采购需求

我院“数智哲社”项目旨在构建上海哲学社会科学研究智能化场景，其中主要包括:多学科垂类大模型建设、赋能多种学术和决策咨询成果生产的智能体模型群构建、全面支撑数智哲社的智算服务底座建设、供给各类垂类模型预训练和成果生产的语料库建设和语料生产服务等。

一、项目建设主要内容

社会科学高价值数据购置：聚焦社会科学院研究数据资源的获取整合与标准化处理，涵盖中央领导人文化思想最佳实践地建设—上海文化发展文献数据、上海文化地图数据、马克思主义经典著作和党的创新理论数据、中国政策数据、宗教文献与互联网宗教资源数据、历史文献典籍数据、法律研究资料数据、应用经济数据、海外精英人物数据、海外主流媒体报道数据 10 个重点领域。通过统一的数据采集、清洗、分类、更新与合规管控，构建分类清晰、动态更新、安全可控的高价值数据集，解决传统数据碎片化、格式不统一、质量无保障的问题。

社会科学院自有语料集成与管理：目前我院数据量已经持有商业数据与自有知识产权数据，本项目需要针对已有的两类数据进行数据集成与语料化管理，其中商业数据包含：OECD 系列数据库、外文综合学术资源数据库、知网、学术期刊、文学出版各类数据；自有知识产权数据包含：社会科学院论文数据、社会科学院期刊数据、社会科学院自有数据库数据。

社会科学院垂类智算大模型微调服务：基于主流开源大模型，结合社科领域特色语料完成三大垂类模型的精细化训练。党的创新理论垂类大模型以历代领导人思想理论为核心，整合经典著作与实践案例，保障理论解读的准确性与政治合规性；全过程人民民主实践垂类大模型聚焦上海本地实践，整合基层民主、司法等案例，支撑相关报告撰写与政策分析；区域国别计算分析垂类大模型依托海外精英人物、媒体报道、企业出海等数据，适配国际媒体局势研判与全球动态监测需求。模型采用 “预训练 + 多阶段微调 + 知识增强” 技术路线，确保适配社科研究场景，可与后续应用系统无缝对接。

社会科学院科研 AI 应用系统建设：围绕社会科学院科研核心痛点，开发四大类智能应用工具。案件材料速读分析助手聚焦法学研究，实现案件信息高效提取、案情解构与类案匹配，大幅缩短材料处理周期；区域国别人物数据分析工具构建全球精英人物画像，支持关联关系挖掘与观点趋势研判，可视化呈现人物网络；区域国别媒体数据分析工具，基于大模型能力，结合全球主流媒体信息，帮助科研人员可以快速查找外媒信息，并且针对这些信息实现多维分析；通用 AI 工具涵盖多语言互译、文本校对、OCR 提取、我院媒体风险监测功能，覆盖跨学科研究与日常科研办公需求，提升成果产出效率与质量。

社会科学院 AI 语料库治理平台建设：搭建覆盖语料全生命周期的规范化管理平台，核心功能包括数据自动分类、多维度数据清洗、数据共享交换接口、智能检索、数据管理与数据标注。平台参照国家及地方相关标准，建立多层级数据分类目录与唯一编码体系；通过自动化工具与人工校验结合，实现数据去重、纠错、补全与合规性管控；提供标准化 API 接口，支撑与中央厨房数据、智算平台的数据互通；支持关键词、语义、垂类等多模式检索，满足精准查找需求；为模型训练与科研创新提供高质量语料支撑。

二、社会科学高价值数据数据获取与治理

中央领导人文化思想最佳实践地建设—上海文化发展文献数据：包括中央领导人总书记关于文化建设的系列重要讲话、指示批示等，以及对上海文化建设的要求和期许等内容。为落实国家文化战略和《上海市建设中央领导人文化思想最佳实践地行动方案》等相关政策文件，上海在文化体制机制改革、文化产业发展、公共文化服务等方面的举措，上海文化发展成效。上海社会科学院等科研机构关于上海文化领域学习、实践中央领导人文化思想的研究成果，以及关于上海文化的研究成果，包括学术论文、研究报告、著作（包括蓝皮书在内）等。数据获取与治理总量 15－20 万条，其中政策文件类 2－3 万条、学术论文类 8－10 万篇、媒体报道类 5 － 7 万篇。

上海文化地图数据：包括文化创意生产、文化体验空间和文化标识符号数据，数据需要整合了重要公共文化设施、商业文化设施和大型活动等数据，以及城市口碑、国际节点、城市品牌等方面的数据，可用于反映上海的文化设施分布、文

化影响力等情况，为上海文化地图的构建提供了理论和数据支持。数据获取与治理总量 8－12 万条，涵盖文化地标（2－3 万处）、文化场馆（1.5－2 万座）、非遗项目（0.5－1 万项）、文化活动（4－6 万场 / 年，按近 2 年数据采集）。

马克思主义经典著作和党的创新理论数据：收集历代领导人思想理论相关资料，构建语料库、年表及电子化资料，同时收集实践案例主要包括马克思、恩格斯、列宁等经典作家的原著及相关研究文献和中国共产党历代领导人的著作、讲话和重要文献，获取对应数据并清洗入库。数据获取与治理总量 30 － 40 万条，其中经典著作类（马克思主义经典著作及国内研究著作）5－8 万部、理论文章类（期刊论文、会议论文）15－20 万篇、政策解读类（中央及地方政策解读材料）8 － 12 万份，涵盖 1949 年至今。

中国政策数据：包括党和国家的重要会议文件、政策法规、领导人讲话等，如中央领导人总书记关于政治建设、国家治理等方面的重要论述，这些是理解和分析中国政治发展的重要依据，能为报告提供政策导向和理论支撑。数据获取与治理总量 80－100 万条，按层级分为中央政策（30－40 万条）、地方政策（上海及其他省市，50－60 万条），按领域分为经济、文化、社会、生态等，时间跨度 1978 年至今。

宗教文献与互联网宗教资源数据：国内外学者撰写的专业宗教研究书籍，例如：全世界宗教的起源、发展理论基础；关于中国佛教、道教研究的诸多著作等。获取对应数据并清洗入库。数据获取与治理总量 5 － 8 万条，涵盖佛教（2 － 3 万条）、道教（1－1.5 万条）、基督教（1－1.5 万条）、天主教（0.7－1 万条）等宗教的经典文献、研究著作、政策法规、地方宗教活动记录，时间跨度从古代至今。

历史文献典籍数据：主要包括记录中国古代的正史、地方史志、国外的历史文献等。 数据获取与治理总量 10 － 15 万条，涵盖上海地方史文献（3 － 5 万条）、中国通史文献（4－6 万条）、专题史文献（如经济史、文化史，3－4 万条），包含古籍善本（1 － 2 万部，以扫描件为主）、近现代文献（9 － 13 万篇）。

法律研究资料数据：涵盖法律法规，包括全国性和地方性的法律、法规、司法解释等；司法案例，如最高人民法院指导性案例及各级人民法院裁判文书，并

标注相关元数据；学术文献，有法学期刊、综合性期刊涉法文献及学位论文；法律实务资料，像法律意见书、行业调研报告等；国际法律文件，包含国际组织发布的法律文件和条约。对国际法院、国际投资仲裁、国际海洋法法庭、联合国贸法会等机构判例，获取对应数据并清洗入库。数据获取与治理总量 60 - 80 万条，按层级分为法律（1－2 万部，含宪法及部门法）、行政法规（3－5 万部）、地方性法规（上海及全国其他省市，30－40 万部）、规章（部门规章与地方政府规章，26 - 33 万部。

应用经济数据：含经济文献，涉及各领域经济理论和应用的期刊论文、书籍等；统计数据，有国内外官方统计机构和行业协会发布的经济数据；政策文件，包括国家和地方政府的经济政策、发展规划等。数据获取与治理总量 100－120 万条，按领域分为上海重点产业数据（30－35 万条）、长三角经济协同数据（25－30 万条）、国际经济对标数据（20－25 万条）、宏观经济监测数据（25－30 万条），统计周期含月度（40 - 50 万条）、季度（30 - 35 万条）、年度（30 - 35 万条）。

海外精英人物数据：针对中国学所的研究对象获取海外各类精英人物数据，如：青年汉学家、重要国家领导人、议员等。数据获取与治理总量 5－8 万条，涵盖政治领域（1－1.5 万条）、经济领域（2－2.5 万条）、文化领域（1-1.5 万条）、科技领域（1 - 2 万条）、学术领域（0.5 - 0.8 万条），覆盖全球主要国家。

海外主流媒体报道数据：获取国外重要媒体对我国的报道。数据获取与治理总量 20 - 30 万条，覆盖海外主流媒体（英文媒体 15 - 20 万条，含《华尔街日报》《金融时报》等；非英文媒体 5－10 万条，含法国《世界报》、日本《读卖新闻》等），报道主题涉及中国（含上海）政治、经济、文化、社会等。

社会科学高价值数据数据获取与治理			
序号	**数据类别**	**数据描述**	**数据量**
1	**中央领导人文化思想最佳实践地建设—上海文化发展文献数据**	**包括中央领导人总书记关于文化建设的系列重要讲话、指示批示等，以及对上海文化建设的要求和期许等内容。为落实国家文化战略和《上海市建设中央领导人文化思想最佳实践地行动方案》等相关政策文件，上海在文化体**	**预估总量 15 - 20 万条**

		制机制改革、文化产业发展、公共文化服务等方面的举措，上海文化发展成效。上海社会科学院等科研机构关于上海文化领域学习、实践中央领导人文化思想的研究成果，以及关于上海文化的研究成果，包括学术论文、研究报告、著作（包括蓝皮书在内）等。	
2	上海文化地图数据	包括文化创意生产、文化体验空间和文化标识符号数据，数据需要整合了重要公共文化设施、商业文化设施和大型活动等数据，以及城市口碑、国际节点、城市品牌等方面的数据，可用于反映上海的文化设施分布、文化影响力等情况，为上海文化地图的构建提供了理论和数据支持。	预估总量 8 - 12 万条
3	马克思主义经典著作和党的创新理论数据	收集历代领导人思想理论相关资料，构建语料库、年表及电子化资料，同时收集实践案例主要包括马克思、恩格斯、列宁等经典作家的原著及相关研究文献和中国共产党历代领导人的著作、讲话和重要文献，获取对应数据并清洗入库。	预估总量 30 - 40 万条
4	中国政策数据	包括党和国家的重要会议文件、政策法规、领导人讲话等，如中央领导人总书记关于政治建设、国家治理等方面的重要论述，这些是理解和分析中国政治发展的重要依据，能为报告提供政策导向和理论支撑。获取对应数据并清洗入库。	预估总量 80 - 100 万条
5	宗教文献与互联网宗教资源数据	国内外学者撰写的专业宗教研究书籍，例如：全世界宗教的起源、发展理论基础；关于中国佛教、道教研究的诸多著作等。获取对应数据并清洗入库。	预估总量 5 - 8 万条
6	历史文献典籍数据	主要包括记录中国古代的正史、地方史志、国外的历史文献等，获取对应数据并清洗入库。	预估总量 10 - 15 万条
7	法律研究资料数据	涵盖法律法规，包括全国性和地方性的法律、法规、司法解释等；司法案例，如最高人民法院指导性案例及各级法院裁判文书，并标注相关元数据；学术文献，有法学期刊、综合性期刊涉法文献及学位论文；法律实务资料，像法律意见书、行业调研报告等；国际法律文件，包含国际组织发布的法律文件和条约。对国际法院、国际投资仲裁、国际海洋法法庭、联合国贸法会等机构判例，获取对应数据并清洗入库。	预估总量 60 - 80 万条
8	应用经济数据	含经济文献，涉及各领域经济理论和应用的期刊论文、书籍等；统计数据，有国内外官方统计机构和行业协会发布的经济数据；政策文	预估总量 100 - 120 万条

		件，包括国家和地方政府的经济政策、发展规划等；	
9	海外精英人物数据	针对中国学所的研究对象获取海外各类精英人物数据，如：青年汉学家、重要国家领导人、议员等，获取对应数据并清洗入库。	预估总量 5 - 8 万条
10	海外主流媒体报道数据	获取国外重要媒体对我国的报道，获取对应数据并清洗入库。	预估总量 20 - 30 万条

三、社会科学院自有语料集成与管理

目前我院数据量已经持有商业数据与自有知识产权数据，本项目需要针对已有的两类数据进行数据集成与语料化管理，其中商业数据包含：OECD 系列数据库、外文综合学术资源数据库、知网、学术期刊、文学出版各类数据；自有知识产权数据包含：社会科学院论文数据、社会科学院期刊数据、社会科学院自有数据库数据。

社会科学院自有语料集成与管理		
序号	数据种类	需要集成与管理语料内容
1	商业数据	OECD 系列数据库：经济合作与发展组织（OECD）推出的 OECD Statistics、OECD-WTO TiVA、OECD STRI 及 OECD-WTO BaTIS 等一系列全球经济政策与监测数据库，包括宏观统计、价值链贸易、服务贸易限制性指数及平衡服务贸易数据等，用户广泛覆盖世界各国的机构、组织、高校及科研单位。
2		中外文综合学术资源数据库：《CNKI 中国知网学术期刊库》是全球最大的中文学术资源数据库，广泛服务于全球 56 个国家和地区的数万家机构用户及超 1.7 亿个人用户，用户范围覆盖教育科研、党政军机关、公共文化等众多领域，使用量持续位居全球同类网站前列。
3		知网：知网作为国内权威学术资源平台，是项目学术数据获取的核心渠道。其涵盖期刊论文、博硕士学位论文、会议论文、报纸文章、年鉴、专利文献等多种类型资源。
4		学术期刊：聚焦人文社科领域核心、权威学术期刊，包括《中国社会科学》《学术月刊》《法学研究》《文学评论》等国内顶级期刊，以及上海社会科学院主办的《社会科学》《上海经济研究》等特色期刊。这些期刊发表的论文具有较高学术价值，涵盖学科前沿理论、实证研究成果、政策分析报告等内容。

5		**文学出版：主要包括文学类专著、文集、作品集、文化发展报告等正式出版成果，既涵盖经典文学作品、现代文学创作，也包括上海本地文化出版成果。**
6	**自有知识产权数据**	**社会科学院本院发布的各类论文数据：社会科学院本院发布的各类论文数据。**
7		**社会科学院期刊数据：社会科学院本院发布的各类期刊数据。**
8		**社会科学院自有数据库数据：社会科学院本院发布的文学出版数据。**

四、社会科学院垂类智算大模型微调服务

基于主流开源大模型，结合社科领域特色语料完成三大垂类模型的精细化训练。党的创新理论垂类大模型以历代领导人思想理论为核心，整合经典著作与实践案例，保障理论解读的准确性与政治合规性；全过程人民民主实践垂类大模型聚焦上海本地实践，整合基层民主、司法等案例，支撑相关报告撰写与政策分析；区域国别计算分析垂类大模型依托海外精英人物、媒体报道等数据，适配国际媒体局势研判与全球动态监测需求。模型采用 “预训练 + 多阶段微调 + 知识增强” 技术路线，确保适配社科研究场景，可与后续应用系统无缝对接。

党的创新理论垂类大模型微调：基于开源模型进行微调，该模型围绕党的创新理论体系化与学理化，构建语料库、年表及电子化资料，通过系统梳理理论观点、收集实践案例、呈现思想发展脉络训练。该模型通过对大量相关文本数据的学习和分析，能够深入理解党的创新理论的内涵、意义和实践要求。为确保模型既具备大模型通用理解能力，又能精准体现党的创新理论语义特征，微调需采用“预训练模型+多阶段垂类微调+知识增强”的技术路线。微调后大模型理论问答准确率≥80%，政治表述合规率 99%。

全过程人民民主上海实践垂类大模型微调：基于开源模型进行微调，以理论赋能加实践落地为双导向，构建聚焦全过程人民民主上海实践的专业化垂类大模型，为科研分析、政策制定、实践总结提供智能支撑，助力上海打造全过程人民民主最佳实践地。可以帮助用户梳理全过程人民民主在不同领域的实践情况，如立法过程中基层立法联系点收集民意、司法实践中公众参与等案例素材。为用户撰写内容全面、数据准确、分析深入的 “丝路电商” 合作发展报告提供有力支

持，帮助用户清晰展现 “丝路电商” 在推动 “一带一路” 经贸合作中所发挥的重要作用及未来发展趋势。为确保模型既具备大模型通用理解能力，又能精准体现全过程人民民主上海实践语义特征，微调需采用“预训练模型+多阶段垂类微调+知识增强”的技术路线。微调后大模型理论问答准确率≥80%，政治表述合规率 99%。

区域国别计算分析垂类大模型微调：面向区域国别研究的专业化需求，构建融合多源数据与智能分析能力的垂类大模型，实现对全球重点国家 / 区域的动态监测、关联分析与趋势预判，服务国家对外开放战略与上海 “五个中心” 建设。为确保模型既具备大模型通用理解能力，又能精准体现区域国别计算分析特征，微调需采用“预训练模型+多阶段垂类微调+知识增强”的技术路线。微调后大模型理论问答准确率≥80%，政治表述合规率 99%。

五、社会科学院科研 AI 应用系统

围绕科研核心痛点，开发四大类智能应用工具。案件材料速读分析助手聚焦法学研究，实现案件信息高效提取、案情解构与类案匹配，大幅缩短材料处理周期；区域国别人物数据分析工具构建全球精英人物画像，支持关联关系挖掘与观点趋势研判，可视化呈现人物网络；区域国别媒体数据分析工具支持多维度检索与热点、情感分析，助力全球媒体信息精准把握；区域国别企业出海数据分析工具支持多面向多个东南亚及中东地区的中国企业出海相关数据分析，实现助理企业出海情况分析；通用 AI 工具涵盖多语言互译、文本校对、OCR 提取、媒体风险监测、蓝皮书数据监测功能，覆盖跨学科研究与日常科研办公需求，提升成果产出效率与质量。

社会科学院科研 AI 应用系统		
序号	**功能名称**	**功能描述**
1	**案件材料速读分析助手**	**聚焦法学所刑事法学研究、司法实务分析及政策建议输出需求，构建高效信息提取、 深度案情解构、专业研究适配的案件材料速读分析助手，为法学所开展类案研究、司法规律总结、政策建议撰写提供智能化支撑，助力提升法学研究与实务结合的精准度和效率。**
2	**区域国别人物数据分析**	**基于大模型能力，针对全世界各类精英人物搭建分析系统，实现如：青年汉学家、国外政府首脑、议员等人物的关联分析。**

3	区域国别媒体数据分析	基于大模型能力，结合全球主流媒体信息，帮助科研人员可以快速查找外媒信息，并且针对这些信息实现多维分析。
4	区域国别企业出海数据分析	基于大模型能力，结合东南亚与中东部分国家的中国企业出海相关的信息，帮助科研人员可以快速查找外媒信息，并且针对这些信息实现多维分析。
5	通用 AI 工具	多语言互译：基于大模型能力为国际所、新闻所、宗教所、中国学所提供学术研究过程中的便捷翻译工具，实现 20+语种互相翻译。
		文本校对：聚焦社会科学院全院系各类文稿写作纠错，支持发现知识性差错、文字标点差错、内容导向风险识别三大类错误。
		OCR 提取：基于先进的 OCR 引擎，可高效准确地识别多种语言的文本，支持从 PDF、图像等多种格式文档中提取文字，并可将结果导出为多种格式，满足跨语言文档处理需求。
		媒体风险监测：针对本院发布的自媒体信息以及公开网站，进行发文信息的多维统计与风险发现。
		蓝皮书数据监测：基于模型能力，实现面向海量蓝皮书文档的监测与解析，自动分类高价值报告，并实现针对报告的内容分析。

六、社会科学院 AI 语料库治理平台

搭建覆盖语料全生命周期的规范化管理平台，核心功能包括数据自动分类、多维度数据清洗、数据共享交换接口、智能检索、数据管理与数据标注。平台参照国家及地方相关标准，建立多层级数据分类目录与唯一编码体系；通过自动化工具与人工校验结合，实现数据去重、纠错、补全与合规性管控；提供标准化 API 接口，支撑与中央厨房数据、智算平台的数据互通；支持关键词、语义、垂类等多模式检索，满足精准查找需求；基于三级权限模型实现数据安全管控，配套文本分类、实体识别等标注工具，为模型训练与科研创新提供高质量语料支撑。

社会科学院 AI 语料库治理平台		
序号	功能名称	功能描述
1	数据自动分类	参照《国家文化大数据体系建设标准》《上海市公共数据开放暂行办法》，结合项目服务的 14 类数据，制定“社会科学数据分类目录”，明确一级分类、二级分类及三级分类，确保每类数据有清晰归属。

2	多维度数据清洗	对重复数据，通过内容哈希比对，保留最新版本并标记关联关系；对文本类数据中的错别字、标点错误，结合社会科学术语库，自动识别并修正，适配；对缺失数据，通过同源数据补充和标标缺失，确保数据完整性。
3	数据共享交换接口	提供标准化数据 API 接口，支持数据订阅、推送、下载，数据使用审计：记录数据访问、下载、调用日志
4	智能检索	运用先进的算法与索引技术，支持用户通过关键词、语义等多维度快速精准地查找所需语料，提升获取语料的效率。
5	数据管理	具备全面且完善的语料存储、增删改查、更新以及权限管理等功能，可有效保障语料库数据的完整性、安全性与有序性。
6	数据标注	借助智能工具与人工辅助相结合的方式，对语料进行准确的分类、属性标注等操作，便于后续的模型训练及各类分析利用。

七、售后服务要求

本项目从系统验收通过之日起 1 年内提供 7*24 小时免费技术支持和售后服务，1 年后进入有偿维护期。在质量保证期内，供应商将按照售后服务的承诺提供保修和运行维护服务。

在质量保证期内，供应商负责信息系统的运行维护工作，确保信息系统安全、稳定、可靠地运行。本项目涉及的运行维护工作范围为：信息系统漏洞的修复、信息系统正常运行定期巡检。

八、应急响应要求

供应商对系统故障应能够实时响应，若系统发生故障，接到通知后 30 分钟之内响应，专业工程师 1 小时内到达现场。特殊故障与客户沟通协商后，按照协商的方式制定解决方案并进行处理。

具体故障级别及对应的应急响应要求如下：

一级故障：在 1 小时内确诊，总故障解决时间不超过 4 小时。

二级故障：在 2 小时内确诊，并在 4 小时内由专家到达现场确诊并解决，总故障解决时间不超过 8 小时；

三、四级故障：在 4 小时内确诊故障，总故障解决时间不超过 16 小时。

九、培训要求

对系统使用单位提供业务操作培训，应提供详细培训方案。

(1)在 12 个月的质量保证期内，提供 8 次与项目相关的必要培训。

(2)供应商需要开展分层次的人员培训工作，每次培训后应对参加培训人员进行测试，评估培训成果。培训应具有培训教材、培训环境和高水平的培训讲师。

(3)供应商应提供一般用户的基础操作培训和部门信息管理员的日常应用维护的培训，确保用户对象能够掌握对应的操作技能。

十、进度要求

投标人应根据建设内容，分阶段制定合理的时间进度，并且应根据招标方要求进行调整和细化。

总建设周期为 4 个月，分为 4 个阶段。

第一阶段为 0.5 个月，完成项目交付团队的组建与需求调研。

第二阶段为 0.5 个月，完成项目的概要设计与详细设计。

第三阶段为 2 个月，完成项目的整体功能开发与部署。

第四阶段为 1 个月，完成项目的培训与试运行。

十一、项目团队人员要求

1） 投标人须具有稳定的在职技术保障力量，能够提供及时的技术支援或服务，应针对本项目提供不少于 20 人的项目服务团队（包括项目经理、产品经理、技术负责人、研发等），投标单位的相关服务人员需具备相应的服务能力，需提供相关证明。

2） 投标人具有数据采集与治理相关的软件著作权证书或专利专有技术证明材料的为优。

角色	主要职责	人员数量	人员要求
项目经理	负责项目质量和进度控制	1 人	项目经理须具备项目管理师资质
产品经理	负责项目需求评估与产品设计	3 人	产品经理需具备系

			统集成项目管理工程师 1 名
研发	负责项目具体开发与实施	17 人	研发需具备高级系统架构设计师 1 名

十二、技术文件要求

投标人提供的书面技术资料应能确保系统正常运行所需的管理、运营及维护有关的全套文件。技术文件应该全面、完整、详细。投标人提供的技术文件至少应包括：

- 系统功能说明手册；
- 技术手册(安装、测试、操作、维护、故障排除等)；
- 项目文档，应该包括：

(1)软件需求说明书

(2) 系统总体设计说明书

(3) 应用软件功能清单

提供全套技术文件纸介质 3 套以及电子文件 1 套。