

DeepSeek 赋能科研创新

AI 大模型在科研全流程中的实践

韦玮

w.w@taoxiang.org

2025 年 3 月 28 日

讲师简介

● 韦玮

- 上海市科评院专家
- 上海市经济和信息化委员会专项资金评审专家
- 计算机畅销图书作家，高级工程师职称
- 硕士毕业于英国阿伯丁大学人工智能专业
- 曾担任同程数科（同程集团旗下）大数据平台高级研发工程师
- 曾担任重庆韬翔网络科技有限公司联合创始人兼技术部门负责人

● 独立技术顾问

- 为花旗银行、山东高速集团、中国银行、广州证券、东莞农业银行、深圳建设银行、平安寿险等多家知名企业提供内训 & 咨询服务

● 已出版图书

- 《Python 程序设计基础实战教程》
- 《Python 基础实例教程》
- 《精通 Python 网络爬虫》
- 《Laravel 入门与实战》（第一版）（第二版）

● 其他社会任职

- 广东省人工智能产业协会粤港澳大湾区人工智能产业专家
- 中国知网（CNKI）评审专家库专家
- 中国国际经济技术合作促进会标准化工作委员会专家库专家

目录

- 1 DeepSeek 科研能力解析
- 2 DeepSeek 科研场景部署方案
- 3 科研全流程赋能案例
- 4 学术伦理与前沿展望

- **痛点分析:**

- 海量文献难以筛选
- 跨语言文献阅读障碍
- 研究热点难以捕捉
- 文献管理效率低下

- **DeepSeek 解决方案:**

- 智能文献检索: 快速定位相关文献, 支持多语言交叉分析
- 研究热点预测: 生成论文趋势图谱, 实时追踪领域动态
- 文献管理与引用: 自动生成综述, 智能引文规范核查

- **实际案例:**

- 跨语言文献智能综述: 中英文文献交叉分析与智能生成 (案例部分会详细介绍)
- 研究热点预测: 助力实验室调整方向 (案例部分会详细介绍)

实验设计痛点解析

科研实验设计中的常见问题

- **痛点 1：实验方案生成效率低**
 - 传统方法依赖人工经验，耗时长。
 - 多变量组合难以全面覆盖。
- **痛点 2：实验资源分配不科学**
 - 资源分配缺乏数据支持。
 - 导致资源浪费或不足。
- **痛点 3：实验过程缺乏实时优化**
 - 发现问题后调整方案效率低。
 - 依赖人工判断，容易出错。
- **痛点 4：跨学科实验设计复杂**
 - 需要整合多领域知识。
 - 传统方法难以快速生成融合方案。

DeepSeek 在实验设计中的具体应用

结合 DeepSeek 的功能解决科研痛点

● 功能 1：智能实验方案生成

- 基于多目标优化算法，自动生成实验方案。
- 支持多变量、多约束条件。
- **结合 DeepSeek**：利用 DeepSeek V3 的通用模型能力，快速生成科学、高效的实验方案。

● 功能 2：实验资源优化分配

- 根据实验目标和历史数据，智能分配资源。
- **结合 DeepSeek**：通过 DeepSeek R1 的推理能力，实时预测资源需求，动态调整分配策略。

DeepSeek 在实验设计中的具体应用（续）

结合 DeepSeek 的功能解决科研痛点

● 功能 3：实时实验监控与优化

- 提供实时数据分析功能，实验过程中发现问题可即时调整方案。
- **结合 DeepSeek**：利用 DeepSeek 的实时推理能力，实现低延迟的监控与反馈。

● 功能 4：跨学科实验设计支持

- 整合多学科知识库，支持跨学科实验设计。
- **结合 DeepSeek**：基于 DeepSeek 的知识图谱能力，提供跨领域知识融合与智能推荐。

DeepSeek 实验设计流程

从需求输入到结果优化的全流程支持

- **步骤 1：需求输入**

- 用户输入实验目标、变量范围、约束条件等。

- **步骤 2：智能方案生成**

- DeepSeek 自动生成多个实验方案，并提供科学依据和优化建议。

- **步骤 3：资源分配与预演**

- 系统根据方案智能分配资源，并模拟实验过程，预测结果。

- **步骤 4：实时监控与调整**

- 实验过程中，DeepSeek 实时监控数据，发现问题即时调整方案。

- **步骤 5：结果分析与优化**

- 实验结束后，系统自动分析结果，优化下一轮实验方案。

- **更高效的实验设计**

- 持续优化算法，进一步提升实验设计效率。
- 结合 Agent 技术，自动进行实验的执行与反馈。

- **更广泛的应用场景**

- 支持更多学科领域的实验设计需求。
- 推动科研创新，降低科研门槛。

● 痛点 1：数据量大且复杂

- 科研数据通常来自多个来源，格式多样，清洗和整合耗时。
- 高维数据分析（如基因序列、气象数据）对计算资源要求高。

● 痛点 2：分析工具门槛高

- 传统工具（如 R、Python）需要编程基础，学习曲线陡峭。
- 可视化工具（如 Tableau、Matplotlib）配置复杂，难以快速上手。

● 痛点 3：结果可解释性差

- 机器学习模型（如深度学习）常被视为“黑箱”，难以解释其决策过程。
- 缺乏透明化的分析过程，难以满足学术伦理要求。

● 方案 1：智能数据预处理

- 多源数据的自动处理（CSV、Excel、JSON 等）。
- 实现快速自动的数据清洗管道，自动处理缺失值、异常值和重复数据。

● 方案 2：低代码分析平台

- 提供可视化操作界面，支持拖拽式数据分析。
- 集成常用算法库（如 Scikit-learn、TensorFlow），无需编程即可完成复杂分析。

● 方案 3：透明化分析过程

- 提供推理过程暴露功能，实时展示模型决策路径。
- 生成可解释性报告，满足学术伦理要求。

论文写作痛点解析

科研全流程中的关键挑战

● 写作效率低下:

- 传统写作流程繁琐，耗时较长。
- 缺乏自动化工具支持，重复性工作多。

● 结构不清晰:

- 论文结构设计不合理，逻辑不连贯。
- 缺乏有效的结构化写作工具。

● 语言表达不精准:

- 学术语言表达不规范，影响论文质量。
- 缺乏智能化的语言优化工具。

● 图表生成复杂:

- 图表生成与优化过程繁琐，耗时较长。
- 缺乏自动化图表生成工具。

● 排版耗时:

- LaTeX 排版复杂，学习曲线陡峭。
- 缺乏智能排版工具支持。

- **结构化写作助手：**

- 提供论文结构模板，快速生成初稿。
- 支持多级标题自动生成与调整。

- **语言优化工具：**

- 智能检测语言表达，提供优化建议。
- 支持多语言翻译与校对。

- **图表智能生成：**

- 自动生成 Matplotlib 代码，优化图表展示。
- 支持图表样式自定义与一键导出。

- **智能排版 Agent：**

- 自动生成 LaTeX 代码，简化排版流程。
- 支持期刊模板一键适配。

- **期刊匹配系统：**

- 根据论文内容推荐合适期刊。
- 提供影响因子、审稿周期等关键信息。

目录

- 1 DeepSeek 科研能力解析
- 2 DeepSeek 科研场景部署方案
- 3 科研全流程赋能案例
- 4 学术伦理与前沿展望

API 使用

如何通过 API 集成 DeepSeek 到现有应用中?

各云厂商 API 的获取

- 支持主流云平台（如火山引擎、腾讯云）。
- 基于火山引擎部署。
- 实战演示云平台集成部署方式。



云厂商 API 集成

通过代码直接请求 API 使用

- 支持 Python、Java、Node.js 等多种编程语言。
- 演示示例代码（Python）实战。

小型科研团队：API 快速接入

演示 Jupyter Notebook 集成案例

- 步骤 1: DeepSeek API 请求函数的封装
- 步骤 2: 基于 OpenAI 库实现 DeepSeek 的调用
- 步骤 3: 基于 langchain 实现 DeepSeek 的调用（实战案例演示）

私有化部署：企业的优选

为什么私有化部署是企业的优选？

- **数据隐私保护**：企业数据无需外传，确保敏感信息安全。
- **大规模使用降低成本**：长期使用下，私有化部署比 API 调用更经济。
- **算力成本与盈亏平衡点**：需根据企业需求评估算力成本，找到最优解。

如何搭建私有化部署？

- **服务器选择**：
 - 云服务器租用：灵活便捷，按需付费。
 - 采购硬件 GPU 服务器：长期投资，性能更优。
- **部署技术**：
 - 模型选择：根据需求选择合适的模型。
 - 技术框架：如 vLLM，支持高效推理。
 - 多卡 vs 单卡：根据算力需求选择部署方式。

如何搭建私有化部署？（续）

- **RESTful API 封装：**

- 提供标准化接口，便于集成。
- 支持并发请求，提升服务效率。

- **鉴权机制：**

- 选择合适鉴权方式（如 OAuth2、API Key）。
- 确保服务安全，防止未授权访问。

- **监控与维护：**

- 部署监控工具，实时跟踪服务状态。
- 定期更新模型与框架，保持最佳性能。

注意事项

私有化部署需综合考虑成本、性能与安全性，确保满足企业需求。

部署方式对比

对比项	API 部署	私有化部署
适用场景	快速集成、轻量级应用	高安全性、定制化需求
部署复杂度	低，无需额外硬件支持	高，需搭建专用服务器
成本	按需付费，初始成本低	硬件和运维成本高
安全性	依赖第三方安全措施	数据完全自主控制，安全性高
性能	依赖网络环境，可能存在延迟	本地化部署，性能稳定
扩展性	快速扩展，无需额外配置	需手动扩展，灵活性较低
维护	由服务商维护，用户无需干预	需专业团队维护，运维压力大
优缺点	优点：快速、低成本；缺点：安全性依赖第三方	优点：高安全性、定制化；缺点：成本高

Table: API 部署与私有化部署的对比

敏感数据隔离方案

确保数据安全与合规

- **数据隔离策略：**

- **物理隔离：**通过专用服务器或独立硬件设备，确保敏感数据与公共网络的物理隔离。
- **逻辑隔离：**利用虚拟化技术（如 Docker、Kubernetes）创建独立的运行环境，隔离不同项目的数据。

- **访问控制：**

- **身份验证：**采用多因素认证（MFA）确保只有授权人员可以访问敏感数据。
- **权限管理：**基于角色的访问控制（RBAC），细化数据访问权限。

- **数据加密：**

- **传输加密：**使用 TLS/SSL 协议加密数据传输过程。
- **存储加密：**对敏感数据进行 AES-256 加密存储。

敏感数据隔离方案（续）

确保数据安全与合规

- **审计与监控：**

- **日志记录：**详细记录数据访问和操作日志，便于审计。
- **实时监控：**部署实时监控系统，及时发现并应对异常行为。

- **灾难恢复：**

- **备份策略：**定期备份敏感数据，确保数据可恢复性。
- **应急响应：**制定应急响应计划，快速应对数据泄露或系统故障。

- **合规性：**

- **法律法规：**遵守《网络安全法》和《数据安全法》等相关法律法规。
- **行业标准：**符合 ISO 27001 等国际信息安全标准。

实验数据预处理管道搭建

PyTorch + Docker 实战演示

- 目标：

- 构建高效、可复现的实验数据预处理管道。
- 通过容器化技术确保环境一致性。

- 工具：

- **PyTorch**：用于数据加载、转换和预处理。
- **Docker**：用于环境隔离与部署。

- 步骤：

- ① 数据加载与清洗。
- ② 数据转换与标准化。
- ③ 数据分割与批处理。
- ④ 容器化部署与测试。

实验数据预处理管道搭建（续）

PyTorch + Docker 实战演示

- **数据加载与清洗：**

- 使用 PyTorch 的 `DataLoader` 加载原始数据。
- 清洗无效数据，处理缺失值。

- **数据转换与标准化：**

- 应用数据增强技术（如旋转、缩放）。
- 标准化数据以加速模型收敛。

- **数据分割与批处理：**

- 将数据集分为训练集、验证集和测试集。
- 使用 `BatchSampler` 进行批处理。

实验数据预处理管道搭建（续）

PyTorch + Docker 实战演示

- 容器化部署与测试：

- 编写 Dockerfile 定义环境依赖。
- 构建 Docker 镜像并运行容器。
- 测试预处理管道的功能与性能。

- 优势：

- 可复现性：确保实验环境一致。
- 可移植性：轻松部署到不同平台。
- 高效性：并行处理加速数据预处理。

- 实战演示：

- 1.：基础容器镜像的创建。
- 2.：标准化 SOP 处理代码的编写与部署。
- 3.：容器镜像的再次封装。

目录

- 1 DeepSeek 科研能力解析
- 2 DeepSeek 科研场景部署方案
- 3 科研全流程赋能案例
- 4 学术伦理与前沿展望

跨语言文献智能综述

中英文文献交叉分析

- **多语言文献整合：**

- 支持中英文文献的自动翻译与对齐，确保跨语言文献的无缝整合。
- 利用深度学习模型进行语义匹配，提升文献关联度分析的准确性。

- **智能综述生成：**

- 自动提取文献核心观点，生成结构化综述。
- 支持自定义综述模板，满足不同学科需求。

- **案例演示：**

- 展示中英文文献交叉分析的实际案例，演示 DeepSeek 在跨语言文献处理中的能力。

研究热点预测

论文趋势图谱生成

- **趋势分析：**

- 基于海量文献数据，利用时间序列分析预测未来研究热点。
- 提供可视化趋势图谱，直观展示研究领域的发展动态。

- **热点识别：**

- 通过关键词共现分析，识别当前研究热点。
- 支持多维度热点分析，包括学科交叉、国际合作等。

- **案例演示：**

- 展示某领域研究热点预测的实际案例，演示 DeepSeek 在趋势分析中的能力。

- **学术不端检测：**

- 利用 AI 技术检测论文中的抄袭、篡改等学术不端行为。
- 提供详细的检测报告，帮助作者及时修正问题。

- **引文规范核查：**

- 自动核查引文格式是否符合期刊要求。
- 支持多种引文格式（APA、MLA、Chicago 等），确保引文规范。

- **案例演示：**

- 展示学术伦理守护的实际案例，演示 DeepSeek 在学术诚信维护中的能力。

● 实验设计优化：

- **自动化参数调优：**基于历史实验数据与模型预测，自动推荐最优超参数（如学习率、批量大小等）。
- **多目标优化：**支持多目标优化算法（如 NSGA-II），帮助在模型精度、训练时间、资源消耗之间找到平衡。
- **实验方案生成：**根据研究目标（如模型性能提升）自动生成实验方案，减少人工设计时间。

● 实验流程管理：

- **实验进度跟踪：**实时监控训练过程，自动生成实验报告（如损失曲线、准确率变化）。
- **资源调度优化：**智能调度 GPU/CPU 资源，提高资源利用率。

● 案例演示：

- **案例 1：**基于 DeepSeek 的深度学习模型超参数优化。
- **案例 2：**分布式训练中的资源调度优化。

故障诊断：实验异常数据实时分析预警

确保实验数据的可靠性与稳定性

● 异常检测：

- **实时监控**：实时监控训练数据流，自动检测异常数据点（如梯度爆炸、NaN 值）。
- **多维度分析**：结合损失函数、梯度分布等多维度数据进行异常分析。
- **预警机制**：设置阈值（如损失值突增），自动触发预警通知，及时处理异常。

● 故障诊断：

- **根因分析**：通过机器学习算法，自动分析异常数据的根本原因（如数据分布偏移、模型过拟合）。
- **修复建议**：提供针对性的修复建议（如调整学习率、增加正则化），帮助快速解决问题。

● 案例演示：

- **案例 1**：深度学习训练中的梯度爆炸检测与修复。
- **案例 2**：数据预处理中的异常值检测与处理。

智能实验平台集成

无缝集成现有实验平台

- **平台兼容性:**

- **API 接口:** 提供标准化 API 接口, 支持与主流深度学习框架 (如 PyTorch、TensorFlow) 无缝集成。
- **插件支持:** 支持 Jupyter Notebook、VS Code 等开发环境的插件开发, 方便快速接入。

- **数据安全:**

- **数据加密:** 采用高级加密标准 (AES) 确保数据传输与存储安全。
- **权限管理:** 细粒度的权限控制, 确保数据访问的安全性。

- **用户界面:**

- **可视化仪表盘:** 提供直观的可视化仪表盘, 实时展示训练状态与数据分析结果 (如损失曲线、资源占用)。
- **自定义报告:** 支持自定义实验报告模板, 自动生成符合要求的报告 (如模型性能对比、资源消耗分析)。

- **结构化写作助手：**

- 自动生成论文大纲，提供写作框架。
- 根据输入数据生成初稿，支持多轮迭代优化。
- 提供语法检查、风格建议和内容优化。

- **图表智能生成：**

- 自动生成 Matplotlib 代码，支持多种图表类型。
- 提供图表优化建议，提升可视化效果。
- 支持数据更新后自动重新生成图表。

图表智能生成案例

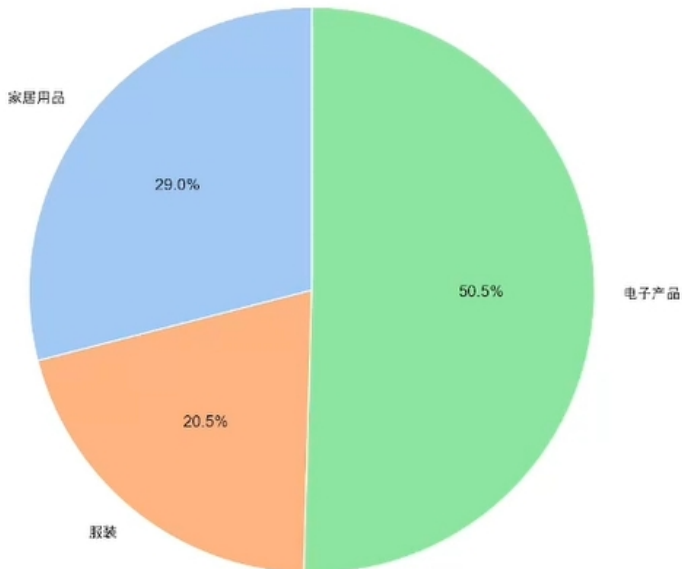
数据预处理与清洗、自动数据可视化实战

核心功能

- **数据预处理与清洗：**DeepSeek 提供智能化的数据清洗工具，自动识别并处理缺失值、异常值、重复数据等问题，确保数据质量。
- **自动数据可视化：**通过 DeepSeek，用户可以一键生成多种可视化图表（如折线图、柱状图、散点图等），快速洞察数据趋势和规律。
- **高效分析：**DeepSeek 内置强大的机器学习算法，支持自动化的数据分析和预测，显著提升办公效率。

图表智能生成案例

数据预处理与清洗、自动数据可视化实战



- **期刊匹配系统：**

- 根据论文内容推荐合适期刊。
- 提供影响因子、审稿周期和风格契合度三维推荐。
- 支持多期刊对比分析，帮助选择最佳投稿目标。

- **智能排版 Agent：**

- 自动生成 LaTeX 排版代码，支持多种期刊模板。
- 提供排版优化建议，确保符合期刊要求。
- 支持一键导出 PDF，简化投稿流程。

论文智造实战演示

从数据到初稿的全流程演示

- **数据输入：**
 - 导入实验数据，选择分析目标。
- **初稿生成：**
 - 自动生成论文大纲和初稿。
 - 提供语法检查和内容优化建议。
- **图表生成：**
 - 自动生成 Matplotlib 代码，优化图表效果。
- **期刊匹配：**
 - 推荐合适期刊，提供对比分析。
- **智能排版：**
 - 自动生成 LaTeX 排版代码，导出 PDF。

目录

- 1 DeepSeek 科研能力解析
- 2 DeepSeek 科研场景部署方案
- 3 科研全流程赋能案例
- 4 学术伦理与前沿展望

● 模型优化与扩展层面

- 提升模型的多语言处理能力，支持更多小语种文献分析。
- 优化模型在跨学科领域的适应性，增强对复杂数据的理解能力。
- 更高效的推理模型，更低的计算成本。

● 科研生态建设层面

- 构建开放的科研 Agent 工具平台，支持更多第三方插件与工具集成。
- 丰富科研社区，促进学术资源共享与协作。

● 伦理与合规层面

- 推出更严格的学术伦理检测工具，确保科研过程的合规性。
- 加强数据隐私保护，推出敏感数据隔离与加密方案。
- AI 科研伦理标准的制定，国际学术伦理的标准制定。

- 互动环节：解答参与者问题。
- 总结：回顾 DeepSeek 在科研中的关键应用, 更多可能性的探索。