、创新项目应用推广等方面形成合力。三是倡导国有资本向数字经济产业领域倾斜，进一步降低对国有基金的投资回报周期要求，释放国有背景投资机构牵引力，让民间投资机构“站队跟随”。

（三）注重新技术能力培养，提升数字经济就业质量

一是围绕人工智能、工业互联网、集成电路等重点领域，鼓励职业院校、行业企业等主体联合打造数字技能培训体系，动态更新培训教材，优化实习实训资源，为产业输送高水平、高素养数字人才。二是引导互联网信息服务、电子商务等行业企业结合新技术应用和新业态拓展，设立新型数字人才岗位，促进数字人才深度融入数字经济各环节各领域。三是分行业分专业组织数字人才供需对接活动，引导教育机构与行业企业建立联系渠道，实现数字人才和就业岗位的精准匹配。

加快人工智能语料库建设 促进大模型性能实现飞跃

赛迪智库电子信息研究所
谢馨慧 赵燕 陈渌萍

人工智能（AI）语料库是汇集大量来自书籍、学术文章、社交媒体等渠道的文本、图片、音频、视频数据集合，是人工智能领域研究和应用的基础数据。目前，国际主流大模型训练语料库以英文语料为主，中文语料占比不超过5%。中文人工智能语料库匮乏制约了我国大模型性能飞跃和技术创新。赛迪智库电子信息研究所建议加快专业语料建设，提升语料数据质量；优化基础设施建设，维护语料数据安全；完善语料生态环境，构建评估作价体系。

国内外AI语料库存在差异

大规模、高质量的语料数据是训练和评估模型的基础。一是从海量语料数据中提取语法结构、语义特征能够提升模型泛化性和准确性。OpenAI基于3000亿个单词和超过40TB语料训练GPT-3模型，能够准确理解用户问题并生成自然流畅的文本内容。谷歌使用涵盖书籍、新闻等广泛领域的海量文本训练BERT模型，使其文本翻译、情感识别等任务的准确度提高。二是高质量语料数据可以提高模型性能和训练效率。谷歌PaLM2模型采用包含多种语言和科学数据的改进语料库训练，其翻译、推理、代码生成能力得到显著提升。三是专业领域语料库驱动AI技术创新和应用落地。通用语料库难以满足特定专业领域需求，通过收集医疗、金融等专业领域的术语和概念扩展专业领域语料库，加速相关领域算法创新和应用推广。

国外语料库在数据规模、开源建设和应用场景方面具有先发优势。一是英文语料库数据规模庞大，语料来源渠道丰富。GPT-3训练语料Common Crawl广泛收集了来自网页文本、书籍和学术论文等多渠道的文本数据，数据规模达到拍量级（1PB=220GB）。华盛顿大学等高校机构组织构建的开源数据集MINT-1T，包含1万亿个文本构建块和30亿个图像。二是英文AI语料库在标准化建设和开源共享方面拥有优势。欧洲语言资源协调机构通过制定数据采集、标注和共享标准，整合欧洲各国及全球范围内的语料资源，推动语料库规范化发展。谷歌、微软等科技巨头允许开发者通过应用程序开发接口访问其语料库。三是国外企业和研究机构正加大对多模态AI语料库的建设力度。多模态AI语料库能够提升模型处理复杂任务和跨领域应用的能力。Meta借助社交平台积累多模态语料提升模型对图像的理解能力，并将其集成在智能眼镜上。亚马逊通过构建语音语料库，推动其语音助手在智能家居和语音交互领域的应用。

国内企业和研究机构积极跟进中文AI语料库建设。一是中文AI语料库在数据规模和多样性方面取得显著进展。中国大模型语料数据联盟发布“书生·万卷”多模态语料库，涵盖来自网页、书籍、百科等不同来源的清洗后预训练语料，数据规模超2TB。智源研究院联合多家数据单位建设全球最大中文语料数据库WuDaoCorpora，涵盖1.2TB中文文本数据、2.5TB中文图文数据。二是特定行业或专业领域的中文AI语料库建设已初具规模。科大讯飞构建用于训练和优化语音识别模型的语料库，包含多种语言、方言和口音的数据。上海交通大学创建包含6种语言和21种医学子课题的多语言医疗语料库，用于提高医疗诊断模型的准确度。南京大学以法律文书、司法考试为基础构建法律领域对话数据集，以提高模型对法律内容的理解能力。三是高质量中文语料短缺是当前语料库建设亟待解决的问题。现有中文语料来源广泛但质量参差不齐，未经清洗包含错别字、语法错误和价值观偏见的语料会影响模型训练效果。此外，我国语料库建设规范性不足，数据标注标准不一、语料库结构差异明显以及相关企业共享意愿不足，导致高质量中文语料积累薄弱。

AI语料库面临三大挑战

语料收集受限于数据来源、版权以及隐私保护法规。一是语料来源的单一性限制了对多样化、高质量文本数据的获取。尤其在特定专业领域语料资源匮乏的情况下，难以收集足够的文本数据来训练更具泛化性的AI模型。二是版权问题进一步增加了语料收集的难度。文本资源通常受到版权保护，未经授权的使用可能引起法律纠纷，也限制了研究人员和开发者对语料的获取和使用。三是隐私保护法规对语料收集提出了严格要求。例如，欧盟《通用数据保护条例》规定在处理涉及个人信息的数据时，必须确保匿名化或得到数据主体的明确同意，否则将面临法律风险，同时增加了语料收集的成本。

语料数据的清洗和标注需要投入大量人力成本。一是语料清洁性是语料库建设、流通和使用的前提。对收集到的原始语料进行过程繁琐的去噪、去重、标准化等清洗操作，以确保输入模型数据的准确性和一致性。二是专业语料标注通常依赖人工标注。语料标注的专业性、复杂性要求标注者具备专业知识，能够对语料进行初步分析和判断，如词性标注、句法结构标注、情感分析等。三是语料标注容易受到标注者主观判断的影响。自动化标注工具虽有所发展，但其在处理复杂语义或细微语境时的精度和可靠性尚不能完全替代人工标注，而不同标注者的主观判断标准不同，将导致标注不一致或标注错误。

海量语料存储、同步处理和安全管理的难度大。一是大规模语料库需要庞大算力设施支撑。语料库规模不断扩大，企业和研究机构需要购买大量分布式存储系统、图形处理单元和云计算平台等技术设备，而中小型企业和研究机构往往难以承担基础设施建设和维护的成本。二是分布式存储系统面临不同节点语料同步处理困难的问题。存储节点分散、语料分布不均、网络传输延迟等因素导致分布式存储系统难以完成对实时性要求高的任务。三是语料库面临网络攻击、数据泄露等安全隐患。海量语料中可能包含大量敏感、有价值的数据，分布式存储环境增加了语料库被黑客攻击的风险。

不断提升语料数据质量

加快专业语料建设，提升语料数据质量。一方面，加大对专业领域语料库的建设投入。通过设立专项基金或项目资金补贴等方式支持专业领域语料库建设和运营，同时，引导企业、科研机构、高校等主体形成合作共建联合体，推进跨领域、跨机构合作的数据资源共享，实现专业领域语料的有效整合，提高语料资源的利用率。另一方面，优化数据收集与标注流程。结合自动化工具与人工审查，定期对语料进行更新扩充、监测维护，并形成优质的标准化语料库和完备的数据生命周期管理体系，确保语料数据的质量。

优化基础设施建设，维护语料数据安全。一方面，优化计算资源配置与基础设施建设。采用混合云架构、自动化调度和负载均衡技术，根据训练任务需求合理规划资源配置，提高语料库使用效率。另一方面，加强对语料安全与隐私保护技术的研发。采用加密技术、访问控制等手段，确保数据的安全和用户的隐私。鼓励企业建立数据安全管理体系，定期进行安全评估和漏洞检测，确保语料库的安全性。

完善语料生态环境，构建评估作价体系。一方面，从国家层面建立大规模、公开的语料库。面向社会各界征集高质量语料资源，通过给予奖励和补贴等形式鼓励优势企业和研究机构参与中文国家AI语料库建设，推进具有科研价值的公共语料资源的开放力度。另一方面，建立语料产品评估标准和作价体系，明确语料版权归属。鼓励行业内企业和科研机构共同探索数据合作机制与商业模式，促进语料资源在合法合规前提下的开放共享与交易。