贾天瑞

北京市 | j18614269732@163.com | jiatianrui1.github.io

出生年月：2002.12　籍贯：北京

教育经历

北京理工大学 | 数学 / 硕士研究生（在读）　2024.09—2027.06

北京理工大学 | 数学与应用数学 / 学士　2020.09—2024.06

- **主修课程**：系统学习机器学习、深度学习、数学等课程。
- **荣誉奖项**：北京理工大学奖学金（一等奖、二等奖）；全国大学生数学竞赛省级三等奖；全国大学生英语竞赛三等奖。

技术能力与特长

- **大模型算法与微调**：深入理解 **Transformer** 架构及 Attention 机制；熟练掌握 **SFT**、**LoRA**、**P-Tuning v2**、**Adapter** 等微调技术；具备 **Qwen**、**LLaMA**、**GLM** 等主流模型的微调与推理优化经验。
- **RAG 与 Agent 开发**：熟练掌握 **RAG 全链路开发**（文档解析、分块策略、混合检索、Rerank）；掌握 **LangChain**、**LlamaIndex** 框架；掌握 **Function Call**、**ReAct**、**CoT** 等 Agent 规划策略；熟悉 **Milvus**、**Faiss** 向量数据库及 **Neo4j** 图数据库应用。
- **多模态与前沿架构**：熟悉 **Diffusion Model**（含 **Rectified Flow**）原理及应用；具备音频信号处理（**EMD**、**频谱分析**）及多模态融合（**EEG+Audio+Text**）建模经验。
- **框架与工程化**：熟练掌握 **PyTorch** 与 **GBDT** 等机器学习模型，熟悉 **C++/MATLAB**；掌握模型压缩（**INT8/INT4 量化**、**剪枝**、**蒸馏**）；熟悉 Linux 环境开发与 **Docker** 容器化部署。
- **英语读写能力**：CET-4 588，CET-6 542；具备较强的英文论文阅读与写作能力。

实习经历

理想汽车 | 大模型实习生 / 智能汽车群组-软件本体-理想同学-理想同学智能体部　2026.02-2026.04

车载音乐个性化推荐检索排序系统

- **项目描述**：面向车机端音乐个性化推荐场景，基于用户侧 **Query + 用户画像 + 播放场景**与歌曲侧内容特征，构建 **双塔召回 → ANN 粗排 → Diffusion + DIN 精排**的多阶段推荐链路，实现低延迟、高相关推荐。
- **数据理解与特征工程**：整合 **11.86 万**用户画像、**526 万**播放行为及 **21.44 万**歌曲特征，将用户侧多源信息拼接为统一音乐语义表达，并与歌曲侧文本对齐，用于双塔训练与线上检索。
- **Embedding 模型训练与 ANN 粗排**：基于 **Qwen3-Embedding** 进行双塔监督微调，训练中采用 **In-Batch Negative**；离线生成 **21 万 +** 歌曲向量并构建 **Faiss IVF-PQ** 索引缓存，支撑千万级向量检索与 TopK 粗排召回。
- **文本改写增强**：基于**全参数微调 Qwen2.5-1.5B** 对 Query/Item 文本进行改写增强，补充音乐风格、情绪与场景语义，提升 Query-Item 语义对齐效果（**Recall@200 +4.6%**）。
- **生成式推荐与精排优化**：精排阶段引入 **Diffusion** 生成用户偏好语义表示，提升长尾内容分发质量；结合**历史搜索/播放序列**设计 **DIN** 兴趣激活机制，对候选进行相关性加权建模（**NDCG@10 +4.2%**）。
- **阶段性效果（离线/回放口径）**：相比基线模型，**NDCG@10 提升 6.8%**、**MRR@10 提升 9.3%**，**Recall@200 提升至 87.4%**；召回延迟稳定控制在 **100ms** 内。

理想同学音乐 Agent 深度规划

- **训练数据管理**：收集开源 Agent / Tool-Use 数据集，采用 **LLM** 进行过滤筛选，保留符合车载音乐场景的任务数据；围绕搜歌、点播、推荐、切换播放模式、上下文续播等业务场景，采用 **back-instruct** 方法构造多步任务数据，累计合成高质量训练数据 **1W+**。
- **规划模型算法设计**：协助完成基于 **MCP Server** 的 **Plan-Execute** 规划算法设计，涵盖初次规划、反思规划、工具调用、结果校验及思维链提示词设计；针对音乐场景下的多轮确认、槽位补全、上下文承接等任务，开展 **SFT**、**GRPO**、**GSPO**、**DAPO** 等多种训练方式实验，提升复杂音乐指令下的规划准确率与任务成功率。
- **投机采样模型训练**：为满足车载端低延迟响应需求，基于 **eagle-3** 算法完成适配音乐 Agent 的投机解码头训练，提升推理吞吐与响应速度，保障复杂音乐任务在端侧场景下的实时交互体验。

项目经历

NeurIPS 2025 Google Code Golf **锦标赛** | **自动化代码压缩** Agent | 核心开发者 2025.09—2025.11
github.com/JiaTianrui1/codegolf

- **项目描述**：针对 ARC-GEN 任务构建自动化代码压缩与验证框架，最终获得全球第 **13** 名（银牌区第 **1** 名）。
- **Agent 闭环设计**：设计基于 ARC-GEN 集成 Codex Agent 的多智能体协作流程，实现“候选解生成 → 正确性校验 → 最优解筛选”的自动化闭环，累计完成 **2000+** 次自动迭代。
- **提示工程与压缩策略**：设计 **zlib + Base85** 通用压缩壳并结合 Prompt 工程压缩 solver 代码，驱动 **DeepSeek**、**GPT-5**、**Qwen** 等多模型多轮代码逻辑重构。
- **工程落地**：搭建基于 **GitHub Actions** 的 CI/CD 流水线，实现 PR 自动合并、版本管理及 Kaggle API 自动提交。

EduRAG | **基于** RAG **的教育场景智能问答系统** | 独立负责人 2025.03—2025.08
github.com/JiaTianrui1/integrated-qa-system

- **项目描述**：从 0 到 1 构建面向教育场景的智能问答系统，整合课程、政策等多源异构知识，实现高准确率自动应答。
- **检索增强架构**：基于 **Milvus** 构建向量索引，设计 PDF/Word/PPT 多格式文档清洗与 OCR 提取流水线，累计处理 **50 万** + 知识片段；实施稀疏 + 稠密混合检索与 **BGE-Reranker** 重排序策略，提升长尾知识召回效果。
- **意图识别与生成**：基于 **Sentence-BERT** 构建意图分类模型（**F1=91.27%**）；设计历史对话融合与 **HyDE** 的 Query 重写模块；利用 **Qwen2.5-7B** 进行指令微调，控制回答格式与幻觉。
- **部署与评估**：构建 **RagAS** 自动评估（答案相关性 **0.91**，忠实性 **0.93**），基于 **1000** 条标注测试集验证整体准确率 **89.74%**；通过 **FastAPI** 封装服务并上线，自动应答率达 **81.56%**。

科研成果

- **[BSPC, Q2 | 已录用 | 第一作者] Breaking Through Data Scarcity: A Novel Diffusion Model Approach for Snoring Sound Augmentation**
  提出基于扩散模型的原始波形增强框架，并首次将 **Rectified Flow** 引入音频生成，构建端到端模型 **RFSSDiff** 及数据增强策略，在 MPSSC 数据集上将 UAR 提升至 **62.5%**（**+5.8%**），显著提升低资源场景下的鼾声分类性能。
- **[INTERSPEECH 2025, CCF-B | 主要贡献者] Exploring EMD for Sensing the Sound of Silence: Mice Autism Detection via Ultrasonic Vocalization**
  构建融合 **EMD** 与频谱图的多分支网络，实现超声叫声的多尺度建模与可解释特征提取，验证高频 IMF1 是区分 ASD/WT 的关键声学标志，并在 MADUV 数据集上取得显著优于基线的方法效果（$\mathbf{p<0.05}$）。
- **[ESWA, Q1 TOP | 共同一作 | 在投] Robust EMG Forecasting for IoT Healthcare via Adaptive Mapping & Normalization-Free Transformer**
  提出基于 **MNN** 的 EMG 分解映射方法，并结合 **RLP-RIME** 自动超参搜索与 **DyT** 无归一化时序建模框架，使模型在多个任务上平均 R^2 提升 **15.5%**、推理延迟降低 **61.3%**，整体性能超越 TimeMixer 等 SOTA 方法。
- **[Pattern Recognition, Q1 TOP | 学生二作 | 大修] Bayesian Knowledge-Guided Confidence-Aware Method for Depression Detection**
  提出 **BLADE** 抑郁识别框架，将 LaBraM（EEG）与 Emo2Vec（Audio）进行知识引导下的多模态融合，并引入贝叶斯不确定性建模与潜空间扩散机制，在 MODMA 数据集上取得 **100%** 被试级和 **88.12%** 样本级准确率。

校园经历

北京理工大学校足球队 | 副队长 2022.09-2023.09

- 负责协助队伍日常训练组织、队内沟通协调与比赛准备，推动球队训练与参赛安排有序开展。
- 曾带领队伍参加北京市比赛并获得第四名；组织球队参加区级比赛并夺冠，体现了较强的团队协作、组织协调能力。

北京理工大学寒假社会实践团 | 队长 2023.01-2023.03

- 负责团队组织与任务分工，协调成员与指导教师之间的沟通，推动实践活动按计划开展。
- 统筹实践过程中的安排与总结工作，独立完成最终总结材料，具备一定的组织管理、沟通协调与落地执行能力。

自我评价

- **数理基础扎实**：数学系科班出身，对深度学习背后的优化理论、概率统计有扎实理解，能快速复现顶会论文算法。
- **全栈算法能力**：具备从数据清洗、模型微调（SFT/RLHF）、RAG 构建到 Agent 开发、量化部署的全链路实战经验。
- **持续学习热情**：密切关注 LLM 前沿动态，具备良好的英文论文阅读与技术调研能力。
- **团队协作能力**：沟通顺畅，能在复杂任务中主动承担责任并推动项目进展。