陈熹

153-2229-4107 | jtydfcx@163.com | https://xichennn.github.io

个人简介

拥有系统工程博士学位与统计学双硕士背景，兼具深厚的数学建模功底与 5 年+ 机器学习系统工程化经验。近年来专注于 LLM 与 RAG 系统研发，主导 LexisNexis 法律智能问答系统中多个核心功能交付，实现超 10%性能提升。曾在深度学习，统计机器学习等方向开展多项研究，既能深入底层模型，又能驾驭复杂的上层应用编排，致力于构建精准、可靠的下一代生成式 AI 应用。

教育经历

亚利桑那大学 Tucson, AZ, USA

- 博士, 系统与工业工程 (全额奖学金，4.0/4.0) 2018.8 - 2024.8
- 硕士, 统计与数据科学(4.0/4.0) 2020.8 - 2023.5

北京航空航天大学 北京, 中国

- 硕士, 控制科学与工程 2015.9 - 2018.3
- 本科, 质量与可靠性工程 2011.9 - 2015.6

北京大学 本科, 国家发展研究院经济学双学位 2014.9 - 2017.6

工作经历

高级算法工程师 - LexisNexis 2024.9 – Present

- 参与构建基于 **RAG** 框架的法律**智能问答**系统，实现对用户上传法律文档的自动问答。开展**文本分块**策略，**混合检索**与**重排**序实验，提升检索精度；设计并优化生成**提示词**，使生成回答的**有用性 (Usefulness)** 指标提升超 10%。
- **主导时间轴（Timeline）功能交付**。带领一个小团队完成事件时间轴生成功能交付。与产品经理紧密协作明确用户意图，通过真实用户查询分析，将 LLM API 调用次数减少 **80%**。该功能在内部答案质量评分较竞品领先 **15%**，在 Precision 和 Recall 上均提升超过 **10%**。
- **主导文档定位（Doc Locator）功能交付**。**从 0 到 1 架构**并落地 Doc Locator 功能。通过快速原型迭代与多方案基准测试，协调跨职能团队完成最终落地实施， 性能超过人工审核基线标准 **13%**。将该模块作为“**Citation Agent**”集成至自构建的 **ReAct** 式**智能体**系统中，通过引入验证闭环，确保生成的答案严格**锚定于具体的文档坐标**，实现精准溯源。
- **自动化评测体系**：构建时间轴和文档定位对应的**自动评估**方案，与领域专家一起构建评估数据集，并开发基于 Prompt 的程序化自动评测工具，将**开发迭代周期缩短了 70%**。
- 代表团队向全公司进行技术汇报，系统性展示整体架构与成果。

科研经历

不确定性量化与多源序列建模 2023.6-2024.5

- 提出了一种端到端**序列建模**框架，融合车路协同 (V2I) 场景下的多视角基础设施数据，显著增强了车联网环境中的协同感知 (Cooperative Perception) 能力。
- 设计了基于共形预测 (Conformal Prediction) 的**不确定性量化**模块，用于校准模型的置信区间。该方法为模型输出提供了统计学层面的覆盖保证 (Statistical Coverage Guarantees)，这是在安全关键领域 (Safety-Critical Domains) 部署 AI 系统的关键要素。
- 工程化实现了一种自定义神经网络架构，集成了用于空间编码的图神经网络 (**GNNs**)、用于特征交互建模的跨图注意力模块 (**Cross-Graph Attention**) 以及**多模态**解码器。
- 在 V2X-Seq 基准测试中，相较于业内最先进 (SOTA) 的基线模型，将误差率 (FDE/MR) 降低了 4%。
- 构建了分布式训练流水线，利用 SLURM 在 HPC 集群上进行调度，实现了大规模超参数调优与高效实验。项目代码已开源：https://github.com/xichennn/V2I_trajectory_prediction

基于 Transformer 的多智能体轨迹预测 Intel 校企合作 2021.6-2023.6

- 构建支持**车联网**环境下的多模态轨迹预测深度模型，融合车载与通信感知数据。 在高联网车渗透率场景中，预测精度提升 **5%**，优于现有主流方法，如 LSTM, VectorNet, HiVT。
- 提出基于混合交通场景的**深度学习**框架。使用以智能体为中心的数据表示方法统一特征对齐。应用 **Transformer encoder** 获取时序编码，图注意力网络(**GAT**)获取空间交互及地图编码。设计**多模态解码器**，输出混合高斯模型 (GMM) 分布，实现了对轨迹多模态特性的概率化建模。
- 训练数据采集自 CARLA 仿真器。引入感知误差与通信噪声模拟真实世界数据。

自动驾驶系统中的故障传播建模研究 2021.8-2021.12
- 用 **OpenCV** 向摄像头图像中注入雨天、玻璃破裂、高斯噪声等**视觉**扰动，模拟真实驾驶场景下的输入故障。
- 集成 **YOLO** 目标检测模型评估感知系统在视觉退化条件下的鲁棒性，并量化检测性能的下降趋势。
- 监测自动驾驶系统各模块（感知、规划、控制）在故障输入下的输出偏移变化，捕捉系统级响应。
- 将故障引起的行为漂移建模为**泊松过程**（Poisson Process），用于预测故障在系统中的传播概率与时序分布，支持系统稳定性评估与容错机制设计。

Bayesian Optimization & Tensor Regression 2018.8-2021.5
- 提出一种新型的**高斯过程协方差核函数**，能够有效保留张量形状空间数据的结构信息，用于标量-张量回归任务中的**监督学习**。 将高斯过程与**贝叶斯优化**结合，在有限样本下高效搜索最优参数设计。
- 模型成功应用于 3D 打印天线设计预测。在仿真设计打印出的 3D 实体上进行性能测试，验证了模型在保留空间结构相关性的同时显著提高了预测准确性与样本利用效率。为结构化数据建模与工程优化提供了可行路径。

大模型项目经历

基于 Agentic RAG 的法规时效性校验与自动修正系统
- 构建基于 **LangGraph** 的“**时效性校验**”工具 (Temporal Verification Pipeline)，用于自动审查历史判例（Caselaw）片段。系统通过识别**关键数值实体**（如金额上限、诉讼时效周期），并与实时的“现行法规” API 进行交叉验证，从而精准识别已发生修订的法律条款。
- 设计基于 AsyncIO 的“**惰性验证**” (Lazy Verification) 机制，仅在检测到高风险实体时触发外部工具调用。该机制通过跳过对非法规性或纯定性文本片段的验证循环，显著**降低了系统延迟**。
- 开发自动化的“比对与标注” (Compare-and-Annotate) 状态模块，计算判例引用的历史数值与现行法规标准的差异。系统会在 Context Window 中**动态注入警示信息**，有效防止大模型将已过时的法律先例误判为有效依据，导致答案出现错误信息。

应用 LoRA 技术微调多模态 BLIP 模型以适应特定的自驾数据集
- 在一个小型图像-文本自驾数据集上，**微调** BLIP 模型，以提高模型在特定驾驶场景下的 VQA 任务。采用 **LoRA** 技术高效微调，减少内存占用并优化训练性能
- 通过集成 Hugging Face 的 Transformers 库和 **PyTorch** 框架，实现数据预处理，模型训练和评估的端到端流程

专业技能

编程语言: Python, SQL, R, MATLAB, Git, CMake, ROS, SAS
大模型： GPT (OpenAI, Anthropic, etc.), Hugging Face Transformers & Datasets, LLaMA, FAISS, prompt engineering, LangChain, AutoGen, ReAct agents, LoRA fine-tuning
深度学习: PyTorch, TensorFlow, Keras, PySpark, Pandas, NumPy, SciPy, scikit-learn
数据可视化: Matplotlib, Plotly, Seaborn, Power BI, Tableau, ggplot

Xi Chen

153-2229-4107 | jtydfcx@163.com | https://xichennn.github.io

SUMMARY

Ph.D.-trained **Data Scientist** with 5+ years of experience in developing scalable machine learning systems across autonomous driving, LLMs, and RAG. Adept in bridging academic research and real-world deployment, with hands-on expertise in sequence modeling, uncertainty quantification, and agent-based architectures. Delivered high-impact solutions in industry, leading cross-functional ML product development with measurable gains in accuracy, efficiency, and user value.

EDUCATION

The University of Arizona, Tucson, AZ

- Ph.D., Systems and Industrial Engineering (full scholarship, 4.0/4.0) Aug 2018 - Aug 2024
- M.S., Data Science and Statistics (4.0/4.0) Aug 2020 - May 2023

Beihang University, Beijing, China

- M.S., Control Science and Engineering Sep 2015 - Mar 2018
- B.S., Quality and Reliability Engineering Sep 2011 - Jun 2015

Peking University, Beijing, China

- B.S., Economics (Dual Degree) Sep 2014 - Jun 2017

WORK EXPERIENCE

Senior Data Scientist – LexisNexis Sep 2024 - Present

- **RAG System Architecture**: Architected a legal chatbot using a Retrieval-Augmented Generation (RAG) pipeline. Executed hybrid search (keyword + semantic) and custom re-ranking strategies to maximize retrieval precision. Iteratively optimized generation prompts, increasing answer usefulness metrics by >10%.
- **Timeline Feature Leadership**: Led a squad to deliver a high-impact "Timeline" generation feature. Collaborated with Product Managers to map user intent to query patterns, reducing LLM API costs by 80%. The feature boosted internal quality ratings by 15% over competitors, improving both Precision and Recall by >10%.
- **Doc Locator & Citation Agent**: Prototyped and benchmarked the "Doc Locator" feature, which outperformed the Human Review Test Baseline by 13%. Integrated this module as a Citation Agent within a custom ReAct-style workflow, implementing a verification loop to strictly ground generated answers in specific document coordinates.
- **Evaluation & Ops**: Built a domain-specific evaluation dataset and implemented a programmatic, prompt-based evaluation tool, shortening development cycles by 70%.
- **Cross-Functional Impact**: Coordinated cross-functional teams to drive the Doc Locator from prototype to production under tight deadlines. Presented technical deep-dives of the overall solution and outcomes to the entire company, representing the team.

RESEARCH EXPERIENCE

Research Associate - The University of Arizona Jun 2023 - Aug 2024

Uncertainty Quantification & Multi-source Sequence Modeling

- Proposed an end-to-end **sequence modeling** framework that fuses multi-view infrastructure data from Vehicle-to-Infrastructure (V2I) sources to enhance cooperative perception in connected driving environments.
- Designed a rigorous uncertainty module using **Conformal Prediction** to calibrate model confidence intervals. This ensured statistical coverage guarantees for model outputs, a critical factor for deploying AI in safety-critical domains.
- Engineered a custom architecture combining **Graph Neural Networks (GNNs)** for spatial encoding, a **Cross-Graph Attention** module for feature interaction, and a multimodal decoder.
- Achieved a **4% reduction** in error rates (FDE/MR) compared to State-of-the-Art baselines on the V2X-Seq benchmark.
- Implemented distributed training pipelines using **SLURM on an HPC cluster**, enabling large-scale hyperparameter tuning. The work is open-sourced at: https://github.com/xichennn/V2I_trajectory_prediction

Research Assistant - U of Arizona x Intel Autonomous Driving Research Collaboration May 2021 - Dec 2022

Transformer-based Multi-agent Trajectory Prediction

- Designed a deep learning framework for multimodal trajectory forecasting using heterogeneous sensor and communication data.
- Fused inputs through a **cross-attention** module, modeled temporal dependencies with **Transformers**, and spatial interactions with **Graph Attention Networks** (GAT).
- Simulated driving scenarios in CARLA, introducing synthetic sensor noise and communication dropouts to emulate real-

world imperfections.

- Implemented an agent-centric data representation to maintain alignment across time steps and agent roles.
- Demonstrated a **5%** improvement over baseline models in prediction accuracy, particularly under high connected-vehicle penetration scenarios.

Robustness Testing & Error Propagation Aug 2021 - Dec 2021

- Simulated real-world visual faults by injecting image-level corruptions (e.g., rain, broken windshield, Gaussian noise) into camera inputs using **OpenCV.** Integrated the **YOLO** object detection model to evaluate the robustness of the perception system under degraded visual conditions, quantifying performance degradation.
- Monitored output deviations across the full autonomous driving stack, including perception, planning, and control components. Modeled fault-induced behavior drift as a **Poisson proces**s, enabling statistical prediction of error propagation likelihood and failure onset under input perturbations.

Bayesian Optimization & Tensor Regression Aug 2019.8 - Apr 2021

- Proposed a novel **kernel function** for **Gaussian Processes** that preserves the structural information of tensorial spatial data, enabling effective learning in scalar-on-tensor regression tasks.
- Integrated **Bayesian Optimization** with the Gaussian Process model to efficiently explore the **design space** and identify high-performing parameter configurations.
- Applied the model in a **supervised learning** context to predict antenna performance based on 3D-printed geometric design tensors. Validated the result by comparing with the printed antenna performance.
- Demonstrated **8%** performance improvement on prediction accuracy and sample efficiency in antenna design tasks, leveraging the model's ability to exploit spatial correlations.

SPECIAL PROJECTS

Agentic RAG for Temporal Statute Validation & Correction

- **Engineered a "Temporal Verification" pipeline** using **LangGraph** to autonomously audit historical caselaw snippets. The system detects quantitative entities (monetary limits, statute of limitations) and validates them against a live "Current Statute" API to identify amended law.
- **Designed a "Lazy Verification" protocol** using **AsyncIO** that triggers external validation only when high-risk entities are detected. This reduced latency by bypassing the validation loop for non-statutory or qualitative text chunks.
- **Implemented an automated "Compare-and-Annotate" state** that calculates the delta between the cited (historical) figure and the current statutory limit. The system dynamically injects warning headers into the context window, preventing the LLM from hallucinating the validity of outdated legal precedents.

Fine-tuning the Multimodal BLIP Model with LoRA for VQA on Autonomous Driving Data

- Fine-tuned a **BLIP** multimodal model using **LoRA** to improve performance on Visual QA tasks in domain-specific autonomous driving scenarios.
- Designed and implemented an end-to-end pipeline, covering data preprocessing, model training, and evaluation by integrating Hugging Face Transformers, **PyTorch**, and custom data loaders.
- Improved model performance while minimizing memory and compute overhead through efficient parameter-efficient tuning with LoRA.

TECHNICAL HIGHLIGHTS

Programming Languages: Python, R, SQL, MATLAB, Git, CMake, ROS, SAS

LLM & RAG: GPT (OpenAI, Anthropic, etc.), Hugging Face Transformers & Datasets, LLaMA, FAISS, prompt engineering, LangChain, AutoGen, ReAct agents, LoRA fine-tuning

Deep Learning: PyTorch, TensorFlow, Keras, PySpark, Pandas, NumPy, SciPy, scikit-learn

Data Visualization: Matplotlib, Plotly, Seaborn, Power BI, Tableau, ggplot