ICS 35.240.99
CCS L70

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIAS 031—2025

科学智能语料库建设导则

Guidelines for the Construction of Artificial Intelligence for Science Corpora

2025-07-24 发布　　2025-07-25 实施

上海市人工智能行业协会　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由上海市人工智能行业协会提出并归口。

本文件起草单位：上海库帕思科技有限公司、上海市人工智能行业协会、上海科学智能研究院、上海人工智能实验室、上海创智学院、上海算法创新研究院大模型中心、北京深势科技有限公司、鸿之微科技（上海）股份有限公司、上海商汤智能科技有限公司、联通（上海）产业互联网有限公司、中国电信股份有限公司上海分公司、联通数据智能有限公司、上海宝信软件股份有限公司、上海工业自动化仪表研究院有限公司、国创智造科技（上海）有限公司、上海阶跃星辰智能科技有限公司、东华大学、上海岩芯数智人工智能科技有限公司、国家管网集团储能技术有限公司、国家工业信息安全发展研究中心人工智能所、上海联影智能医疗科技有限公司。

本文件主要起草人：山栋明、黄海清、漆远、邹亮、程远、石伯明、丁晓东、熊飞宇、孔令和、张林峰、钟俊浩、曹荣根、施佳樑、邓思文、杨闻博、曹宇、饶雪、赵春昊、汤洁、李孟渚、李吉羊、张谦、张颖慧、宋纯锋、白磊、李萌、郑啟豪、王宇涵、李佳秾、唐诗翔、薛东雨、任昱宸、黄维然、冯恺睿、李昊、罗烨、唐波、志宇、魏文强、蔡晓晨、李永超、龚奎、史涛涛、王建、肖玲、胡顺波、刘国鑫、陆知雨、陈若曦、路少卿、堵炜炜、虞祝豪、朱莉琴、杨文恺、郑更河、华静、宋佳琪、汤凯锋、章伟、张驰、潘登、赵兴华、杜镇泽、胡兵、张洪、张伟、郭爱华、沈彦、肖红练、贺仁龙、郑茂宽、陈巧慧、孙雯、王娜、沈滨、杨华，周永星，胡银银、李建君、张火箭、段冲、李卫、陈翔宇、谭晓坤、郭汉杰、魏飞、陈磊、曲晓婷、林一琪。

本文件首次制定。

首期执行单位：上海库帕思科技有限公司、上海市人工智能行业协会、上海科学智能研究院、上海人工智能实验室、上海创智学院、上海算法创新研究院大模型中心、北京深势科技有限公司、鸿之微科技（上海）股份有限公司、上海商汤智能科技有限公司、联通（上海）产业互联网有限公司、中国电信股份有限公司上海分公司、联通数据智能有限公司、上海宝信软件股份有限公司、上海工业自动化仪表研究院有限公司、国创智造科技（上海）有限公司、上海阶跃星辰智能科技有限公司、东华大学、上海岩芯数智人工智能科技有限公司、国家管网集团储能技术有限公司、国家工业信息安全发展研究中心人工智能所。

引　言

人工智能是新一轮科技革命和产业变革的重要驱动力量，语料则是人工智能研究和应用不可或缺的资源，高质量的语料库更是人工智能赋能新质生产力的关键。

在人工智能的浪潮中，科学智能作为前沿科技的代表，正受到国家和上海市的高度重视。《科学智能语料库建设导则》的编纂，正是在这一背景下应运而生，旨在为该领域的发展提供坚实的数据基础和标准化指导。

国家层面，科技部等六部门关于印发《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》，明确提出要围绕高水平科研活动打造重大场景，充分发挥人工智能技术在文献数据获取、实验预测、结果分析等方面作用，推动人工智能技术成为解决数学、化学、地学、材料、生物和空间科学等领域的重大科学问题的新范式，为科学智能语料库的建设提供了政策指导。

上海市层面，上海市经济和信息化委员会、上海市发展和改革委员会等部门颁布了《上海市推动人工智能大模型创新发展若干措施（2023-2025年）》，强调要开展科学智能等领域的大模型研究、推进科学智能大模型应用，支持相关主体建设科学智能创新中心、算法创新基地等平台，协调算力资源和科研数据集，推动科学智能大模型在生命科学、工程计算、气象等领域应用，打造科学研究新范式。进一步为科学智能的发展明确了重点领域和发展方向。此外，上海科学智能研究院、浦江实验室等科研机构积极探索科学智能相关技术和工具的研究应用，为科学智能发展和语料库建设提供了坚实的技术支持。

本导则积极响应国家和上海市的政策要求和发展方向，为语料库的建设提供科学、系统、标准化的指导。本导则将详细从科学智能数据资源的属性及其采集、清洗、标注、测试、应用等各个环节进行阐述，以满足科学智能在不同学科场景下的通用需求，保证数据的多元可靠。我们期待本导则能够进一步推动科学智能技术的创新和应用，为科学智能的可持续发展注入不竭动力。

科学智能语料库建设导则

1 范围

本文件提供了建设科学智能模型训练数据内容、数据采集、数据清洗、数据标注、语料测试、语料使用和数据安全方面的技术指导方法。

本文件适用于科学智能语料库的研究、开发、维护、应用、评估等工作。其它与科学智能语料库建设相关的工作也可参照使用。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 4894—2009 信息与文献 术语

GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求

GB/T 22240—2020 信息安全技术的网络安全等级保护定级指南

GB/T 35273—2020 信息安全技术 个人信息安全规范

GB/T 36073—2018 数据管理能力成熟度评估模型

GB/T 42755—2023 人工智能 面向机器学习的数据标注规程

GB/T 45577—2025 数据安全技术 数据安全风险评估方法

YD/T 4245—2023 电信网和互联网数据脱敏技术要求和测试方法

T/SAIAS 015—2024 语料库建设导则

3 术语和定义

T/SAIAS 015—2024界定的以及下列术语和定义适用于本文件。

3.1

科学智能 Artificial Intelligence for Science

利用大语言模型、机器学习、深度学习、数据挖掘和计算建模等人工智能技术，加速科学研究、发现新知识或解决复杂科学问题的新兴领域。

3.2

数据资源 Data Resource

以电子化形式记录和保存的具备原始性、可机器读取、可供社会化再利用的数据集合。

3.3

科学智能语料库 AI for Science Corpora

经过系统化采集、清洗、标注、测试，可用于科学研究相关的数据库和知识库构建、模型训练和智能体研发等科学研究领域的多源异构数据集合及其管理系统。

3.4

合成数据 Synthetic Data

通过计算机仿真、生成模型等技术人工生成的数据。

3.5

研究型数据 Research Data

各类科研活动和项目中产生的原始数据及其衍生形式

注：研究型数据包括但不限于实验观测、模拟结果、调查记录、代码、算法等所有支撑研究结论的原始材料。

3.6

管理型数据 Management Data

各类科学研究活动相关的支撑性、过程性、管理性的记录。

注：管理型数据包括但不限于项目审批文件、伦理审查记录、数据使用协议、资源分配日志等，服务于科学研究的合规性、资源调配及流程监控，并不直接参与科学研究活动。

4 缩略语

下列缩略语适用于本文件。

AI:人工智能（Artificial Intelligence）

CNN:卷积神经网络 (Convolutional Neural Network)

CoT:思维链（Chain of Thought）

FAIR:可发现性、可访问性、互操作性和可重用性（Findability, Accessibility, Interoperability and Reusability）

MeSH:医学主题词(Medical Subject Headings)

SciBERT:科学文献语言模型(A Scientific Cased BERT)

SFT:有监督微调（Supervised Fine-Tuning）

XRD:X射线衍射（X-ray Diffraction）

5 科学智能语料

5.1 科学智能数据分类

科学智能数据包括研究型数据和管理型数据。

5.2 科学智能数据的来源

科学智能数据来源包括真实数据及合成数据。

5.3 科学智能元数据

科学智能元数据涵盖多种类型，各学科宜依据自身特点选择对应元数据类型。

表1 科学智能元数据类型

类型	主要内容
描述性元数据	描述数据内容、主题及资源标识的信息，用于资源的快速发现和多模态数据关联对齐
技术元数据	描述数据的物理存储结构、格式及访问接口技术细节，优化存储效率与读写性能
管理性元数据	记录数据管理生命周期信息，包括质量评估、安全策略及版本控制
情景元数据	记录数据生成环境、实验参数及预处理过程，支撑实验复现和特征工程
业务元数据	将数据映射到学科领域的业务规则、术语及计算口径
溯源元数据	记录数据血缘关系，包括来源、转换过程及迭代路径
语义元数据	通过本体论和知识图谱描述数据的语义关系，支持机器推理

5.4 科学智能数据储存格式

科学智能数据应存储为通用、开放、可扩展、可互操作、长期可读的数据格式，便于后续处理和分析。依照数据模态不同，主要数据信息储存格式应满足文本与代码、表格、图像、视频、多维数组或网格数据、时序数据、结构化数据、3D模型/点云/网格数据等模态的要求。

5.5 科学智能语料的总体原则

科学智能语料建设的总体原则应包括以下内容：

a) 应覆盖对应科研场景的实际需求；

b) 宜符合 FAIR 的规定原则；

c) 可具备多种模态，包括但不限于文本、图像、视频等；

d) 不宜设置加密或编辑限制、访问限制等内容操作权限，确保数据可用；
e) 应满足科学研究生态伙伴的兼容性和可扩展性。

6 科学智能数据采集

6.1 真实数据采集

6.1.1 来源

真实数据的主要来源于科学研究的各类场景和活动，包括但不限于文献与专利数据、书籍与报告数据、专业数据库数据、实验观测数据、科学装置及基础设施数据等。

6.1.2 采集要求

真实数据采集符合下列要求：
a) 数据应确保真实准确，反映出符合科学规范的信息和性质；
b) 数据宜覆盖学科领域及主要科研场景；
c) 数据应在学科或专业的数据有效期或更新周期内采集。

6.2 合成数据采集

6.2.1 合成类型

合成数据采集类型主要包括以下类型：
a) 由数值仿真算法模拟真实系统行为生成的数据，严格满足物理定律，适合机理明确的物理系统；
b) 将观测数据和仿真数据结合，同时嵌入领域知识生成的增强数据；
c) 利用生成式AI模型生成的数据或符合GB/T 42755-2023要求标注的数据；
d) 利用机器学习模型模拟个体行为的交互规则生成群体行为数据，需要环境的交互与反馈，常用于复杂系统研究；
e) 依据领域专家设定的逻辑规则或统计分布生成数据，无需物理模型或训练过程。

6.2.2 采集要求

合成数据采集符合下列要求：
a) 计算仿真过程应模拟真实科研活动，保证数据符合科学规律；
b) 计算仿真数据应涵盖学科研究及实验场景，并基于不同场景和多种参数条件随机化处理。

6.2.3 合成数据的意义

合成数据在科学智能中的意义主要包括以下内容：
a) 针对真实科学研究中个别极端情况难以采集的问题，计算仿真生成极端情况数据可增强数据的鲁棒性；
b) 计算仿真数据通过科研场景泛化来模拟各种科研活动中的变化情况，通过引入不同的科研活动场景的参数条件为模型提供更丰富的训练样本。

6.3 数据资源提交方式

6.3.1 数据文件标识

数据(资源)文件应通过文件名称进行标识并遵循以下要求：
a) 文件名称=文件名+文件扩展名；
b) 命名通常不含有中文字符和不合法字符；
c) 在后续使用过程中不能对数据集重命名，以免数据无法回溯或导致数据丢失。

6.3.2 数据资源的提供

本文件数据资源提交方式宜满足T/SAIAS 015—2024中的数据资源提交方式。数据资源的提交方式包括以下几种：

a) 实体存储介质方式是指将数据资源文件按一定的格式和组织形式（如压缩）存入实体存储介质后进行的数据交换方式命名通常不含有中文字符和不合法字符；
b) 云盘传输方式是指将数据资源文件按一定的格式和组织形式（如压缩）后通过公有或私有云盘转储所实施的数据交换方式；
c) 直连在线方式是指数据资源供给和接收双方通过光纤专线点对点进行数据传输，这一方式具有较高的安全性和可靠性；
d) 数据空间是互相信任的合作伙伴之间的数据关系，每一方都对其数据的存储和共享适用相同的标准和规则。

7 科学智能数据清洗

7.1 数据清洗流程

7.1.1 数据评估与规划

对数据的完整性、准确性、一致性等进行评估，确定数据中存在的问题；根据数据的特点和分析需求，制定相应的清洗策略，明确清洗的目标和方法。

7.1.2 缺失值处理

对每个字段计算其缺失值比例，并根据字段的重要性差异采取以下处理方式：

a) 对于不重要的字段或缺失率过高的数据，可以直接删除该字段；
b) 对于重要的字段或缺失率尚可的数据，可通过业务知识或过往经验进行推测填充、使用同一指标数据的均值、中位数等进行填充或根据不同指标数据之间的关系进行填充等处理方法。

7.1.3 异常值处理

检查字段内容中是否存在不合逻辑的字符、不合法的输入值或与字段应有语义不一致的情况，同时可借助散点图方法、箱线图方法等方法识别异常值检测。对于异常值应进行剔除或修改。

7.1.4 数据格式化

检查多源数据在格式和内容上是否存在不一致的问题并进行统一。同时将文本转换为统一的小写或大写，以确保一致性。

7.1.5 数据去重

使用字符串模式识别技术清理数据集中的重复记录。

7.1.6 数据去噪

使用正则表达式去除语料中的特殊字符、数字等非文本内容。如果语料的特殊字符比例超过预定义的阈值，则删除该语料。

7.1.7 数据脱敏

对数据中涵盖的个人信息、敏感位置信息、敏感实验操作信息、商业机密信息等内容，须进行脱敏处理，数据脱敏策略包括但不限于替换、泛化、屏蔽、加密、数据扰动等。

7.1.8 数据修正

数据修正的内容包括但不限于停用词去除、词性标注、词性校正、词形还原等。

7.2 数据规范

7.2.1 数据规范管理

所有的数据资源需根据6.3.1中所规定的文件标识进行统一命名，并以6.3.1中所规定各数据表征模式的文件格式之一形式存在。

7.2.2 元数据规范管理

科学智能元数据宜涵盖基础标识层、内容描述层和科学情境层的相关内容。

表2 科学智能元数据层级

层级	主要内容
基础标识层	唯一标识符：为每条语料分配全局唯一ID，确保全生命周期可追溯 物理属性：标注格式（文本/图像/音视频）、大小、存储路径及编码类型 权限信息：声明数据使用许可协议、访问权限和安全等级
内容描述层	学科分类：按学科树状标签标注领域 多模态关联：标注图文/音视频的对应关系 术语标准化：映射领域本体
科学情境层	数据来源：记录采集方式、仪器型号及参数 时空信息：标注数据时空戳 研究关联：链接相关学术论文、专利或项目编号

7.3 数据清洗方法

7.3.1 规则驱动型清洗

基于预定义的逻辑规则、语法模式或业务标准，通过程序化脚本或表达式直接处理语料中的错误、噪声和不一致。

表3 规则驱动型清洗方法示例

方法	技术实现	适用场景
结构化校验	正则表达式（日期/单位/编码格式） 字典匹配（预设值域校验）	实验表格字段规范化
冗余剔除	哈希指纹去重（MD5/SHA256） Jaccard相似度计算（文本/日志类数据）	重复实验记录合并
缺失值填补	数值型：中位数插补 + 回归模型预测（XGBoost） 类别型：知识库推理（如化学物质类别继承）	仪器参数缺失修复
逻辑冲突检测	DROOLS规则引擎（如“熔点 > 沸点”矛盾判定）	实验记录逻辑纠错

7.3.2 模型增强型清洗

利用机器学习模型自动识别语料中的语义错误、边界模糊问题，并通过学习历史数据动态优化清洗策略。

表4 模型增强型清洗方法示例

任务	科学数据案例
异常检测	电镜图像异常斑点识别 光谱数据噪声峰过滤
实体纠错	文献中专业术语表述修正 电镜样品污染区域标注
跨模态对齐	文献描述与XRD谱图匹配度验证
语义补全	材料合成流程段落自动补写

7.3.3 多模态协同清洗

整合文本、图像、表格等多源数据，通过跨模态对齐检测不一致性，验证图像与文本（例如电镜图像和实验文本）、图像与表格（例如光谱曲线与实验属性表格）等数据组合的匹配性，借助多模态大模型联合优化清洗过程。主要的协同方式包括静态-静态对齐、静态-动态对齐、动态-动态对齐、知识增强型对齐等。

示例1：电镜图像-实验文本对齐案例，见图1。
示例2：光谱曲线-实验表格对齐案例，见图2.

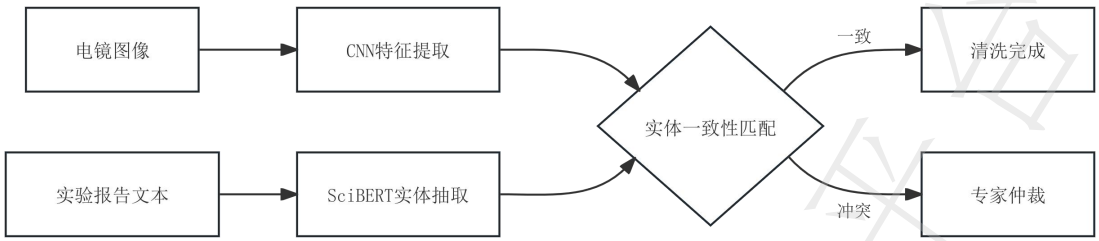

图1 电镜图像-实验文本对齐案例

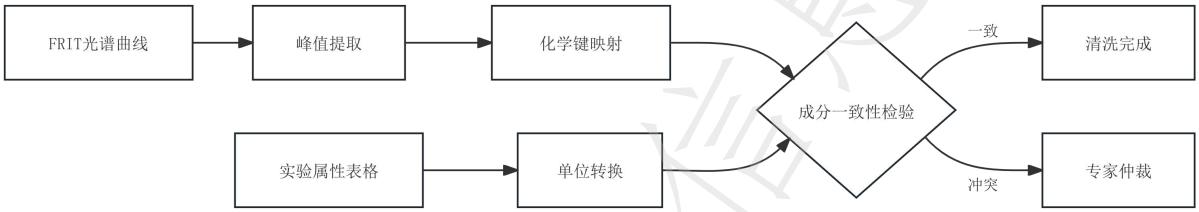

图2 光谱曲线-实验表格对齐案例

8 科学智能数据标注

8.1 数据标注通用流程

8.1.1 准备阶段

8.1.1.1 确定标注需求

明确语料模态及其场景，确定需要标注的内容范围，设计或选择合适的标签体系，确保标签具有明确的定义和层次结构。

8.1.1.2 确定标注方法

详细定义每个标注类别的含义、特征和示例，对应说明标注的具体方法和步骤，并列出标注过程中需要注意的事项。

8.1.1.3 选择标注工具

科学智能标注工具应具备专用性、安全性与可扩展架构以适应科研数据处理需求。基于数据的不同模态宜选择合适的自动化标注工具，处理规则明确、重复性高的任务，快速生成标注结果。

8.1.1.4 开发标注指南

标注前宜开发详细的标注指南，明确标注标准和流程，在后期依据实际情况定期评估更新标注规则。

8.1.1.5 建立数据标注人才培训体系

针对数据标注要求构建完善的数据标注人才培训体系，并对标注人员进行系统性培训。

表5 多层级的数据标注人才体系构成

层级	核心要求
通用标注员	基础工具操作、术语识别等
垂类标注员	具备特定领域知识及规则解读能力
行业专家	规则制定、冲突仲裁、模型优化
质检工程师	错误模式分析、质量闭环控制

8.1.2 标注阶段

8.1.2.1 试标注

选取部分具有代表性的语料进行试标注。根据试标注过程中遇到的问题，调整和完善标注指南。

8.1.2.2 正式标注

将语料合理分配给标注人员，由标注人员仔细阅读语料，识别其中的实体、关系、事件等信息，并按照规定的标签体系进行标注。标注人员应定期检查标注进度和质量，及时发现并纠正问题。

8.1.3 审核阶段

审核阶段包含标注员自查、标注者间质检、行业专家审核。

8.1.3.1 标注员自查

由标注员对所完成标注内容进行自查，要求对照标注指南质检，覆盖全部标注数据。

8.1.3.2 标注者间质检

通过计算标注者间一致性指标评估不同标注者之间的标注结果是否一致。如果一致性较低，需要分析原因，并组织标注人员进行讨论，对不一致的标注结果进行修正。

8.1.3.3 行业专家审核

由行业专家对标注结果提出反馈意见，指出存在的问题和改进建议。标注人员根据专家的反馈，对标注结果进行修正。

8.1.3.4 内容安全性审核

对危害人类健康及社会稳定的数据内容应当重点审核标注，明确其危险性。必要时需要将危险信息从语料中剔除。

8.1.4 优化阶段

8.1.4.1 标注指南更新

根据审核过程中发现的问题，对标注指南进行更新和优化，以提高后续标注工作的质量和效率。

8.1.4.2 数据清洗与整合

检查标注结果，去除错误标注和重复数据，将标注结果进行整合，形成统一的标注语料库。

8.1.5 交付阶段

8.1.5.1 标注结果交付

提供标注语料的格式说明，同时提供标注指南和相关说明。

8.1.5.2 反馈与改进

根据语料使用方的反馈，对标注过程和结果进行改进和优化。

8.2 学科特色数据标注

依据不同学科的特色知识内容和需求针对性调整标注流程和步骤。各类学科的标注体系应通过具备学科特色的场景验证与动态迭代，确保持续适应学科领域的科研需求。

8.3 人工标注

8.3.1 SFT 语料标注

a) 由人工标注员基于行业知识提出问题，经专业团队审核，确保问题的专业有效，明确问题的背景和核心诉求；

b) 在自动化标注工具生成答案后，人工标注员对答案进行打分，并对概念、逻辑、知识点等内容进行校正修改，经专业团队审核，确保答案符合科学逻辑；
c) 循环修正检查后将问答对提交至自动标注工具，优化后续标注工作。

8.3.2 CoT 语料标注

a) 由人工标注员基于行业知识提出强推理问题，经专业团队审核确保问题专业有效，明确问题的背景和核心诉求；
b) 由人工标注员撰写逻辑清晰的推理步骤和回答，其中回答应当与思维链的推理步骤一一对应。经专业团队审核，确保答案符合科学逻辑；
c) 将含有推理过程的问答对提交至自动标注工具，优化后续标注工作。

9 科学智能语料测试

9.1 通用测试内容

9.1.1 输入输出规范性检测

验证语料存储格式是否符合领域标准，并检查元数据完整性。

9.1.2 科学概念一致性检测

通过领域知识图谱或本体库匹配语料中的专业术语，验证术语的准确性和上下文一致性。

9.1.3 多模态关联性检测

测试跨模态数据的对齐能力，确保多模态数据在模型训练中的协同有效性。

9.1.4 模型输入压力检测

将语料输入目标模型，监测异常输出并溯源至语料缺陷，识别语料中隐含的噪声或逻辑冲突。

9.1.5 语料分布及覆盖性检测

统计语料对于真实科研场景的覆盖度，确保各类语料的比例分布符合实际科研场景，并验证罕见样本是否纳入语料库。

9.1.6 语料价值对齐检测

测试科学智能语料的价值倾向，确保语料符合科学伦理和社会道德。

9.2 学科特色测试

9.2.1 学科专用指标设计

依据不同学科类型确定核心可量化的检测指标。

9.2.2 学科领域知识图谱验证

a) 检查知识图谱的拓扑结构合理性，包括实体关系路径连通性、属性值完整性、层级关系合理性；
b) 验证实体关系是否符合领域规则、属性值范围合理性；
c) 解决跨来源数据的冲突，需实体对齐和关系消歧。

9.2.3 学科知识动态更新验证

科学智能语料库须依照所在学科的发展进度动态更新，须追踪知识版本迭代时的逻辑延续性，记录变更溯源链。

9.2.4 行业企业应用检测

针对行业重点企业定期调查科学智能语料库在实际应用中的可用度并收集用户反馈，依据行业应用情况验证科学智能语料库的有效性。

9.3 人工检测

9.3.1 人工检测流程

科学智能语料的人工检测主要包括以下内容：

a) 初级检测员检查语料的基础错误并开展一致性验证，生成问题清单并注明错误类型；
b) 同一语料内容由至少两位高级检测员独立检测，重点对科学智能语料的逻辑性矛盾和科学事实符合情况进行验证，若低于既定阈值需要启动质量仲裁；
c) 专家对交叉审核中的分歧样本进行质量仲裁，针对高频错误类型针对性对语料提出优化建议。

9.3.2 检测人员资质要求

科学智能语料检测需结合领域知识和技术能力，对检测人员进行分层管理：

a) 初级审核员须具备相关学科教育背景或1-3年从业经验，主要执行初审环节；
b) 高级审核员须具备3年以上相关领域工作经验，具备相关专业能力认证，主导交叉审核环节；
c) 专家审核员需具备相关学科领域的专业权威性，具备高级职称或博士学历，且拥有5年以上行业内工作经验，且发表过相关领域论文或参与标准制定，主要负责终审仲裁环节。

10 科学智能语料使用

10.1 科研大模型

10.1.1 预训练

基于跨学科文献、结构化数据库与非结构化科学文本等语料，通过多模态对齐预训练任务、自监督学习，构建通用科学语义表征，帮助模型构建对于科学领域的基础理解，支撑跨任务迁移学习、降低领域适配成本。

10.1.2 领域适应

基于具体科学领域的语料和学科领域知识图谱，优化模型的专业知识表达能力，实现专业术语的精确解析，减少跨学科语义鸿沟。

10.1.3 指令微调

基于人工撰写的科研任务指令-响应问答对、实验标准流程等语料，训练模型拆解复杂科学问题和操作指令，提升模型对齐科研任务、执行复杂操作的能力。

10.1.4 推理优化

基于经过标注逻辑关系和推理过程的CoT语料，训练模型生成可追溯的推理步骤，使其能输出具备因果关联的科研推论，并通过对抗语料训练抑制事实性幻觉。可基于故意植入科学谬误的文本及修正说明等语料，提升模型对于错误推理的敏感性。

10.1.5 模型测评

基于行业专家验证语料、多源交叉验证语料等模型评测语料，评估模型在科学问答、文献综述生成等任务中的事实准确性与可靠性，训练模型识别矛盾信息的能力。基于对抗语料的训练增强模型鲁棒性。

10.2 科研数据库与学科知识库

10.2.1 科学数据库

科学智能语料可用于科学数据库建设的以下方面：

a) 基于多模态科研数据整合，实现跨模态对齐，支持跨模态检索，确保数据调用时语义完整性，避免信息割裂；

b) 基于高质量的真实科学智能语料，通过仿真计算进一步生成新的符合科学逻辑的语料，扩充原有数据库的内容和维度；
c) 通过对特定领域科学智能语料的深度学习和知识图谱构建，揭示科学领域内实体之间复杂的、隐含的关联和模式，同时对新兴交叉学科领域的数据库结构提出初始建议。

10.2.2 科学知识库

科学智能语料可用于科学知识库建设的以下方面：
a) 对科学知识进行实体识别、关系抽取、事件抽取、属性抽取等操作；
b) 对科学知识进行跨来源冲突消解、多模态知识对齐、层级关系和跨本体映射发现、知识图谱填补；
c) 基于科学智能语料背景知识、前提假设和科学原理，关联知识库片段提供隐含知识上下文；
d) 对已有科学知识库进行信息抽取、关系抽取、事件抽取，科学模型可从最新语料中识别新发现、新数据、新关系、新结论。

10.3 科研智能体

10.3.1 文献阅读

科研智能体可以通过科学领域认知、精准定位和语义搜索、批判评估与趋势挖掘，有效提升文献阅读的效率和质量，其中科学智能语料的主要应用包括以下方面：
a) 高质量科学智能语料库为智能体提供预训练的素材，使其对科学智能的基础理解能力提升；
b) 高质量科学智能语料库可训练智能体较强的语义理解与检索能力，可根据研究目标精准定位相关文献段落、图表、方法描述，提升跨文献关联能力；
c) 基于科学智能语料库中蕴含的方法对比和科学逻辑，智能体可评估文献的可靠性、创新性、局限与短板等，并依据相关文献的研究内容分析研判未来研究趋势。

10.3.2 仿真计算

科研智能体可融合历史仿真数据与物理规则语料，优化参数搜索空间，实现仿真收敛速度提升，并保障计算结果的物理合理性，其中科学智能语料的主要应用包括以下方面：
a) 科学智能语料库中涵盖大量以往计算案例及相关参数配置，科研智能体通过对语料库的分析，可定制化推荐适合当前计算问题的过往方法、参数范围、减少计算试错成本并提升效率；
b) 科研智能体可基于科学智能语料库中的计算知识优化计算方案的具体步骤和流程，并根据相关科研文献中的通用规则和理论预期，识别计算异常并提出优化建议；
c) 科学智能体可理解现有计算逻辑，生成计算任务的代码片段，同时从科学智能语料库中寻找其它符合场景逻辑的计算代码片段，提升开发效率。

10.3.3 实验设计

科研智能体可输入文献方法描述语料，解析实验参数约束，生成可复现的操作指令集，进而生成具备实操性的实验方案，其中科学智能语料的主要应用包括以下方面：
a) 基于科学智能语料库中以往成功实验方案数据，以及其它详细的实验过程参数，科研智能体可依据实验目标提供实验初步方案，并结合相关文献材料对已有试验方案提出优化方向建议；
b) 科研智能体可理解学习科学智能语料库存储的实验步骤、实验参数、仪器操作、安全事项等信息，辅助实验操作者执行实验，并全程实时监控实验过程数据，预警潜在问题；
c) 针对科学实验产生大量原始数据，科研智能体可利用科学智能语料库中学习到的分析方法、特征识别、实验规律等知识对实验数据进行快速处理分析，提供初步实验结果解读；
d) 针对失败或结果误差较大的实验，科研智能体可通过分析科学智能语料库中记录的类似失败案例及其参数记录，辅助推测实验失败的原因并提出改进建议。

11 数据安全

11.1 安全原则

科学智能语料库建设方须依据各个学科特点，围绕组织、制度、人员、平台等方面制定安全策略、执行覆盖数据资源和语料产品全生命周期的安全控制措施。

11.2 安全性评价

数据采集、清洗、标注、测试、使用的全过程均需要安全性评价，考虑的因素包括但不限于内容合规性、数据处理过程安全性、个人信息保护合规性等。应依据GB/T 45577-2025、YD/T 4245-2023规定定期开展数据安全风险评估，同时根据 GB/T 22240-2020对承载数据的系统进行定级备案。

11.3 制度要求

数据采集、加工、测试和使用的全过程考虑建立的制度包括数据管理规范、访问控制及权限管理制度、安全事件应急响应预案和第三方合作安全管理制度等。

11.4 技术要求

科学智能语料库建设在技术层面应确保数据采集、清洗、标注、测试、使用的全周期安全，整体技术体系的设计应遵循GB/T 22239-2019相应级别的要求。

11.4.1 数据处理要求

为保障数据在存储、使用和流转等状态下的机密性、完整性和可用性，应对数据采取数据加密、数据备份与恢复、数据脱敏等处理技术。

11.4.2 访问控制要求

建立严格的身份验证与授权机制，确保任何对数据的访问和操作都经过合法授权、可追溯且符合最小权限原则。

11.4.3 安全审计与监控要求

具备全面的行为记录和实时监控能力，以便于事后追溯、发现潜在威胁和及时响应安全事件。

11.4.4 基础设施安全

进行合理的网络区域划分，并部署防火墙等设备进行边界防护和访问控制，同时对操作系统、数据库及各类应用软件及时安装安全补丁，关闭不必要的服务和端口，并部署主机恶意代码防范软件。

11.4.5 软硬件设施

数据采集、清洗、标注、测试、使用的全过程所涉及信息系统考虑的软硬组件包括但不限于数据网关、数据加密与备份装置、安全防范和监控系统、私域数据存储与传输系统、数据资产管理工具以及满足GB/T 22239-2019所需的其它设备或系统。

11.5 人员要求

参与数据采集、清洗、标注、测试、使用的任何人员都须签署与其职责相对应的保密协议，同时这些人员所属机构须建立独立、专业的信息安全团队。

附 录 A
（资料性）
科学智能代表学科领域语料特点

A.1 生命科学

A.1.1 数据内容及特点

A.1.1.1 数据内容

生命科学数据本质上是多模态和多组学的，旨在全面捕捉从微观分子到宏观表型的生命活动信息。生命科学领域的科学智能数据主要包括以下几种类型：

a) 序列数据：包括DNA序列、RNA序列等，用于研究基因的结构、功能和表达调控；
b) 结构数据：涉及蛋白质的三维空间结构，如蛋白质晶体结构、核磁共振结构等；
c) 生物医学影像数据：包括但不限于磁共振成像、计算机断层扫描等，用于诊断和研究疾病；
d) 临床试验数据：包括患者的病历、症状、治疗反应、药物剂量等，用于评估药物疗效和安全性；
e) 生理信号数据：如心电图、脑电图、肌电图等，用于监测生物体的生理状态；
f) 多组学数据：包括基因组学、转录组学等，用于全面研究生物系统的复杂性；
g) 计算仿真数据：通过计算模拟得到的数据，如分子动力学模拟的蛋白质折叠过程、细胞内信号传导路径的模拟、药物与受体的结合动力学等，用于研究生物系统的动态行为和机制。

A.1.1.2 数据特点

生命科学领域中的科学智能数据特点如下：

a) 高维度与复杂性：数据量大，涉及多个维度和复杂的相互关系；
b) 动态性与时效性：生物体的生理状态和疾病发展是动态变化的，数据需要及时更新和分析；
c) 异构性与多样性：不同来源和类型的生物数据格式和标准不一致，需要进行整合和标准化处理；
d) 噪音和不确定性： 生物系统的特异性和实验测量过程的人为误差导致数据存在噪音；
e) 多尺度性：数据设计从分子尺度（如基因，蛋白质)到细胞尺度（如细胞功能，细胞间信号传导）再到组织和器官尺度（如器官功能、疾病理论）的多层次信息数据；
f) 隐私性与敏感性：涉及个人健康信息，需要严格保护数据隐私。

A.1.1.3 数据挑战

生命科学领域中的科学智能数据面临如下挑战：

a) 高维度与小样本问题：即数据特征（如基因）数量远超样本（如患者）数量，对算法的稳健性和泛化能力提出了极高要求；
b) 数据内在关联性问题：生命系统本身是一个复杂的调控网络，因此数据特征之间存在着复杂的内在关联性，这要求算法能够超越简单的统计相关，去捕捉非线性的生物学因果关系；
c) 标准化和可追溯问题：由于实验过程中的技术偏差，所有数据必须高度依赖严格的标准化流程和可追溯性，确保数据质量和研究可复现性。

A.1.2 数据生产加工环节特色

A.1.2.1 数据采集

a) 特色内容：生命科学领域的数据采集具有高度的专业性和复杂性，与通用科学智能数据采集存在显著差异。
   1) 多源异构数据融合：生命科学数据来源广泛，包括基因测序仪、质谱仪、显微镜、电子病历系统等。这些数据不仅类型多样，如文本、图像、序列、数值等，而且格式和标准也各不相同。例如，基因测序数据通常以特定专业格式存储，而电子病历数据则以结构化表格或非结构化文本形式存在。为了充分利用这些数据，需要开发专门的工具和方法来整合多源异构数据，实现数据的无缝对接和交互。例如，通过建立统一的数据仓库，将不同来源的数据进行标准化处理，转换为统一的数据模型，以便后续的分析和挖掘；

2) 生物样本的特殊处理：在采集生物样本数据时，需要严格遵循生物安全和伦理规范。例如，在采集人体组织样本时，必须获得患者的知情同意，并确保样本的采集、运输和存储过程符合生物安全标准，以防止样本污染和交叉感染。此外，生物样本的处理还需要考虑其生物学特性，如细胞的活性、蛋白质的稳定性等。例如，在提取蛋白质样本时，需要在低温环境下操作，并使用适当的缓冲液和保护剂，以保持蛋白质的结构和功能完整性；
3) 动态监测与实时反馈：生命科学中的许多研究对象是动态变化的生物系统，如细胞的生长、分化，疾病的进展等。因此，数据采集需要具备动态监测和实时反馈的能力。例如，在细胞培养过程中，通过实时荧光显微镜成像技术，可以动态观察细胞的形态变化、细胞周期进程以及细胞间相互作用等。这些实时数据可以为研究人员提供即时反馈，帮助他们及时调整实验条件，优化实验方案。同时，动态监测数据也为研究生物系统的动态行为和机制提供了重要依据，如通过时间序列分析，可以揭示基因表达的动态模式和调控网络。

b) 数据采集核心准则
1) 必须实施稳健的质量控制样本策略，例如使用混合样本以在整个分析过程中持续监控仪器性能的稳定性，包括信号强度、保留时间等关键指标。
2) 必须严格控制和记录从样本采集到处理的每一个分析前步骤，以最大限度地减少人为引入的变异；推荐使用内标对样品提取效率和分析过程中的基质效应进行校正，并定期分析空白样本以识别和控制系统污染及样品间的交叉污染；
3) 对于依赖扩增步骤的技术，必须评估并控制可能引入的扩增偏好，因为它会影响定量的准确性和下游分析的可靠性。

c) 实验数据采集的质量控制：许多在数据分析阶段难以解决的问题源于上游实验设计的缺陷。因此数据质量控制须在实验设计阶段开始。
1) 在任何数据采集开始之前，必须进行前瞻性的、周密的实验设计。这包括明确定义科学问题、计算并确定足够的样本量、合理设置生物学重复和技术重复、规划对照组，并采用随机化策略来分配样本的处理顺序和位置，以最大限度地减少已知的和未知的混杂因素；
2) 实验设计必须避免生物学变量与技术变量（如处理批次）完全混淆。例如，将所有处理组样本在一个批次中完成，而所有对照组样本在另一个批次中完成，这种设计将导致生物学信号与批次效应无法通过计算方法分离，从而使产生的数据失去科学价值；
3) 在数据采集的同时，必须强制要求同步记录全面、详尽的元数据。元数据的记录应遵循国际公认的“最低信息”（Minimum Information, MI）标准，以确保实验的可解释性、可重复性和可验证性。

表 A.1 生命科学代表性实验数据质量控制准则

代表性实验	数据质量控制遵循原则
通用高通量测序实验	遵循 MINSEQE (Minimum Information about a high-throughput SEQuencing Experiment) 指南，记录样本描述（物种、组织、细胞类型）、样本处理方法、实验变量、文库制备策略、测序平台与参数、以及数据处理流程等核心信息
基因表达微阵列实验	遵循 MIAME (Minimum Information About a Microarray Experiment) 指南
蛋白质组学实验	遵循 MIAPE (Minimum Information About a Proteomics Experiment) 指南，详细记录样本制备、分离技术、质谱参数、数据分析软件和参数等
单细胞实验	遵循 MinSCe (Minimum Information about a Single-Cell Experiment) 指南。该指南是对MINSEQE的扩展，额外要求记录单细胞悬液制备方法、细胞解离方案、细胞活力评估、单细胞捕获技术（如微流控或FACS）、细胞条形码（barcode）和唯一分子标识符（UMI）策略等特有元数据

表 A.2 生命科学多组学领域数据要求

组学领域	主要生成技术	原始数据格式	处理后数据格式	注释/结果格式
基因组学	NGS (Illumina, etc.)	FASTQ	BAM/SAM (比对后)	VCF (变异), BED/GFF/GTF (区域注释)
转录组学	NGS (RNA-Seq, scRNA-Seq)	FASTQ	BAM/SAM (比对后), Expression Matrix (定量后, e.g., CSV, TSV)	GFF/GTF (可变剪接), Differential Expression Lists
蛋白质组学	Mass Spectrometry (MS/MS)	mzML (推荐标准)	mzIdentML (mzi (鉴定结果)	Protein/Peptide Quantification Matrix (定量结果), PSI-MI XML (相互作用)
代谢组学	MS, NMR	mzML (MS), nmrML (NMR)	Peak List/Intensity Table (CSV, TSV)	Compound Identification List (with confidence levels), mzTab-M
医学影像组学	CT, MRI, PET	DICOM (临床标准)	NIfTI (研究常用, 简化元数据)	Feature Matrix (提取的影像特征)

d) 数据采集案例：某研究团队正在进行一项关于阿尔茨海默病的多组学研究，旨在通过整合基因组学、蛋白质组学和生物医学影像数据，探索疾病的早期诊断标志物和发病机制。

表 A.3 生命科学数据采集案例

步骤	具体做法
基因测序数据采集	使用RNA-Seq技术对患者的脑组织样本进行转录组测序，获取基因表达数据
生物医学影像数据采集	使用3T MRI对患者的脑部进行成像，获取高分辨率的结构影像和功能影像数据 使用PET扫描对患者的脑部进行代谢成像，获取葡萄糖代谢数据
临床试验数据采集	通过电子病历系统（EMR）和临床试验管理系统（CTMS）记录患者的详细病历信息，包括症状、认知评分、药物使用等 通过定期随访收集患者的治疗反应和病情进展数据
生理信号数据采集	使用多导联心电图（ECG）和脑电图（EEG）设备，记录患者的生理信号数据，用于评估患者的心脏和大脑功能

A.1.2.2 数据清洗

a) 特色内容：生命科学数据的清洗不仅要处理通用的噪声和错误，还要针对其特有的数据形式和内容进行特殊处理。
   1) 数据标准化与规范化：生命科学领域中，数据标准化与规范化是确保数据质量与可用性的关键环节。由于生命科学数据来源广泛、格式多样，且涉及多维度的生物学信息，因此需要通过一系列专门的处理步骤来实现数据的统一与一致性。数据标准化涉及将不同来源的数据转换为统一的格式和编码，例如将基因测序数据从 FASTQ 格式转换为 BAM 格式，或将临床数据中的文本描述转换为标准化的医学编码（如 ICD-10）；
   2) 生物数据的复杂噪声处理：生命科学数据中存在多种复杂的噪声，如基因测序中的随机测序错误、生物医学影像中的伪影和运动伪迹等。这些噪声不仅影响数据的质量，还可能掩盖重要的生物学信息。因此，需要开发专门的算法和工具来识别和去除这些复杂噪声。例如，在基因测序数据中，可以使用基于贝叶斯统计的错误校正算法，通过建模测序错误的概率分布，对低质量的测序读段进行校正。在生物医学影像数据中，可以采用基于深度学习的去伪影算法，通过训练卷积神经网络来识别和去除影像中的伪影和运动伪迹；
   3) 数据一致性校验：由于生命科学数据涉及多个维度和层次，不同来源的数据之间需要进行一致性校验。例如，在多组学研究中，基因表达数据和蛋白质表达数据之间应具有一定的相关性。如果发现两者之间存在显著的不一致性，可能意味着数据中存在错误或偏差。因此，需要开发专门的校验方法，通过比较不同数据集之间的相关性，检测和纠正数据中的不一致性。例如，可以使用基于主成分分析的方法，对多组学数据进行降维处理，然后通过比较不同数据集在低维空间中的分布情况，识别和校正数据中的偏差；

4) 生物数据的动态更新与迭代清洗：生命科学数据通常是动态变化的，随着研究的深入和新数据的不断产生，需要对数据进行动态更新和迭代清洗。例如，在疾病研究中，随着对疾病机制的进一步了解，可能会发现新的基因或蛋白质与疾病相关。这时，需要将这些新发现的数据纳入到现有的数据集中，并重新进行清洗和分析。同时，迭代清洗还可以根据新的分析结果，进一步优化清洗策略，提高数据的质量。例如，通过机器学习算法，可以根据已清洗数据的特征，自动调整清洗参数，实现对新数据的高效清洗。

b) 数据清洗案例：某研究团队在进行一项关于肿瘤基因组学的研究，旨在通过分析肿瘤组织样本的基因测序数据，识别与肿瘤发生和进展相关的基因变异。

表 A.4 生命科学数据清洗案例

步骤	具体做法
基因测序数据清洗	使用FastQC工具对测序数据进行质量评估，生成质量报告 使用Trimmomatic工具去除低质量的测序读段和接头序列，提高数据质量 使用LoFreq工具对校正后的数据进行变异检测，识别低频突变并校正测序错误
生物医学影像数据清洗	使用NL-means去噪算法对MRI影像数据进行去噪处理，减少影像中的随机噪声 使用基于深度学习的去伪影算法（如DeepMRI）对影像中的伪影和运动伪迹进行识别和去除
临床试验数据清洗	使用Pandas库对临床数据进行预处理，填补缺失值，使用均值填补方法对数值型变量进行填补，使用K近邻填补方法对分类变量进行填补 使用基于统计学的异常值检测方法（如Z-score、IQR）识别并处理异常值，确保数据的准确性

c) 数据清洗通用准则：数据清洗的核心目标是将原始数据转化为可进行生物学解释的、标准化的定量矩阵。此过程应遵循以下通用原则：

1) 必须遵循领域内公认的最佳实践工作流程，并对多样本队列研究采用联合分析策略，以增强信号检测的灵敏度并有效缓解样本间的批次效应；
2) 数据分析必须基于权威、全面且版本一致的参考数据库（如基因组、蛋白质序列库），并在数据库中包含常见污染物序列以供筛查。在结果认定时，必须在多个层级上应用严格的统计控制，例如将假发现率控制在 1%或 5%等公认阈值以下；
3) 针对仪器信号漂移等技术性变异，必须利用贯穿实验的质量检测样本进行有效的计算校正，以保证最终定量结果的准确性和可比性。

A.1.2.3 数据标注

a) 特色内容：生命科学数据的标注不仅需要专业知识，还要考虑其生物学意义和复杂性。其特色内容如下：

1) 专业领域知识的深度标注：生命科学数据的标注需要具备深厚的生物学专业知识，以确保标注的准确性和可靠性。例如，在基因序列数据的标注中，需要对基因的功能、表达调控、变异类型等进行详细标注。这不仅要求标注人员熟悉基因组学的基本知识，还需要了解特定基因在生物学过程中的具体作用。在蛋白质结构数据的标注中，需要对蛋白质的三维结构、活性位点、结合位点等进行精确标注，这需要标注人员具备结构生物学的专业背景。因此，生命科学数据的标注通常需要由专业的生物学家或生物信息学家来完成，或者在标注过程中引入领域专家的知识和经验；
2) 多层次标注体系：生命科学数据具有多层次的结构和信息，因此需要建立多层次的标注体系。例如，在生物医学影像数据的标注中，不仅要对病变区域的位置、大小、形状等进行标注，还需要对病变的类型、分级、发展阶段等进行标注。在多组学数据的标注中，需要对基因、转录本、蛋白质、代谢物等不同层次的分子进行标注，并建立它们之间的关联关系。这种多层次标注体系可以为后续的分析和建模提供更丰富的信息，帮助研究人员更全面地理解生物系统的复杂性。例如，通过多层次标注，可以构建基因-蛋白质-代谢物的调控网络，揭示生物系统的分子调控机制；
3) 动态标注与更新：与静态数据标注不同，生命科学数据的标注需要根据研究进展和新知识进行动态更新。例如，随着对某种疾病的认识不断深入，可能会发现新的生物标志物或疾病机制。这时，需要对已标注的数据进行重新评估和更新，以反映最新的研究成果。同时，

动态标注还可以根据新的实验数据和分析结果，对标注内容进行优化和完善。例如，在药物研发过程中，随着对药物靶点的进一步研究，可能需要对药物与受体结合位点的标注进行更新，以指导药物的优化设计。

b) 数据标注案例：某研究团队在进行一项关于心血管疾病的多模态数据研究，旨在通过整合基因组学、生物医学影像和临床数据，开发疾病诊断和预后模型。

表 A.5 生命科学数据标注案例

步骤	具体做法
生物医学影像数据标注	由专业的放射科医生对MRI影像中的病变区域进行标注，包括病变的位置、大小、边界等信息。 使用3D Slicer软件进行精确标注，确保标注的准确性和一致性。
基因测序数据标注	使用SnpEff工具对基因变异进行功能影响预测和标注，包括变异类型、位置、功能影响等信息。 通过与已知的基因变异数据库（如ClinVar）进行比对，验证变异标注的准确性。
临床试验数据标注	对患者的症状、治疗反应、药物剂量等进行详细标注，确保数据的可读性和可用性。 使用结构化的标注模板，确保标注的一致性和标准化

A.1.2.4 数据测试

a) 特色内容：生命科学数据的测试需要考虑其生物学背景和实验验证的复杂性。

1) 生物学实验验证：生命科学数据的测试通常需要通过生物学实验进行验证。例如，对于基因序列数据中预测的基因功能，需要通过基因敲除、基因过表达等实验方法进行功能验证。对于生物医学影像数据中开发的疾病诊断模型，需要通过临床试验进行验证，评估模型的准确性和可靠性。这些实验验证不仅需要严格的设计和实施，还需要考虑实验条件的控制、样本的选择和统计分析等因素。例如，在进行基因功能验证实验时，需要选择合适的细胞系或动物模型，控制实验条件的一致性，并进行多次重复实验以确保结果的可靠性。同时，实验验证的结果还需要与数据测试的结果进行对比和分析，以评估数据测试方法的有效性和准确性；

2) 多维度性能评估：生命科学数据的测试需要从多个维度进行性能评估。例如，在疾病诊断模型的测试中，不仅要评估模型的准确性、敏感性和特异性，还需要考虑模型的泛化能力、鲁棒性和临床实用性。对于药物研发中的数据测试，需要评估药物的活性、选择性、毒性等多个指标。这些多维度的性能评估可以为研究人员提供更全面的评估结果，帮助他们更好地选择和优化数据测试方法。例如，通过构建多维度的性能评估指标体系，可以对不同的疾病诊断模型进行综合评估，选择最适合临床应用的模型；

3) 动态测试与反馈机制：生命科学数据的测试需要建立动态测试和反馈机制。随着研究的进展和新数据的产生，需要对数据测试方法进行动态调整和优化。例如，在疾病研究中，随着对疾病机制的进一步了解，可能需要对疾病诊断模型的测试方法进行调整，以更好地反映疾病的生物学特征。同时，测试结果也需要及时反馈给研究人员，以便他们根据测试结果调整研究方向和策略。例如，通过建立实时反馈系统，研究人员可以在数据测试过程中及时获取测试结果，并根据结果调整模型参数或实验方案，提高研究效率和质量。

b) 注释与语义富集：标准化的定量数据需通过注释和语义富集转化为具有生物学意义的知识。此过程的主要原则有：

1) 应对语料库中的核心特征（如基因、变异、分子）进行全面的功能注释，内容应涵盖其物理位置、预测的生物学功能、在参考群体中的频率，以及与已知表型（特别是临床表型）的关联，并链接到权威数据库（如 ClinVar, UniProt）；

2) 应将鉴定出的差异分子列表映射到生物学通路数据库（如 GO, KEGG, Reactome），并执行功能富集分析，以从系统层面揭示受扰动的核心生物学过程和信号通路；

3) 对于需要结构验证的分子（如代谢物），必须根据社区公认的置信度标准进行分级报告，明确区分经过化学标准品验证的“已鉴定”实体和仅通过数据库谱图匹配的“假定注释”实体。

c) 数据测试案例：某研究团队开发了一个基于基因测序数据和生物医学影像数据的心血管疾病诊断模型，旨在通过多模态数据融合提高疾病的诊断准确性。

表 A.6 生命科学数据测试案例

步骤	具体做法
基因测序数据测试	通过实验验证基因测序数据中预测的基因功能是否准确，使用CRISPR-Cas9基因编辑技术进行功能验证。 通过与已知的基因变异数据库（如ClinVar）进行比对，验证变异检测的准确性。
生物医学影像数据测试	通过临床试验验证基于生物医学影像数据开发的诊断模型的有效性，使用交叉验证方法评估模型的准确性和泛化能力。 通过与已知的正常脑部影像进行比对，验证去噪和去伪影处理的效果。
多模态数据融合模型测试	使用多中心临床试验数据对多模态数据融合模型进行验证，评估模型的准确性和泛化能力。 通过与单一模态数据模型进行对比，验证多模态数据融合模型的优势

A.1.3 语料使用及特色

A.1.3.1 使用方向

生命科学数据的使用需要紧密结合生物学研究目标和实际应用需求。其使用方向包括但不限于：

a) 个性化医疗与精准医学：生命科学数据的一个重要应用方向是个性化医疗和精准医学。通过分析患者的基因序列、生物医学影像、临床病历等多维度数据，可以为患者提供个性化的诊断和治疗方案。例如，基于基因测序数据的肿瘤诊断和治疗方案推荐系统，可以根据患者的基因变异特征，预测患者对不同药物的敏感性，为患者选择最合适的治疗方案。这种个性化医疗模式不仅可以提高治疗效果，还可以减少药物副作用，提高患者的生活质量。同时，精准医学还需要考虑患者的个体差异，如年龄、性别、家族病史等因素，进一步优化治疗方案；

b) 药物研发：借助科学智能语料训练的大模型，可以预测药物的活性、毒性和药代动力学等特性，加速药物研发进程。例如，在药物筛选阶段，通过分析大量的化合物结构数据和已有的药物活性数据，模型可以预测新化合物的潜在活性，从而帮助研究人员快速筛选出有潜力的药物候选分子。实操路径是将药物研发过程中的相关数据（如化合物结构、生物活性实验数据等）输入到模型中，模型会基于语料中的知识进行分析和预测，为药物研发人员提供决策支持；

c) 生物系统的建模与模拟：生命科学数据可以用于构建生物系统的数学模型和计算机模拟。例如，通过分析基因表达数据和蛋白质相互作用数据，可以构建基因调控网络模型，模拟基因表达的动态过程和调控机制。通过分析生物医学影像数据和生理信号数据，可以构建器官系统的生理模型，模拟器官的生理功能和疾病发生发展过程。这些模型和模拟不仅可以帮助研究人员更好地理解生物系统的复杂性，还可以为疾病的预防、诊断和治疗提供理论支持。例如，通过模拟肿瘤的生长和转移过程，可以为肿瘤的早期诊断和治疗提供新的思路和方法；

d) 因功能预测与疾病关联研究：利用科学智能语料中的基因序列和功能注释数据，可以预测基因的功能以及基因与疾病之间的关联。例如，通过分析基因序列的相似性和已知基因功能数据，模型可以预测未知基因可能的功能。在疾病关联研究中，结合临床数据和基因数据，可以挖掘出与疾病相关的基因变异。实操路径是将基因数据和临床数据输入到模型中，模型会根据语料中的知识进行关联分析，找出可能与疾病相关的基因变异位点，为疾病的遗传学研究提供线索。

A.1.3.2 使用案例

某研究团队在进行一项关于阿尔茨海默病的多组学研究，旨在通过整合基因组学、蛋白质组学和生物医学影像数据，开发疾病早期诊断模型和治疗方案推荐系统。

表 A.7 生命科学数据使用案例

步骤	具体做法
疾病诊断模型开发	使用基于生物医学影像数据和基因测序数据训练的大模型，开发AD早期诊断模型。
治疗方案推荐系统开发	借助科学智能语料训练的大模型，开发治疗方案推荐系统。 通过分析患者的基因变异、临床症状和治疗反应，为患者提供个性化的治疗方案。
生物医学研究	利用科学智能语料训练的知识库，为研究人员提供生物医学知识查询和分析工具。 通过查询基因功能、疾病机制等相关知识，为科研人员提供研究思路和参考依据。

A.1.4 数据安全措施

典型的生命科学领域数据安全措施包括但不限于：

a) 人类遗传信息保护措施：为避免个人遗传信息隐私问题，应建立多层次身份认证机制，实施基于角色的访问控制，对所有遗传数据进行假名化或匿名化处理，并采用强加密技术保护数据传输和存储安全；

b) 实验数据完整性与保密性保障：为保护新药或新疗法临床试验数据的知识产权，应与所有接触数据的相关方签订严格的保密协议，并建立不可更改的审计日志，详细记录所有数据访问和修改行为，确保全流程可追溯。同时建立数据备份与恢复机制，采用数字签名技术确保数据完整性；

c) 生物多样性与物种遗传资源保护：针对特定物种基因组信息和地理分布数据，应建立分级授权机制，根据数据敏感级别设置不同访问权限，建立数据使用审批流程，并与相关科研机构签署数据共享协议，明确使用范围和责任边界。

A.2 物质科学

A.2.1 数据内容及特点

在电池材料科学领域，科学智能数据是推动该领域研究和应用的关键资源。

A.2.1.1 实验数据

实验数据是电池材料科学研究中最基础且重要的数据来源。它涵盖了从材料合成到性能测试的各个环节，为研究人员提供了直接的实证依据。

a) 数据类型

1) 材料合成数据：包括材料的制备方法（如溶胶-凝胶法、共沉淀法等）、反应条件（温度、压力、时间等）、原料配比等信息,对于理解材料的形成过程和优化合成工艺至关重要；

2) 结构表征数据：通过 X 射线衍射（XRD）、透射电子显微镜（TEM）、扫描电子显微镜（SEM）等技术获得的材料晶体结构、形貌等信息。这些数据有助于揭示材料的微观结构与宏观性能之间的关系；

3) 电化学性能数据：如比容量、循环稳定性、倍率性能、内阻等。这些数据直接反映了电池材料在实际应用中的性能表现，是评估材料优劣的关键指标；

4) 热性能数据：包括材料的热稳定性、热膨胀系数等，这些数据对于电池的安全性和可靠性具有重要意义。

b) 数据特点

1) 高精度：实验数据通常需要高精度的仪器设备来采集，以确保数据的准确性。例如，高分辨率的 XRD 仪器可以精确测量材料的晶体结构参数；

2) 多维度：电池材料的性能受到多种因素的影响，如材料的成分、结构、制备工艺等。因此，实验数据具有多维度的特点，需要综合考虑这些因素来分析数据；

3) 与实验条件密切相关：实验条件的变化（如温度、压力、合成方法等）会导致实验数据的显著差异。因此，在记录和分析数据时，必须详细记录实验条件，以便进行准确的数据解读和重复实验。

A.2.1.2 模拟数据

模拟数据是通过计算模拟方法得到的数据，它在电池材料科学中有着大量的使用，尤其是在理论研究和材料设计阶段。

a) 数据类型
   1) 密度泛函理论（DFT）计算数据：用于预测材料的电子结构、能带结构、态密度等，帮助理解材料的电化学性能和化学稳定性；
   2) 分子动力学（MD）模拟数据：用于研究材料在原子尺度上的动态行为，如原子间相互作用、扩散系数等，有助于预测材料的热性能和机械性能；
   3) 蒙特卡洛（MC）模拟数据：用于研究材料的统计性质，如相变、吸附行为等，适用于复杂体系的模拟。
b) 数据特点
   1) 基于理论模型和计算参数：模拟数据的生成依赖于特定的理论模型和计算参数。不同的模型和参数设置会影响模拟结果的准确性。因此，在使用模拟数据时，需要对模型和参数进行仔细选择和验证；
   2) 预测性和假设性：模拟数据具有一定的预测性，可以为实验研究提供理论指导和方向。然而，由于模型的假设和近似，模拟结果可能存在一定的误差，需要通过实验数据进行验证；
   3) 计算成本高：尤其是对于复杂的电池材料体系，如多组分复合材料，计算模拟需要大量的计算资源和时间。因此，在实际应用中，需要合理分配计算资源，优化计算策略。

A.2.1.3 文献数据

文献数据是从学术文献、专利、技术报告等来源提取的关于电池材料的性能、制备方法、应用案例等信息。这些数据为研究人员提供了丰富的背景知识和参考信息。

a) 数据类型
   1) 性能数据：包括材料的比容量、循环寿命、倍率性能等，这些数据通常以表格、图表或文字描述的形式出现在文献中；
   2) 制备方法数据：描述材料的合成方法、工艺参数等，为研究人员提供了具体的实验参考；
   3) 应用案例数据：介绍材料在实际应用中表现和效果，如在电动汽车、储能系统中的应用案例。
b) 数据特点
   1) 来源广泛：文献数据来自不同的研究机构和作者，涵盖了各种研究背景和方法。这种多样性为研究人员提供了丰富的信息来源，但也增加了数据筛选和整合的难度；
   2) 信息碎片化：文献中的数据往往是分散的，需要通过文献综述和数据挖掘等方法进行整合和分析，以形成完整的知识体系；
   3) 质量参差不齐：由于不同研究的质量和标准存在差异，文献数据的质量也参差不齐。因此，在使用文献数据时，需要进行严格的筛选和验证，以确保数据的可靠性和有效性。

A.2.2 数据生产加工环节特色

A.2.2.1 数据采集

在电池材料科学领域，数据采集的特色体现在设备精度和稳定性、试验条件记录及模拟数据采集等方面。

a) 特色操作与注意事项
   1) 设备精度和稳定性：在电池材料实验中，数据采集需确保设备的高精度和稳定性。例如，使用高分辨率的 XRD 仪器采集材料晶体结构数据时，要定期校准设备，避免因设备误差导致数据偏差。对于电化学性能测试，高精度的电化学工作站是必不可少的，需确保其在长时间运行中的稳定性；
   2) 实验条件记录：详细记录实验条件，如温度、压力、合成方法等参数的变化。这些条件对实验结果有显著影响，记录详细信息有助于后续的数据分析和实验重复；
   3) 模拟数据采集：对于模拟数据采集，需合理选择计算模型和参数。不同的模型和参数设置会影响模拟结果的准确性。例如，在 DFT 计算中，选择合适的交换相关泛函和基组是关键。在 MD 模拟中，需合理设置时间步长、温度控制方法等参数。

b) 实操案例：某研究团队在采集锂离子电池正极材料的电化学性能数据时，采用高精度的电化学工作站，设置合适的电流密度和电压范围进行充放电测试。在测试过程中，实时监控设备运行状态，确保数据采集的连续性和准确性。同时，记录详细的实验条件，如温度、湿度等环境参数，以便后续分析数据时考虑这些因素的影响。

表 A.8 物质科学数据采集案例

步骤	具体做法
设备准备	选择高精度的电化学工作站，确保设备经过校准且运行正常。
实验设置	根据材料特性，设置合适的电流密度（如0.1C、1C等）和电压范围（如2.5V-4.2V）。
数据采集	在充放电过程中，实时记录电压、电流、容量等数据，同时记录环境温度、湿度等参数。
数据记录	将采集到的数据存储在电子表格中，并详细记录实验条件和设备状态。

A.2.2.2 数据清洗

数据清洗是确保数据质量的关键步骤，通过去除噪声和错误数据，提高数据的可用性。在电池材料科学领域，数据清洗具有以下特色操作和注意事项：

a) 特色操作与注意事项
   1) 异常值识别与处理：对于实验数据，需识别和处理异常值。例如，在电池循环测试数据中，可能存在因设备故障或操作失误导致的异常容量数据。通过统计分析方法（如箱线图）识别并剔除这些异常值；
   2) 数据一致性检查：文献数据清洗时，要对数据进行去重和一致性检查。由于不同文献可能对同一材料性能的描述存在差异，需统一数据格式和单位，确保数据的一致性和可比性；
   3) 数据完整性检查：检查数据是否完整，是否存在缺失值。对于缺失的数据，需根据实际情况进行合理填充或删除。

b) 实操案例：在处理一批电池材料的循环寿命数据时，研究人员首先绘制数据的箱线图，发现部分数据点明显偏离正常范围。经过排查，发现这些异常数据是由于在实验过程中电极极片未充分压实导致的。于是，将这些异常数据剔除，并重新计算剩余数据的统计指标（如平均值、标准差等），以获得更准确的循环寿命评估结果。

表 A.9 物质科学数据清洗案例

步骤	具体做法
数据可视化	绘制箱线图，观察数据分布情况。
异常值识别	通过箱线图识别出明显偏离正常范围的数据点。
异常值处理	分析异常数据产生的原因，如电极极片未充分压实，将这些异常数据剔除。
数据重新计算	对剩余数据重新计算统计指标，如平均值、标准差等，以获得更准确的循环寿命评估结果。

A.2.2.3 数据标注

在电池材料科学领域，数据标注具有以下特色操作和注意事项。

a) 特色操作与注意事项
   1) 详细标注材料信息：在电池材料性能数据标注中，需明确标注数据对应的材料成分、结构和测试条件。例如，对于不同掺杂元素的电池材料电化学性能数据，要详细标注掺杂元素种类、含量以及测试时的充放电倍率等信息；
   2) 模拟数据标注：对于模拟数据标注，要注明所使用的计算模型和参数设置。这有助于理解数据的来源和可靠性，以及在模型训练时考虑模型偏差对数据的影响；
   3) 分类标注：根据数据的性能特征或应用场景，对数据进行分类标注。例如，将具有高比容量的材料数据标注为“高比容量”类别，便于后续的模型训练和分析。

b) 实操案例：在对一批不同粒径的电池材料颗粒进行性能标注时，研究人员详细记录了每种粒径对应的材料比表面积、晶体结构类型以及测试时的电解液体系等信息。在标注过程中，还对数据进行了分类，将具有相似性能特征的数据归为同一类别，便于后续利用机器学习算法进行性能预测模型的训练。

表 A.10 物质科学数据标注案例

步骤	具体做法
材料信息记录	详细记录每种粒径对应的材料比表面积、晶体结构类型、掺杂元素等信息。
测试条件记录	记录测试时的电解液体系、充放电倍率等测试条件。
材料分类	根据材料的性能特征，如比容量、循环稳定性等，对数据进行分类标注。例如，将比容量高于某一阈值的数据标注为“高比容量”类别。

A.2.2.4 数据测试

在电池材料科学领域，数据测试具有以下特色操作和注意事项。

a) 特色操作与注意事项
   1) 多种测试方法交叉验证：在电池材料数据测试中，需采用多种测试方法进行交叉验证。例如，除了常规的电化学性能测试外，还可结合材料的微观结构表征（如 TEM 观察）来验证数据的可靠性。不同测试方法从不同角度反映材料的性能和特性，通过交叉验证能够提高数据可信度；
   2) 与基准数据对比测试：对于新采集的数据，需与已有的基准数据进行对比测试。例如，将新合成的电池材料性能数据与市场上同类材料的标准性能数据进行对比，评估新数据的合理性。若新数据与基准数据存在较大偏差，需进一步分析原因，可能是实验操作问题或材料本身存在缺陷；
   3) 数据更新与验证：定期更新数据集，以确保模型能够反映最新的材料性能和发展趋势。每次更新后，需对新数据进行验证，确保其质量和可靠性。
b) 实操案例：某研究机构在测试一种新型电池材料的倍率性能时，除了使用常规的电化学工作站进行充放电测试外，还利用SEM观察材料在不同倍率下的微观结构变化。通过对比两种测试结果，发现材料在高倍率下内部结构出现裂纹，这与电化学性能下降的结果相一致。因此，确认了倍率性能数据的可靠性，并据此对材料的结构稳定性进行了进一步优化研究。

表 A.11 物质科学数据测试案例

步骤	具体做法
电化学性能测试	使用电化学工作站进行充放电测试，记录不同倍率下的比容量和循环稳定性数据。
微观结构表征	利用SEM观察材料在不同倍率下的微观结构变化，重点关注材料内部是否存在裂纹等缺陷。
数据对比分析	将电化学性能数据与微观结构表征结果进行对比分析，确认数据的可靠性。
结构优化研究	根据测试结果，对材料的结构进行优化，以提高其在高倍率下的性能表现。

A.2.3 语料使用及特色

A.2.3.1 使用方向

a) 材料性能预测：利用科学智能语料中的性能数据，预测新型电池材料的电化学性能，如比容量、循环稳定性、倍率性能等。这对于加速材料研发进程、降低研发成本具有重要意义。
   1) 数据收集与预处理：收集大量的电池材料性能数据作为语料，对这些数据进行预处理和标注。预处理包括数据清洗、异常值处理、数据归一化等步骤，确保数据的质量和一致性；
   2) 模型训练：利用深度学习算法（如神经网络）训练性能预测模型。选择合适的网络结构和训练参数，通过大量的数据训练模型，使其能够学习材料成分、结构与性能之间的映射关系；
   3) 模型评估与优化：通过交叉验证等方法评估模型的性能，根据评估结果对模型进行优化。优化方法包括调整网络结构、改进训练算法、增加训练数据量等；
   4) 实际应用：在实际材料研发过程中，输入待预测材料的相关参数，模型即可输出其性能预测结果。根据预测结果，指导实验合成和优化，加速材料研发进程。
b) 材料结构优化：利用科学智能语料中的结构信息和性能关联知识，指导电池材料的微观结构设计和优化。这对于提高材料的性能和稳定性具有重要意义。

1) 数据提取与分析：从语料中提取材料结构与性能之间的映射关系，构建结构优化模型。通过数据分析方法（如相关性分析、主成分分析等）提取关键特征，建立结构与性能之间的数学模型；
2) 模型训练与验证：利用机器学习算法（如支持向量机、随机森林等）训练结构优化模型。通过大量的数据训练模型，使其能够根据目标性能要求反向推导出优化后的材料结构参数；
3) 结构优化设计：在实际材料研发过程中，根据目标性能要求，利用模型反向推导出优化后的材料结构参数。结合实验验证，对材料的微观结构进行优化设计；
4) 实验验证与优化：通过实验验证模型预测的结构优化方案，根据实验结果对模型进行调整和优化，确保模型的准确性和可靠性。

c) 故障诊断与预警：基于电池材料在使用过程中的性能数据和故障案例语料，训练智能诊断模型，实现对电池故障的早期诊断和预警。这对于提高电池的安全性和可靠性具有重要意义。
1) 数据收集与预处理：收集电池在不同故障状态下的性能数据（如电压异常、内阻突变等）以及相应的故障原因和处理方法作为语料。对这些数据进行预处理和标注，确保数据的质量和一致性；
2) 模型训练与验证：利用机器学习算法（如支持向量机、决策树、神经网络等）训练故障诊断模型。通过大量的故障数据训练模型，使其能够学习不同故障特征与故障类型之间的映射关系。采用交叉验证等方法评估模型的性能，根据评估结果对模型进行优化；
3) 实时监测与预警：在电池使用过程中，实时监测其性能数据，如电压、电流、内阻等。将监测到的数据输入故障诊断模型，模型能够快速判断故障类型并发出预警信号。根据预警信号，及时采取相应的处理措施，避免故障的进一步扩大；
4) 模型更新与维护：随着电池使用过程中的数据积累和新的故障案例的出现，定期更新故障诊断模型，以提高模型的准确性和适应性。同时，对模型进行维护和优化，确保长期稳定运行。

A.2.3.2 使用案例

某电池制造企业在开发新一代高能量密度电池时，根据电动汽车的实际使用需求，从科学智能数据平台中筛选出与低温性能和快速充电性能相关的电池材料数据。利用这些数据训练性能预测模型，通过模型预测筛选出几种具有潜力的材料组合。结合实验室的小试验证，最终确定了一种满足企业产品要求的新型电池材料配方，并将其应用于新产品开发中。

表 A.12 物质科学数据使用案例

步骤	具体做法
数据筛选	根据电动汽车的实际使用需求，从数据平台中筛选出与低温性能和快速充电性能相关的电池材料数据。
模型训练	利用筛选出的数据训练性能预测模型，选择合适的机器学习算法（如神经网络）进行模型训练。
模型验证	通过实验室的小试验证，对模型预测结果进行验证，确保模型的准确性和可靠性。
产品开发	根据模型预测结果和实验验证，确定新型电池材料配方，并将其应用于新产品开发中。

A.2.4 数据安全措施

典型的物质科学领域数据安全措施包括但不限于：

a) 核心知识产权信息保护：针对新型材料的化学成分、原子结构、合成工艺及加工参数等核心知识产权信息，应建立严格的数据分级分类制度，采用端到端加密技术，实施零信任网络架构，并建立完善的员工背景调查和保密培训制度；
b) 材料性能测试数据完整性保障：为确保材料性能测试数据（力学强度、导电率、疲劳寿命等）的准确性，宜建立多重数据校验机制，采用区块链技术记录数据修改历史，实施实时数据监控和异常检测，并建立数据质量评估体系；
c) 军民两用材料数据管控：针对高性能材料数据的敏感性，应建立出口管制合规检查机制，对数据访问人员进行安全审查，实施地理位置限制和时间窗口控制，并建立定期安全评估和风险监测体系。

A.3 大气海洋科学

A.3.1 数据内容及特点

A.3.1.1 数据内容

大气海洋科学领域中的科学智能数据主要包括以下几种类型：

a) 卫星遥感数据：包括来自各类气象卫星（如风云系列、GOES、Himawari）、海洋卫星（如Jason系列、Sentinel系列）和雷达卫星（如Sentinel-1、TerraSAR-X）的数据。这些数据涵盖了大气温度、湿度、云、降水、气溶胶、海表温度、海面高度、海面风场、海冰、叶绿素浓度等多种地球物理参数。例如，高分辨率的卫星云图和雷达回波图，用于识别和追踪天气系统；卫星高度计数据，用于监测海平面变化和海洋环流；合成孔径雷达（SAR）数据，用于海面风场反演和海冰监测；

b) 全球再分析数据：通过将历史观测数据与数值预报模型同化而生成的全球一致的格点数据。例如，欧洲中期天气预报中心（ECMWF）的ERA5再分析数据、美国国家环境预报中心（NCEP）的CFSR再分析数据等。这些数据提供了长时间序列、高空间分辨率的大气和海洋状态变量，如温度、气压、湿度、风速、海温、盐度、海流等，是气候研究和模式验证的重要基础数据；

c) 数值预报数据：来自各类数值天气预报（NWP）模型和海洋环流模型（OGCM）的输出数据。包括全球模式（如GFS、ECMWF）、区域模式（如WRF、CMA-MESO）和集合预报系统的数据。这些数据提供了未来不同时间尺度的天气和海洋要素预报，如降水、温度、风、台风路径、海浪、潮汐、海洋锋面等。例如，台风路径预报数据，用于灾害预警；海浪预报数据，用于航运和海洋工程；

d) 地面与海洋观测数据：

   1) 地面气象站数据：包括温度、湿度、气压、风速风向、降水等常规气象要素的实时和历史观测数据；

   2) 高空探测数据：包括探空仪、系留气球、风廓线雷达等获取的高空温度、湿度、气压、风速风向等数据，用于分析大气垂直结构；

   3) 海洋浮标数据：包括定点浮标、漂流浮标、Argo 浮标等获取的海表温度、盐度、海流、波浪、潮位等数据，用于监测海洋环境；

   4) 船舶观测数据：来自商船、科考船等平台的气象和海洋观测数据。

e) 岸基雷达数据：用于监测近海海流、波浪和风场；

f) 模式模拟数据：通过高分辨率大气模型、海洋模型、气候系统模型、地球系统模型等进行模拟实验产生的数据。这些数据用于研究大气海洋过程的机理、气候变化情景预测、极端事件归因等。例如，区域气候模型模拟的未来区域降水变化，地球系统模型模拟的碳循环过程；

g) 地理空间数据：包括地形、地貌、水深、海岸线、土地利用类型、植被覆盖等静态地理信息数据。这些数据是大气海洋模型和数据分析的重要背景信息。例如，高分辨率的数字高程模型（DEM）用于地形对气流影响的研究；全球海岸线数据用于海洋边界的定义；

h) 历史灾害事件数据：包括台风、洪水、干旱、风暴潮、海啸、极端高温/低温等历史灾害事件的发生时间、地点、强度、影响范围、经济损失和人员伤亡等数据。这些数据对于灾害风险评估、预警模型训练和防灾减灾策略制定至关重要；

i) 文本数据：包括气象海洋领域的科研论文、报告、预报产品、新闻报道、专家意见等非结构化文本数据。这些数据包含了丰富的领域知识和专家经验，可以通过自然语言处理（NLP）技术进行挖掘和利用，例如，提取特定天气现象的描述、灾害影响评估的文本信息等。

A.3.1.2 数据特点

大气海洋科学领域中的科学智能数据特点如下：

a) 海量性与高维度：大气海洋数据通常具有极大的体量和高维度特征。例如，卫星遥感数据覆盖全球，时间分辨率高，每个像素点包含多个光谱通道信息；数值模式输出数据则在三维空间和时间维度上提供大量的物理量。这些数据不仅量大，而且包含温度、湿度、气压、风速、海温、盐度、海流等多种物理变量，形成了复杂的高维数据空间。处理和分析这些海量高维数据对计算资源和算法效率提出了极高要求；

b) 多源异构性：大气海洋数据来源于多种不同的观测平台和模拟系统，包括卫星、雷达、浮标、探空、地面站以及各类数值模式。这些数据在采集方式、空间分辨率、时间分辨率、数据格式、质量控制标准等方面存在显著差异。例如，卫星数据通常是格点化的，而浮标数据是离散点的；不同型号的雷达数据可能采用不同的扫描模式和数据格式。这种异构性使得数据整合和融合成为一个巨大的挑战，需要开发专门的数据同化和融合技术；

c) 强时空相关性：大气和海洋现象具有显著的时空演变特征。例如，天气系统的移动、海洋涡旋的生成和消亡、气候模式的周期性变化等都体现出强烈的时空连续性和相关性。这意味着数据点之间并非独立同分布，其时空邻近性蕴含着重要的物理信息。在构建科学智能模型时，需要充分考虑这种时空相关性，例如采用时空序列模型或图神经网络等方法；

d) 动态性与实时性：许多大气海洋现象是快速变化的，特别是天气过程和海洋上层动力学过程。因此，数据采集和处理需要具备动态监测和实时反馈的能力，以支持短时临近预报、灾害预警等应用。例如，雷达数据需要实时传输和处理以监测强对流天气；海洋浮标数据需要实时回传以监测海洋环境变化。数据的实时性要求对数据传输、存储和计算效率提出了更高要求；

e) 不确定性与噪声：大气海洋数据不可避免地存在各种不确定性和噪声。这包括观测误差（传感器精度限制、环境干扰）、模型误差（物理过程参数化不完善、数值离散误差）、数据缺失（观测设备故障、传输中断）以及代表性误差（点观测无法完全代表区域情况）。例如，卫星数据可能受到云层遮挡的影响，雷达数据可能受到地物杂波的干扰。科学智能模型在处理这些数据时，需要具备鲁棒性，能够有效识别、量化和处理不确定性，例如通过不确定性量化（UQ）方法。

f) 多尺度性：大气海洋系统是一个典型的多尺度系统，其过程涵盖了从微观湍流（米级、秒级）到中尺度天气系统（百公里级、小时级）再到大尺度环流（千公里级、月/年级）乃至全球气候变化（万公里级、十年/百年级）的广泛时空范围。这意味着在不同尺度上，主导的物理过程和数据特征可能完全不同。科学智能语料库的建设需要能够支持多尺度数据的整合和分析，例如通过多分辨率数据融合或尺度感知模型。

g) 物理约束与可解释性：大气海洋科学是基于物理定律的学科，数据背后蕴含着复杂的物理机制。科学智能模型在处理这些数据时，不仅要追求预测精度，还需要考虑模型结果是否符合基本的物理定律（如质量守恒、能量守恒、动量守恒）以及是否具有良好的可解释性。例如，一个预测台风路径的AI模型，其内部机制要能与台风的物理形成和移动规律相对应。这要求在模型设计中融入物理先验知识，或开发可解释的AI方法，以增强模型的科学性和可信度。

A.3.2 数据生产加工环节特色

A.3.2.1 数据采集

a) 特色内容：大气海洋科学领域的数据采集具有其独特的复杂性和挑战性，与通用科学智能数据采集存在显著差异。

1) 多平台协同观测与数据融合：大气海洋数据来源于极其多样化的观测平台，包括地面气象站、海洋浮标、探空仪、船舶、飞机、无人机、卫星（气象卫星、海洋卫星、地球资源卫星）、地基雷达、激光雷达、声呐等。这些平台在空间覆盖、时间分辨率、观测要素和数据格式上各不相同。例如，卫星提供大尺度、高频次的覆盖，但垂直分辨率有限；浮标提供定点、高精度的海洋剖面数据；雷达提供高时空分辨率的降水和风场信息。科学智能语料库的建设需要能够有效整合这些多源异构数据，进行时空配准、格式转换和初步质量控制，以形成统一的数据集。这涉及到复杂的数据融合算法，例如，将不同卫星传感器的观测数据进行交叉定标和融合，或者将离散的浮标观测数据与格点化的再分析数据进行融合，以弥补单一数据源的不足，提供更全面的地球系统状态描述；

2) 实时数据流的获取与处理：大气海洋现象的快速演变特性决定了许多数据需要实时或近实时地获取和处理，以支持天气预报、海洋预警和灾害应急响应。例如，气象雷达数据、闪电定位数据、实时卫星云图等都是连续不断的数据流。科学智能语料库的建设需要建立高效的数据传输、接收和实时处理系统，能够处理高并发、大数据量的流式数据。这包括流数据解析、实时质量控制、实时特征提取等技术，确保数据能够及时进入语料库并用于模

型的训练和推理。例如，在强对流天气预警中，需要实时处理雷达回波数据，快速识别对流单体并预测其发展趋势；

3) 极端环境下的数据采集挑战：大气海洋环境的复杂性和极端性给数据采集带来了巨大挑战。例如，在台风、暴雨、极地冰区、深海等极端条件下，观测设备的部署、运行和维护都面临严峻考验。传感器可能受到强风、巨浪、低温、高压、腐蚀等影响，导致数据质量下降甚至设备损坏。科学智能语料库的建设需要考虑这些极端环境下的数据特性，例如，开发针对恶劣天气条件下的数据去噪和校正算法，或者利用多传感器冗余观测来提高数据可靠性。此外，对于一些难以直接观测的区域（如深海、高空），还需要依赖遥感技术或间接推断方法进行数据采集，这增加了数据的不确定性；

4) 模式同化与再分析数据的生成：大气海洋科学中一个重要的数据来源是模式同化和再分析数据。这并非简单的观测数据采集，而是将观测数据与数值模型相结合，通过数据同化系统（如变分同化、集合卡尔曼滤波）将不规则分布的观测信息融入到物理模型中，生成时空连续、物理协调的格点化数据集。再分析数据是历史时期模式同化产品的最佳估计，为气候研究提供了宝贵的、一致的数据集。科学智能语料库的建设需要理解这些数据的生成机制，并能够有效利用这些经过物理约束和优化处理的数据，例如，将再分析数据作为AI模型的输入特征，或者用于模型的验证和评估。

b) 数据采集案例：某国家级气象中心正在建设一套基于科学智能的短时临近预报系统，旨在提高未来0-6小时的降水预报精度。该系统的数据采集环节需要整合多源实时数据。这些数据通过高速网络实时传输到数据中心，经过初步的质量控制、时空配准和格式转换后，进入科学智能语料库，供AI模型进行实时训练和推理，以生成高精度的短时临近降水预报产品。数据采集系统需要具备高吞吐量、低延迟和高可靠性，以应对海量实时数据流的挑战。

表 A.13 大气海洋科学数据采集案例

数据名称	数据内容
地面自动气象站数据	全国数万个站点每分钟更新的温度、湿度、气压、风速风向、降水等数据。
新一代天气雷达数据	全国数百部多普勒天气雷达每6分钟更新一次的三维反射率和径向速度数据，用于识别降水回波和风场结构。
闪电定位数据	全国闪电定位网实时提供的闪电发生时间、位置和强度数据，作为强对流天气的指示。
静止气象卫星数据	高时间分辨率（如每10分钟）的可见光、红外、水汽图像，用于监测云系发展和移动。
数值预报模式输出	高分辨率区域数值预报模式（如CMA-MESO）的最新预报产品，作为AI模型的背景场和初始条件。

A.3.2.2 数据清洗

a) 特色内容：大气海洋科学数据的清洗不仅要处理通用的噪声和错误，还要针对其特有的数据形式和内容进行特殊处理。其特色内容如下：

1) 多源异构数据的标准化与统一化：大气海洋数据来自多种观测平台和数值模式，其数据格式、单位、坐标系、时间戳等往往不一致。数据清洗的首要任务是进行标准化和统一化处理，确保所有数据在进入语料库前具有一致的表示形式。例如，将不同卫星传感器的辐射亮度数据统一到相同的物理单位；将不同观测平台的时间戳统一到协调世界时（UT；将不同投影方式的地理空间数据统一到相同的地理坐标系。这通常需要开发复杂的数据解析器、转换工具和元数据管理系统，以实现数据的无缝集成和互操作性；

2) 复杂噪声与异常值的识别与去除：大气海洋数据中存在多种复杂的噪声和异常值，这些可能来源于传感器故障、传输错误、环境干扰或极端物理现象。例如，雷达数据中的地物杂波、晴空回波、传播异常；卫星数据中的云污染、太阳耀斑干扰；浮标数据中的鸟粪、生物附着或设备漂移导致的异常值。数据清洗需要开发专门的算法来识别和去除这些噪声和异常值，以提高数据质量。例如，利用机器学习方法识别雷达数据中的非气象回波；采用统计学方法（如中位数滤波、三倍标准差法）或物理约束方法（如基于大气垂直廓线的温度梯度检查）来检测和修正异常值；对于缺失数据，可以采用时空插值、经验正交函数（EOF）重构或基于深度学习的补全方法进行填充；

3) 数据一致性校验与偏差校正：由于大气海洋数据通常是多源的，不同数据源之间可能存在系统性偏差。例如，不同型号的温度计可能存在微小的测量偏差；卫星反演的海表温度可能与浮标实测值存在系统性差异。数据清洗需要进行数据一致性校验和偏差校正，以确保不同数据源之间的一致性和可比性。这通常涉及到交叉定标、偏差订正和数据同化等技术。例如，利用高精度的参考观测数据对卫星产品进行偏差校正；通过数据同化系统将观测数据与数值模式背景场进行融合，同时校正观测和模式的偏差。此外，还需要对数据的物理一致性进行检查，例如，检查温度、气压、湿度等物理量是否满足热力学定律；

4) 动态更新与迭代清洗：大气海洋数据是动态变化的，新的观测数据和模式产品不断生成，同时对历史数据的理解也在不断深入。因此，数据清洗是一个动态和迭代的过程。随着新的数据进入语料库，需要定期或实时地进行清洗和质量控制。同时，随着对数据特性和噪声模式的更深入理解，清洗算法和参数也需要不断优化和更新。例如，当发现某种新型的传感器故障模式时，需要更新异常值检测算法；当数值模式进行升级时，需要重新评估其输出数据的偏差特性。这种迭代清洗机制有助于持续提高语料库的数据质量和可靠性。

b) 数据清洗案例：某海洋科学研究机构正在构建一个用于海洋涡旋识别和追踪的科学智能语料库。通过清洗步骤，确保进入语料库的海洋数据是高质量、一致且可靠的，从而为后续的海洋涡旋识别AI模型的训练提供坚实的基础。其数据清洗环节面临的挑战和清洗过程如下。

表 A.14 大气海洋科学数据清洗案例

数据名称	具体挑战及清洗过程
卫星高度计数据	原始数据中存在轨道误差、潮汐误差、大气延迟误差等，需要进行精细的地球物理校正。此外，还可能存在由于仪器故障或数据传输问题导致的异常值和数据缺失。 清洗过程包括：应用最新的地球物理模型进行校正；使用中位数滤波和基于统计阈值的方法去除异常值；对于小范围的数据缺失，采用二维插值方法进行填充。
Argo浮标数据	浮标在海洋中漂移和剖面观测，数据可能受到生物附着、传感器漂移、传输中断等影响，导致温度、盐度剖面数据出现尖峰、平台或漂移。 清洗过程包括：自动识别和标记可疑数据点（如通过梯度检查、范围检查）；结合历史数据和邻近浮标数据进行一致性校验；对于明显的异常剖面，进行人工复核和修正。
海洋模式输出数据	模式输出可能存在数值噪声或边界效应。 清洗过程包括：对模式输出进行平滑处理以去除高频噪声；检查模式输出与观测数据的一致性，并进行偏差订正。

A.3.2.3 数据标注

a) 特色内容：大气海洋科学数据的标注不仅需要深厚的专业知识，还要考虑其独特的时空特性和物理意义。其特色内容如下：

1) 专业领域知识的深度标注：大气海洋数据的标注需要具备深厚的气象学、海洋学、气候学等专业知识，以确保标注的准确性和可靠性。例如，在卫星云图的标注中，需要识别不同类型的云（如层云、积云、卷云）、天气系统（如锋面、气旋、反气旋）及其发展阶段。在海洋遥感图像的标注中，需要识别海洋锋面、涡旋、上升流区域、海冰类型和密集度等。这不仅要求标注人员熟悉大气海洋的基本概念和现象，还需要了解特定物理过程的特征和演变规律。因此，大气海洋数据的标注通常需要由专业的科学家或经验丰富的预报员来完成，或者在标注过程中引入领域专家的知识和经验，例如通过专家系统或半监督学习方法辅助标注；

2) 时空连续性与多尺度标注体系：大气海洋现象具有显著的时空连续性和多尺度特征，因此需要建立能够反映这些特性的多层次标注体系。例如，在台风数据的标注中，不仅要标注台风中心的位置、强度、移动路径，还需要标注其风场、雨带、眼区等精细结构，以及不同发展阶段（如热带低压、热带风暴、强台风）的演变。在海洋涡旋的标注中，需要标注涡旋的中心、边界、旋转方向、生命周期等。这种多层次、时空连续的标注体系可以为后续的 AI 模型训练提供更丰富的信息，帮助模型捕捉大气海洋现象的动态演变和多尺度相互作用。例如，通过对历史台风路径和强度进行标注，可以训练 AI 模型进行台风路径和强度预测；通过对海洋锋面和涡旋的标注，可以研究其对海洋生态系统和能量传输的影响；

3) 不确定性与模糊性标注：大气海洋现象往往存在一定的不确定性和模糊性，例如，锋面的位置可能不是一条清晰的线，而是具有一定宽度的过渡带；云的分类可能存在模糊边界。数据标注需要能够反映这种不确定性，例如，采用概率标注、区域标注或多标签标注等方式。此外，对于一些难以直接观测或定义的气象海洋要素，可能需要通过间接推断或专家共识进行标注。科学智能语料库的建设需要考虑如何有效地存储和利用这些带有不确定性信息的标注数据，例如，通过贝叶斯深度学习模型来处理标注的不确定性，或者通过集成学习方法来融合不同专家标注的结果；
4) 动态标注与迭代更新：大气海洋科学研究的不断深入和观测技术的进步，会带来新的发现和对现有现象更精确的理解。因此，数据标注是一个动态和迭代的过程，需要根据研究进展和新知识进行更新。例如，随着对某种极端天气事件机制的更深入了解，可能会发现新的识别特征或分类标准，这时需要对已标注的数据进行重新评估和更新。同时，动态标注还可以根据新的观测数据和模型模拟结果，对标注内容进行优化和完善。例如，当新的高分辨率卫星数据可用时，可以对历史事件的精细结构进行更准确的标注。这种迭代更新机制有助于语料库保持其科学性和时效性。

b) 数据标注案例：某气象研究团队正在开发一个基于深度学习的强对流天气识别和预警系统。其数据标注环节需要对历史雷达数据和卫星数据进行精细标注. 通过这些精细的标注，AI模型可以学习到强对流天气的视觉特征和演变规律，从而实现对未来强对流天气的自动识别和预警。标注过程通常会结合半自动化工具，例如，先由AI模型进行初步识别，再由专家进行修正和确认，以提高标注效率和质量。

表 A.15 大气海洋科学数据标注案例

标注类别	具体内容
雷达回波图像标注	标注人员（经验丰富的气象预报员）需要识别并勾勒出雷达回波图中的强对流单体（如超级单体、飑线）、冰雹回波区、龙卷风涡旋特征等。标注内容包括：对流单体的边界、中心位置、强度等级、移动方向和速度；冰雹回波的识别标志（如三体散射、高反射率核）；龙卷风涡旋特征（如中尺度涡旋、钩状回波）的区域。
卫星云图标注	标注人员需要识别并勾勒出卫星云图中的对流云团、飑线系统、高架雷暴等特征。标注内容包括：云团的形态、发展阶段、顶高、亮温特征；对流云团内部的过冲云顶、V形缺口等强对流指示特征。
事件标签标注	除了图像区域标注，还需要为每个强对流事件附加事件标签，包括：事件类型（如雷暴、冰雹、龙卷风、短时强降水）、发生时间、持续时间、影响区域、造成的灾害类型和强度等。这些标签通常需要结合地面灾情报告、自动站观测数据和专家判断进行综合确定。

A.3.2.4 数据测试

a) 特色内容：大气海洋科学数据的测试需要考虑其物理背景、模型验证的复杂性以及对预报预警时效性的要求。其特色内容如下：
1) 基于物理规律和领域知识的测试：大气海洋科学是基于物理定律的学科，因此数据测试不仅要关注统计学指标，更要确保模型结果符合基本的物理规律和领域知识。例如，测试 AI 模型预测的风场是否满足质量守恒定律；预测的海温变化是否符合热力学原理；识别的台风结构是否符合气象学特征。这要求在测试指标中融入物理约束，或通过专家系统对模型输出进行物理一致性检查。例如，可以计算模型预测的能量、动量、质量等守恒量，并与理论值或观测值进行比较，以评估模型的物理合理性。此外，对于一些复杂的大气海洋现象，还需要通过领域专家对模型输出的物理意义进行定性评估；
2) 与传统数值模式结果的对比验证：在许多大气海洋科学应用中，传统数值模式已经积累了数十年的发展和应用经验，其结果具有较高的可信度。因此，科学智能模型的测试常常需要与传统数值模式的结果进行对比验证。例如，比较 AI 模型预测的降水与数值天气预报模式的降水预报；比较 AI 模型反演的海表温度与卫星或浮标观测数据同化后的模式分析场。这种对比不仅可以评估 AI 模型的性能提升，还可以发现 AI 模型在某些特定场景下的优势或劣势，为 AI 与物理模型的融合提供依据。例如，可以采用均方根误差（RMSE）、相关系数、技巧评分等统计指标，对 AI 模型和数值模式的预报结果进行定量比较；

3) 多维度性能评估与预报预警时效性：大气海洋科学数据的测试需要从多个维度进行性能评估，特别是对于预报和预警应用，时效性是关键指标。除了传统的准确率、召回率、F1分数等分类或回归指标外，还需要评估模型的预报时效（提前多长时间发出预警）、预警准确率（是否误报或漏报）、空间分辨率（是否能捕捉精细结构）、鲁棒性（对噪声和异常值的抵抗能力）以及泛化能力（在未见过的数据或不同区域的表现）。例如，在台风路径预报中，除了路径误差，还需要评估强度预报误差和登陆时间误差。在洪水预警中，除了预警准确性，还需要评估预警的提前量。这些多维度的性能评估可以为研究人员提供更全面的评估结果，帮助他们更好地选择和优化数据测试方法；

4) 动态测试与反馈机制：大气海洋系统是一个动态变化的复杂系统，新的观测数据和物理过程的理解不断涌现。因此，数据测试是一个动态和迭代的过程。需要建立持续的测试和反馈机制，以便在模型部署后能够持续监测其性能，并根据实际运行情况进行调整和优化。例如，当新的观测数据可用时，需要重新评估模型的性能；当发现模型在特定天气条件下表现不佳时，需要对模型进行再训练或调整。这种动态测试和反馈机制有助于确保科学智能模型在不断变化的大气海洋环境中保持其有效性和可靠性。

b) 数据测试案例：某研究团队开发了一个基于深度学习的海洋锋面识别模型，旨在利用卫星海表温度数据自动识别海洋锋面。通过全面的测试，确保海洋锋面识别模型在准确性、物理合理性和鲁棒性方面都达到要求，从而为海洋锋面研究和海洋渔业应用提供可靠的数据支持。

表 A.16 大气海洋科学数据测试案例

环节	具体内容
物理一致性测试	检查模型识别出的锋面是否符合海洋动力学特征，例如，锋面两侧的温度梯度是否合理，锋面是否与海流、涡旋等海洋现象存在合理的物理关联。这可能需要结合海洋模式输出的海流数据进行交叉验证。
与专家标注对比	将模型识别出的锋面与海洋学专家手工标注的锋面进行对比，计算识别准确率、漏报率和误报率。这通常通过计算像素级的IoU（Intersection over Union）或豪斯多夫距离（Hausdorff Distance）来量化。
跨区域泛化能力测试	在不同海域（如黑潮区、墨西哥湾流区）的卫星数据上测试模型的性能，评估模型在不同海洋环境下的泛化能力。
时间序列稳定性测试	对同一海域不同时间序列的卫星数据进行测试，评估模型识别锋面的时间连续性和稳定性，确保锋面在时间演变上是合理的。
对噪声的鲁棒性测试	在加入不同程度噪声（如云污染、传感器噪声）的卫星数据上测试模型的性能，评估模型对数据质量下降的抵抗能力。

A.3.3 语料使用及特色

大气海洋科学智能语料库的特色在于强调物理知识与数据驱动方法的深度融合。语料库中的数据不仅是原始观测和模拟数据，还包括经过物理校正、同化处理的数据，以及蕴含物理规律的特征。AI模型在利用语料库数据时，可以融入物理约束、守恒定律和领域先验知识，从而构建出既能从数据中学习复杂模式，又符合基本物理原理的“物理知情AI模型”（Physics-informed AI）。这有助于提高模型的泛化能力、可解释性和科学可信度，避免“黑箱”问题。

A.3.3.1 使用方向

大气海洋科学智能语料库的建设旨在解决本领域内的关键科学问题和实际应用挑战，其使用方向包括但不限于：

a) 提高天气和气候预报的准确性和时效性

1) 传统数值天气预报模型在计算资源、物理参数化和初始条件等方面存在局限性，导致预报精度和时效性仍有提升空间，尤其是在短时临近预报和极端天气事件预报方面。气候模型在模拟复杂地球系统过程和预测区域气候变化方面也面临挑战；

2) 大气海洋科学数据语料库是训练和优化科学智能模型以提高天气和气候预报能力的核心。语料库提供海量的历史观测数据（如卫星、雷达、地面站、探空、浮标数据）和数值模式输出数据，为 AI 模型学习大气海洋系统的复杂非线性关系提供了基础。通过 AI 模型，可

以实现对雷达回波、卫星云图等数据的快速识别和外推，提高短时临近预报的精度；可以对数值模式输出进行偏差订正和超分辨率处理，提升中长期预报的准确性；还可以通过AI模型模拟气候系统中的关键物理过程，改进气候模型的性能；

3) 将多源异构的大气海洋数据（如雷达反射率、卫星亮温、地面气象要素、模式预报场等）输入到基于深度学习的AI模型（如卷积神经网络、循环神经网络、图神经网络）中进行训练。训练好的模型可以用于实时预测降水、风场、温度等要素，或对数值模式预报结果进行后处理。例如，利用AI模型对雷达数据进行外推，实现未来1-2小时的降水预报；利用AI模型对全球数值模式的输出进行降尺度处理，生成区域精细化预报产品；

b) 精细化海洋环境监测与预警

1) 海洋环境复杂多变，对海洋现象（如海洋锋面、涡旋、赤潮、海冰）的监测和预警需要高分辨率、高时效性的数据支持，传统方法难以满足精细化管理和决策的需求；

2) 语料库汇集了大量的海洋观测数据（如卫星海表温度、海面高度、海色数据、浮标数据、水下声学数据）和海洋模式模拟数据。AI模型可以利用这些数据识别和追踪海洋现象，预测其发展趋势。例如，通过AI模型对卫星海表温度和海面高度数据进行分析，可以自动识别海洋锋面和涡旋的生成、发展和消亡过程；通过AI模型对海色数据进行分析，可以监测赤潮的发生和扩散；通过AI模型对海冰密集度数据进行分析，可以预测海冰的融化和漂移；

3) 将海洋遥感图像、浮标观测数据、水下声学数据等输入到AI模型中进行训练。训练好的模型可以用于实时监测海洋环境，识别异常现象，并发出预警。例如，利用AI模型自动识别卫星图像中的赤潮区域，并结合海洋环流模型预测其扩散路径，为渔业和海洋管理部门提供预警信息；利用AI模型分析水下声学数据，识别海洋哺乳动物的声纹，辅助海洋生态保护。

c) 深入理解大气海洋系统演变规律

1) 大气海洋系统是一个高度复杂的非线性系统，其内部物理过程相互作用，并受到外部强迫的影响。传统物理模型在揭示所有复杂机制方面仍有局限，尤其是在多尺度相互作用和极端事件形成机理方面；

2) 语料库提供了长时间序列、多维度、多尺度的历史观测和模拟数据。AI模型可以利用其强大的模式识别和特征提取能力，从海量数据中发现传统方法难以察觉的潜在规律、遥相关和因果关系。例如，AI可以帮助识别气候系统中的关键模态及其相互作用，揭示极端天气事件（如热浪、干旱、洪涝）的形成机制和驱动因素；还可以用于改进地球系统模型中复杂物理过程的参数化方案，从而提高模型的物理合理性和模拟能力；

3) 将历史气候数据（如温度、降水、海表温度指数）、古气候重建数据、地球系统模型模拟数据等输入到AI模型中进行分析。AI模型可以用于发现数据中的潜在模式、进行因果推断、构建代理模型等。例如，利用AI模型分析历史极端降水事件与大气环流异常之间的关系，揭示极端降水事件的形成机制；利用AI模型构建云微物理过程的代理模型，加速地球系统模型的模拟速度并提高其准确性。

d) 支持防灾减灾和应对气候变化

1) 大气海洋灾害频发，对社会经济和人民生命财产造成巨大损失。气候变化背景下，极端事件的频率和强度可能增加，需要更有效的风险评估、预警和适应策略；

2) 语料库中包含的历史灾害事件数据、实时观测数据和气候变化情景数据，为AI模型进行灾害风险评估、预警和影响评估提供了基础。AI模型可以学习历史灾害事件的特征和影响，预测未来灾害的发生概率和潜在影响区域。例如，利用AI模型预测台风的登陆地点和强度，为防灾部门提供决策依据；通过分析历史洪水数据和地形数据，评估洪水风险区域并优化防洪策略。在应对气候变化方面，AI可以用于分析气候变化对极端事件频率和强度的影响，评估气候变化情景下的区域风险，并支持制定适应和减缓气候变化的策略；

3) 将历史灾害数据（如台风路径、洪水淹没范围、干旱指数）、实时观测数据、社会经济数据等输入到AI模型中进行训练。训练好的模型可以用于灾害风险评估、预警和影响评估。例如，利用AI模型预测未来台风可能造成的经济损失和人员伤亡，为应急响应提供支持；利用AI模型评估海平面上升对沿海城市基础设施的影响，为城市规划提供科学依据。

A.3.3.2 使用案例

a) 某国家气象局正在利用科学智能语料库构建一个“智慧气象”平台，旨在提升气象服务能力，特别是针对极端天气事件的预报预警。

表 A.17 大气海洋科学数据使用案例-极端天气预测

应用环节	具体内容
短时临近预报	平台整合了全国雷达组网数据、闪电定位数据、地面自动站数据和高分辨率卫星数据。AI模型（如基于卷积神经网络和循环神经网络的混合模型）利用这些实时数据流，学习强对流天气的发生发展规律，实现未来1-2小时的降水、雷暴、大风等精细化预报。例如，当雷达监测到对流单体快速发展时，AI模型能够立即识别并预测其移动路径和强度变化，为城市防洪和交通管理提供提前量。
台风路径和强度预测	平台汇集了历史台风观测数据（包括卫星、飞机侦察、地面观测）、全球数值模式预报数据和再分析数据。AI模型（如基于图神经网络和注意力机制的模型）通过学习大量历史台风案例，预测台风的未来路径、强度变化和登陆地点。例如，在台风逼近时，AI模型可以提供多路径、多强度的概率预报，帮助决策者评估不同情景下的风险，并提前部署防灾资源。
气候变化影响评估	平台利用全球气候模式模拟数据、历史观测数据和AI技术，评估气候变化对区域极端天气事件（如高温热浪、干旱）的影响。例如，AI模型可以分析历史数据中极端高温事件的频率和强度变化趋势，并结合未来气候情景预测，评估未来不同排放路径下极端高温事件的风险，为城市规划和公共卫生政策制定提供科学依据。

b) 某海洋气象中心正在利用大气海洋科学智能语料库，构建一个面向海上风电场的精细化气象监控系统，以优化风电场的运行和维护，并提高发电效率。

表 A.18 大气海洋科学数据使用案例-风电场气象监控

应用环节	具体内容
数据处理	整合历史和实时观测数据（海上浮标的风速风向、波高、海流数据；风电塔上的测风雷达数据）、高分辨率数值天气预报模式（如WRF）输出数据、卫星遥感数据（如海面风场、海浪数据）；统一质量控制、时空配准和格式转换。
模型训练与优化	训练三类AI模型：风功率预测模型（基于LSTM、Transformer等深度学习算法，融入地形、海面粗糙度等物理特征，预测未来24-72小时风速和风功率）；海浪预报模型（预测未来海浪状态，为海上作业提供安全保障）；极端天气预警模型（识别和预测台风、风暴潮、大雾等极端天气事件及其对风电场的影响）。
系统集成与应用	模型集成到监控系统中，实时接收数据并生成精细化监控产品，通过可视化界面提供：未来风功率曲线（指导风机启停和发电计划）、海上作业窗口期预报（优化运维船只调度）、极端天气预警（提前采取防范措施），最终实现提高发电效率、降低运维成本、有效应对海上极端天气风险。

A.3.4 数据安全措施

典型的大气海洋科学领域数据安全措施包括但不限于：

a) 高分辨率地理环境数据管控：针对具有军事应用价值的高分辨率地形、海洋水文、大气环流数据，应建立国家安全审查机制，实施数据脱敏处理，对外提供降精度版本，并建立用户身份验证和使用目的审核制度；

b) 关键基础设施环境数据保护：为防止沿海关键基础设施环境数据被恶意利用，应建立设施周边数据的特殊保护区域，实施更严格的访问控制，对数据使用者进行深度背景调查，并建立实时威胁监测和预警系统；

c) 气象与气候模型数据质量保障：为确保长期气候观测数据和数值预报模型参数的准确性，宜建立多源数据交叉验证机制，采用数字水印技术标记原始数据，实施版本控制和变更管理，并建立异常数据自动检测和报警系统；

d) 海洋声学与水文数据访问控制：针对海洋温盐深剖面、海底地形地貌等敏感水文数据，应建立基于需求的最小授权原则，对特定海域数据实施地理围栏保护，建立数据使用审批和监督机制，并采用动态脱敏技术处理敏感区域数据。

A.4 空间信息科学

A.4.1 数据内容及特点

空间信息科学数据具有以下主要类型和特点：

a) 数据类型
   1) 预报数据：确定性预报（如 GFS、HRES）、集合预报（如 GEFS、ENS）、次季节-季节预测（S2S）等模型输出数据；
   2) 再分析数据：ERA5 等融合历史观测与模型模拟的综合性数据集；
   3) 静态数据：地形高程（ASTER GDEM）、土地利用（Global Land Cover）等基础地理信息。
b) 数据特点
   1) 时空耦合性：每个数据都具有明确的空间位置（经纬度、高程）和时间标记；
   2) 多尺度嵌套：从厘米级到全球级，不同尺度数据相互关联；
   3) 海量异构性：单景高分影像达 GB 级，年度累积数据达 PB 级，格式多样；
   4) 强关联性：地理要素间存在拓扑、方位、距离等空间关系；
   5) 动态更新性：地表变化监测要求数据持续更新。

A.4.2 数据生产加工环节特色

A.4.2.1 数据采集

在数据采集环节，空间信息科学最大的特色是多源协同采集。

a) 特殊操作
   1) 多源协同采集：整合卫星、无人机、地面传感器数据，确保时空同步；
   2) 几何校正：消除地形起伏、传感器姿态等引起的几何畸变；
   3) 大气校正：去除大气散射、吸收对遥感数据的影响；
   4) 云检测与去除：识别并处理云层遮挡区域。
b) 实操案例：城市建筑物变化监测数据采集

表 A.19 空间信息科学数据采集案例

步骤	具体内容
输入	哨兵2号卫星10米分辨率影像 + 无人机倾斜摄影5厘米分辨率影像
处理流程	卫星影像预处理：辐射定标→大气校正（6S模型）→云掩膜（QA波段）
	无人机影像处理：POS数据解算→空三加密→正射纠正
	多源配准：选取同名点（建筑物角点）→仿射变换→配准精度<2像素
	时相匹配：提取同一季节数据，避免植被物候差异
输出	多尺度时空对齐的建筑物监测数据集

A.4.2.2 数据清洗

数据清洗环节的特色操作主要体现在空间数据特有的拓扑关系处理上。

a) 特殊操作：
   1) 空间拓扑修复：修正悬挂点、自相交、缝隙等拓扑错误；
   2) 投影统一：将不同坐标系数据转换到统一参考系；
   3) 边界融合：处理相邻图幅接边误差；
   4) 异常值空间检测：基于空间自相关识别局部异常。
b) 案例实操：处理全球雷达数据。

表 A.20 空间信息科学数据清洗案例

环节	具体内容
清洗	剔除雷达信号衰减导致的空值；
加工	将二进制原始流转换为ZARR格式，对齐地理网格；
标注	附加时间戳（UTC+8）和经纬度元数据。

A.4.2.3 数据标注

标注环节的特色在于需要考虑空间关系和多尺度特征。

a) 特殊操作：
   1) 尺度标注策略：大目标粗标注→小目标精细标注；
   2) 时序一致性标注：确保同一地物在时间序列上的 ID 一致；
   3) 三维标多注：标注建筑物高度、植被冠层结构；
   4) 不确定性标注：标记混合像元、模糊边界的置信度；
b) 实操案例：高分辨率遥感影像建筑物提取标注。

表 A.21 空间信息科学数据标注案例

环节	具体内容
预标注	使用已有建筑物矢量数据生成初始标注
分级标注	L1级：建筑物/非建筑物二分类 L2级：住宅/商业/工业/公共建筑细分类 L3级：建筑物轮廓精确勾画（误差<1米） L4级：楼层数、建筑高度属性标注
质量控制	空间一致性：相邻建筑物不重叠 属性逻辑性：高度与楼层数匹配 时序连续性：新增/拆除标记合理
不确定性量化	遮挡区域标注置信度0.3-0.7

A.4.2.4 数据测试

数据测试环节强调空间精度和空间分布特征的验证。

a) 特殊操作：
   1) 空间精度验证：野外实测控制点验证几何精度；
   2) 空间自相关检验：Moran's I 指数评估空间分布合理性；
   3) 时序一致性检验：变化检测结果的逻辑合理性；
   4) 尺度效应评估：不同空间分辨率下的精度变化；
b) 实操案例：土地覆盖分类产品精度评估。

表 A.22 空间信息科学数据测试案例

环节	具体内容
样本设计	分层随机采样，每类别至少100个样本点
参考数据	Google Earth高清影像 + 实地调查照片
混淆矩阵计算	总体精度(O：≥85% Kappa系数：≥0.80 生产者精度：各类别≥80% 用户精度：各类别≥80%
空间精度分析	边界位置误差：RMSE≤2个像素 面积一致性：相对误差≤5%
误差空间分布图	识别系统性偏差区域

A.4.3 语料使用及特色

A.4.3.1 使用方向

空间信息科学领域当前面临的核心挑战是如何从海量、多源、异构的空间数据中快速准确地提取有价值的信息并支持科学决策。借助空间索引构建（R-tree、GeoHash等加速空间查询）、金字塔构建（多分辨率数据组织，支持快速可视化）、时空立方体组织（便于时序分析和4D可视化）、分布式存储（GeoTrellis、GeoMesa等支持大规模计算）等技术，空间信息科学智能语料可用于以下代表性领域：

a) 智能遥感解译：传统的人工目视解译方法效率低下，一个经验丰富的解译员一天只能完成几平方公里的精细解译工作，而且结果受主观因素影响较大。通过构建遥感影像智能解译模型，首先利用百万级的多源遥感影像进行基础模型预训练，采用自监督学习方法如MAE或SimCLR学习通用的视觉特征。然后针对具体应用场景进行微调，比如违建监测系统可以实现月度自动更新，及时发现新增建筑物；生态监测系统能够定量评估植被覆盖度的变化趋势；灾害应急响应系统可以在震后数小时内完成建筑物损毁程度的快速制图，为救援决策提供支持。

表 A.23 空间信息科学智能遥感解释实操路径

基础模性预训练	
输入	百万级多源遥感影像（哨兵、高分、Landsat）
方法	自监督学习（MAE、SimCLR）学习通用特征
下游任务微调	
地物分类	基于标注样本微调分类头
变化检测	孪生网络检测多时相差异
目标检测	YOLO/Faster R-CNN检测特定目标
实际应用	
违建监测	月度更新，自动识别新增建筑
生态监测	植被覆盖度变化定量评估
灾害评估	震后建筑物损毁快速制图

b) 地理知识图谱构建与推理：传统的地理信息系统虽然能够存储和管理空间数据，但在知识表达和推理方面存在不足。通过从地名词典、POI数据、地理文献中抽取地理实体，识别"位于"、"相邻"、"包含"等空间关系，整合人口、经济、环境等多源属性，可以构建丰富的地理知识图谱。基于这个知识图谱，系统能够进行复杂的空间推理。例如，当用户查询"距离地铁站500米内的三甲医院"时，系统会自动进行推理：首先定位所有地铁站点，然后生成500米缓冲区，在缓冲区内筛选医院类POI，最后根据医院等级属性过滤出三甲医院，并计算最优到达路径。

表 A.24 空间信息科学地理知识图谱实操路径

知识抽取	
实体识别	从地名词典、POI数据提取地理实体
关系抽取	提取“位于”、“相邻”、“包含”等空间关系
属性融合	整合多源属性（人口、经济、环境）
知识推理	
空间推理	基于拓扑关系推断隐含空间关系
时序推理	预测城市扩张、土地利用转换趋势
因果推理	分析自然灾害链、城市问题成因
应用示例	
输入	“查找距离地铁站500米内的三甲医院”
推理链	地铁站点→缓冲区分析→医院POI筛选→等级属性过滤
输出	符合条件的医院列表及最优路径

c) 时空预测与仿真：传统的物理模型虽然有坚实的理论基础，但往往计算复杂度高，参数确定困难。通过深度学习方法，可以直接从历史数据中学习时空演化规律。在交通流预测中，使用Graph WaveNet等图神经网络模型，输入历史1小时的路网流量数据，可以预测未来2小时的交

通流量，平均绝对百分比误差能够控制在15%以内。在空气质量预测中，结合气象数据、污染源分布和地形信息，使用物理约束神经网络可以实现PM2.5浓度的72小时预报。在城市扩张模拟中，融合历史土地利用数据和社会经济指标，通过元胞自动机与深度学习相结合的方法，能够预测2030年的城市边界，为城市规划提供科学依据。

表 A.25 空间信息科学时空预测与仿真实操路径

时空序列建模			
数据组织	构建(lon, lat, time, features) 张量		
模型架构	ConvLSTM、ST-GCN、Transformer		
训练策略	多步预测、课程学习		
典型应用			
	交通流预测	空气质量预测	城市扩张模拟
输入	历史1小时路网流量	气象数据+污染源+地形	历史土地利用+社会经济数据
模型	Graph WaveNet	物理约束神经网络(PINN)	元胞自动机+深度学习
输出	未来2小时流量预测，MAPE<15%	PM2.5浓度72小时预报	2030年城市边界预测

d) 地理空间智能决策。在应急响应场景中，如地震发生后，系统需要综合考虑震中位置、烈度分布、道路通行状况、救援资源分布、人口密度等多方面信息。通过损失评估模型预测不同区域的受灾程度，使用路径规划算法计算考虑道路损毁情况的最优救援路线，运用多目标优化方法在时间最短和覆盖范围最广之间寻找平衡，并根据实时反馈动态调整救援方案。在国土空间规划中，生态保护红线的划定需要平衡生态保护与经济发展。智能决策系统可以自动进行生态敏感性评价，识别与建设用地、基本农田的潜在冲突，使用遗传算法等优化方法生成多个备选方案，并评估每个方案对生态连通性和经济发展的影响，辅助决策者做出科学选择。

表 A.26 空间信息科学地理空间智能决策实操路径

应急响应决策	
场景	地震救援资源调度
输入	震中位置、烈度分布 道路通行状况（遥感识别） 救援队伍、物资分布 人口密度、建筑物类型
决策过程	损失评估模型预测受灾程度 路径规划算法计算最优救援路线 多目标优化分配救援资源
动态调整	基于实时反馈更新方案
输出	救援方案（时间最短+覆盖最广）
国土空间规划	
场景	生态保护红线划定
智能体功能	生态敏感性评价（坡度、植被、水源地） 冲突识别（与建设用地、基本农田） 方案生成（遗传算法优化边界） 影响评估（生态连通性、经济损失）

A.4.3.2 使用案例

某城市希望引入城市违建监测系统，实现违建的自动发现和实时监管，大幅提升城市管理的精细化水平。

表 A.27 空间信息科学城市违建检测系统案例

应用环节	具体内容
影像获取与预处理	定期获取多源遥感影像数据（哨兵卫星、高分卫星、无人机影像）；与历史建筑矢量数据、城市规划底图、地籍数据等进行空间配准；建立空间索引（R-tree）和金字塔结构支持快速查询，完成几何校正和辐射定标处理。
变化检测与识别	运行训练后的孪生网络模型进行多时相变化检测，自动识别新增建筑物；使用目标检测算法（YOLO、Faster R-CNN）提取建筑物轮廓和属性信息；结合城市规划数据进行合规性自动判断，筛选出违建疑似目标。
违建确认与执法	通过可视化界面展示违建疑似点位、变化热力图、统计报表；执法人员现场核查确认违建；自动生成执法线索和处置建议推送给相关部门；建立违建档案和处置跟踪，形成从发现到处置的闭环管理流程。

A.4.4 数据安全措施

典型的空间信息科学领域数据安全措施包括但不限于：

a) 卫星遥测遥控数据安全保障：针对卫星轨道参数、姿态信息、遥测数据及遥控指令，应建立多重身份认证和授权机制，采用量子加密技术保护关键通信链路，实施实时入侵检测和异常行为分析，并建立卫星数据的安全备份和应急恢复预案。

b) 高分辨率对地观测影像管控：为防止关键区域设施信息暴露，应建立敏感目标识别和自动模糊化处理系统，实施基于内容的访问控制，对高敏感区域影像采用特殊加密保护，并建立影像数据的使用审计和溯源机制。

c) 位置服务数据隐私保护：针对GPS轨迹数据和室内定位信息，应采用差分隐私技术处理个人轨迹数据，建立数据匿名化和去标识化流程，实施数据最小化收集原则，并建立用户隐私设置和数据删除机制。

d) GNSS精密产品数据完整性保护：为确保精密星历与钟差等导航定位参数的准确性，应建立多重数据源交叉验证机制，采用数字签名和时间戳技术保证数据不可篡改，实施实时数据质量监控，并建立异常数据的快速响应和纠正机制。

A.5 工程技术科学

A.5.1 数据内容及特点

a) 工程技术领域的科学智能数据主要模态及来源包括但不限于：

表 A.28 工程技术科学科学智能数据模态及来源

数据类型	典型示例	主要来源场景
文本数据	设计文档、技术手册、维护日志	设计规划、施工记录、设备运维
图像/视频数据	设备照片、结构损伤图像、工业监控视频	现场巡检、缺陷检测、过程监控
时序传感数据	温度、压力、振动等实时采集信号	工业现场监测、设备运行状态采集
结构化数据	CAD模型、BIM建筑信息模型、有限元网格、实验测量表格	仿真分析、实验测试、数字化设计

b) 典型领域数据构成包括但不限于：

1) 土木工程：以结构拓扑关系明确的空间数据（BIM/GIS）为主，辅以结构监测时序数据（应变/加速度）、施工影像及规范文本；

2) 机械工程：涵盖高维异构数据（设备运行传感数据、3D 几何模型、工业视觉图像），需关联多部件性能耦合分析；

3) 自动化控制：强调实时流式数据（PLC 日志、多传感器融合信号），需支持毫秒级时序因果分析；

4) 材料工程：高精度实验数据（应力-应变曲线、显微图像）与仿真数据并存，微观尺度数据精度敏感性强；

5) 电子工程：强结构化数据（电路网表、芯片工艺参数），兼具实时性（信号波形）与安全性（商业机密设计）；
6) 化工与环境：动态非线性工艺数据（温度/浓度曲线），需处理噪声干扰；环境监测数据具备时空多尺度特性（地面传感+遥感）。

c) 核心数据特性

工程技术数据需满足以下基础特性要求：

表 A.29 工程技术科学数据基础特性要求

特性	内涵说明	技术要求
高维度性	多传感器变量、仿真网格节点形成高维时空场数据	支持特征降维与分布式计算架构
多模态融合	文本、图像、时序数据共存，需跨模态关联分析（如振动信号+损伤图像诊断）	建立统一时空基准的异构数据对齐机制
强结构化	设计拓扑（装配树、电路连接）、空间关系（BIM构件）等显式工程语义	采用本体论描述数据关系，保障结构信息可追溯性
实时动态性	工业控制、环境监测等场景需亚秒级流数据处理	支持流式计算框架与时序数据库存储
噪声鲁棒性	传感器漂移、环境干扰导致异常值频繁	集成数据清洗与异常检测管道
安全敏感性	含关键基础设施运行参数、知识产权设计文档等机密信息	建立分级脱敏与访问控制机制

d) 以面向大飞机运维场景为例，其多源数据具有以下典型融合特征：
1) 强时序性与高维度：主要传感器数据以秒级或更高频率持续采样，通道数量多，维度高，构成模型训练的基础输入；
2) 异常稀疏性与高敏感性：故障事件少且时长短，分布不均，常伴有微弱异常征兆，要求模型具备高灵敏度与精度；
3) 模态异构与系统耦合性：涉及数值、文本、图像、语音等多种模态，来源于多个子系统，数据之间具有关联性与上下文依赖；
4) 预处理复杂性与标准一致性：不同来源数据需统一命名规则、归一化方法与格式标准，确保跨任务、跨系统建模时的一致性与可复用性；
5) 运行场景驱动性强：数据表现受航线类型、飞行阶段、天气条件、机型差异等影响显著，需结合运行上下文进行动态建模与解释。

A.5.2 数据生产加工环节特色

A.5.2.1 数据采集

a) 采集内容包括真实环境数据与仿真合成数据两类来源，并满足以下要求：
1) 真实数据采集：需适应工程现场复杂性（如风电叶片检测中融合无人机图像与 SCADA 传感数据）；实施设备防护与校准（如抗风蚀传感器、防抖无人机）；遵守安全规范（如限定无人机飞行高度、设备防爆等级）；
2) 合成数据生成：针对极端工况数据缺失问题（如叶片断裂工况），通过有限元仿真或生成式模型补充（示例：裂纹应力场仿真、损伤图像合成）；
3) 多源异构处理：部署边缘计算网关，统一无人机影像、传感器流、仿真结果等格式；
4) 安全合规存储：原始数据须加密存储于私有环境，实施基于角色的访问控制（RBA。

b) 以面向大飞机运维场景的数据采集为例
1) 按照飞行架次完整采集航程内多通道传感器数据；
2) 所采数据需与航班日志、设备编号、起降时刻严格对应；
3) 采集频率、数据长度、通道完整性需满足最低技术要求。

A.5.2.2 数据清洗

a) 清洗流程需包含以下关键操作

表 A.30 工程技术科学数据清洗流程关键操作

环节	工程领域特殊要求	风电叶片检测案例示例
质量评估	制定多模态评估指标（如图像清晰度、传感器缺失率阈值）	剔除强光反光/抖动模糊的无人机影像
异常处理	联合统计方法与机器学习检测离群点	用箱线图识别异常风速尖峰并剔除
格式规范化	统一时空基准（如 GPS 时间戳对齐图像与振动信号）	将叶片图像与对应时刻应变数据绑定
协同清洗	跨模态关联验证（示例：依据振动谱筛除抖动模糊图像）	结合加速度计数据过滤风振导致的无效影像
隐私脱敏	采用工程专用脱敏策略（如抹除设备铭牌、匿名化运维文本）	删除图像中的风机序列号与厂商标识

b) 以面向大飞机运维场景的数据清洗为例
   1) 清除空值、重复记录与非法数据点；
   2) 对常值通道、失效通道、漂移信号等进行识别与剔除；
   3) 非数值字段可采用掩码机制统一处理，确保数据输入统一性。

A.5.2.3 数据标注

a) 数据标注要点包括但不限于：
   1) 专业参与机制：标注人员须具备工程领域知识（如能分辨叶片裂纹与污渍差异）；
   2) 分层标注体系：构建多级标签（示例：叶片缺陷位置[前缘/根部]+类型[裂纹/剥落]+严重度）；
   3) 工具与质检：采用半自动工具辅助（如 AI 预分割裂纹区域供专家修正）；执行双盲审核（两组专家独立校验标注一致性）；
   4) 本体化管理：预先定义标注规范手册（含缺陷分类标准、信号异常模式词典）；
   5) 元数据追溯：标注结果需关联责任人、时间及修订记录（JSON/XML 标准格式输出）
b) 以面向大飞机运维场景的数据标注为例：
   1) 基于飞行记录与维护日志，构建“是否异常”、“异常类型”、“起止时间”等多级标签；
   2) 标签体系建议采用主类-子类分层结构，以支持多层次模型训练与评估；
   3) 所有标签应通过审核机制确认其一致性与准确性。

A.5.2.4 数据测试

a) 测试验证覆盖以下维度：
   1) 基础完整性检验：检查字段完整性、文件可解析性；
   2) 多模态对齐验证：确保跨源数据时空同步；
   3) 工程合理性核查：基于知识图谱验证标签逻辑
   4) 场景覆盖度评估：检验工况完备性
   5) 应用仿真测试：通过基准模型反推数据缺陷
b) 以面向大飞机运维场景的数据测试为例
   1) 应以航班编号、时间区段或设备编号为单位进行划分，避免样本泄露；
   2) 异常样本应覆盖各主要子类，保持分布相对均衡；
   3) 特殊场景（如跨机型、跨系统）可另设泛化测试集。

A.5.3 语料使用路径及特色

A.5.3.1 使用方向

a) 智能设计优化
   1) 应用目标：通过历史设计方案及性能数据，建立设计参数与性能指标的映射关系，实现 AI 辅助或自动优化设计（如汽车结构轻量化与强度平衡）
   2) 语料构成：需融合多模态数据（CAD 模型、CAE 仿真结果、实验测量数据），覆盖多工况与设计约束，标签需高保真（如精确的应力分布评价）

3) 实施流程：

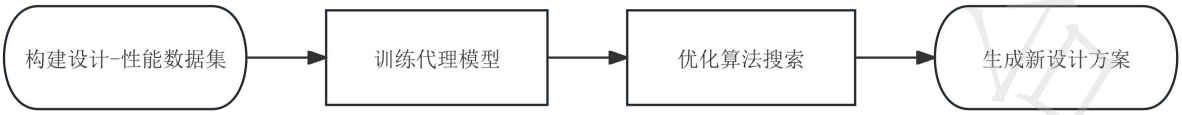

图 A.1 教学视频语料数据关联形式

b) 预测性维护
   1) 应用目标：基于设备运行历史数据与故障案例，训练 AI 模型实现异常预警与故障预测（如风机轴承健康监测）；
   2) 语料构成：传感器时序数据（振动、温度等）、维护记录及精确故障标注（需关联异常信号与具体部件故障）；
   3) 实施流程：

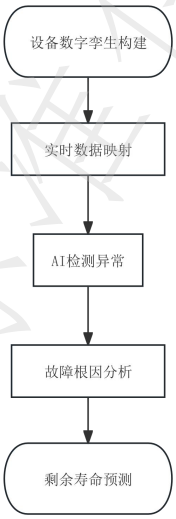

图 A.2 预测性维护实施流程

A.5.3.2 使用案例

在大飞机运维场景下，人工智能系统需支撑多类型健康管理任务，涵盖从故障监测到趋势预测的不同智能决策需求。为保障各类任务的准确性、稳定性与可复用性，需依据任务目标组织相应的数据处理路径，涵盖数据采集、清洗、标注、测试与使用全过程。典型任务包括如下几类：

a) 异常监测任务：实时识别飞行过程中是否存在传感器异常或系统运行异常，为故障定位与保障响应提供先导信息。

表 A.31 工程技术科学异常监测任务数据处理流程

数据处理流程	
数据采集	以飞行架次为单位采集高频传感器数据，覆盖发动机、引气、电气、液压等核心系统
数据清洗	剔除漂移信号、死值通道、异常突变点，保留完整有效的运行区间
标签标注	依据运维系统记录与专家复核标记“是否异常”标签，可采用弱标签机制（如段级标注）
训练测试划分	避免同一飞机或时间段出现在训练与测试集中，异常样本覆盖典型工况与系统
模型使用	用于部署在飞行实时监控系统中，完成滑窗级别的在线异常检测，辅助告警触发

b) 故障诊断任务：在检测到异常后，进一步识别具体的故障类型与系统部位，支撑精准维修与故障溯源。

表 A.32 工程技术科学故障诊断任务数据处理流程

数据处理流程	
数据采集	保留完整异常片段前后的时序数据，同时采集飞行状态参数与工况信息
数据清洗	重点关注异常区域数据质量，排查因采样误差、丢帧等导致的偏差
标签标注	建立主类-子类两级标签体系（如：引气系统 → 过热、泄漏），确保诊断输出的层次性
训练测试划分	按故障类型均衡抽取，确保各子类诊断样本覆盖充分
模型使用	可用于离线诊断报告生成或嵌入远程健康监控平台，用于判断当前异常是否指向某类系统性故障

c) 故障预警任务（架次级）：在当前架次执行前或任务调度阶段，基于历史运行信息，提前识别高风险架次，形成维护或调度建议。

表 A.33 工程技术科学故障预警任务数据处理流程

数据处理流程	
数据采集	基于历史3～5架次的飞行传感器数据构建输入序列，同时采集任务信息、天气条件、飞行状态等
数据清洗	确保每个历史架次数据完整且无缺失，必要时可引入数据插值或补全机制
标签标注	目标为“下一个架次是否发生故障”，标签以航班级别生成，不得基于当前架次滞后标注
训练测试划分	保证训练数据时间早于测试数据，避免未来信息泄露，建议跨机型构建泛化测试集
模型使用	结合趋势分析与风险识别，在航线调度或航前检测阶段部署，生成预警评分或风险等级

d) 剩余寿命预测任务（RUL）：基于历史运行趋势，预测系统或部件剩余可用时长，支持计划性维护与生命周期评估。

表 A.34 工程技术科学剩余寿命预测任务数据处理流程

数据处理流程	
数据采集	需包含完整退化过程或多次小幅异常过程的连续数据，覆盖足够时间跨度
数据清洗	强化平滑处理与特征增强，排除数据缺失对趋势建模的干扰
标签标注	以时间或循环数方式标注剩余寿命（如距故障发生还有X分钟/X架次）
训练测试划分	确保测试样本包含完整退化过程，用于验证预测趋势的一致性与稳定性
模型使用	广泛应用于地面健康管理平台，用于提前规划部件更换周期或进行寿命评估报告生成

A.5.4 数据安全措施

典型的工程技术科学领域数据安全措施包括但不限于：

a) 工业控制系统数据保护：针对工业生产线、能源网络传感器数据和操作日志建立网络隔离和访问控制机制，采用工业防火墙和入侵检测系统，实施操作行为的实时监控和审计，并建立异常模式识别和自动响应系统；

b) 大型工程结构设计数据保密：为保护大型桥梁、水坝、核电设施等工程设计知识产权，建立设计数据的分级保护制度，采用数字水印和版权保护技术，实施CAD文件的加密存储和传输，并建立设计变更的审批和追踪机制；

c) 供应链与物流网络数据安全：针对关键制造业供应链数据，建立供应商安全评估和准入机制，采用区块链技术确保供应链数据的透明性和不可篡改性，实施供应链关键节点的重点保护，并建立供应链中断的风险预警和应急响应机制；
d) 先进制造工艺参数保护：为防止增材制造、半导体制造、复合材料工艺参数泄露，建立工艺数据的严格访问控制和加密保护，采用安全多方计算技术实现数据共享而不泄露原始参数，实施工艺参数的版本管理和变更控制，并建立技术秘密保护的法律和合规框架；
e) 关键基础设施运行数据防护：针对电网拓扑、工业控制协议、基础设施监测数据，建立多层次防御体系，采用网络分段和微隔离技术，实施基于行为的异常检测，建立关键数据的实时备份和灾难恢复机制，并制定网络安全事件的应急响应预案。

参 考 文 献

[1]中国分类主题词表（第二版）
[2]中国图书馆分类法
[3]中国档案分类法
[4]GA/T 1717.3-2020 信息安全技术 网络安全事件通报预警 第3部分：数据分类编码与标记标签技术体系技术规范
[5]SF/T 0153-2023 图片真实性鉴定技术规范
[6]YD/T 3470-2019 面向公有云服务的文件数据安全标记规范