陈文潇

chenwenxiaolive@gmail.com | https://github.com/chenwenxiaolive | +86 18801219115

Google Scholar | 1757+ 引用 | h-index: 10 | i10-index: 11

研究方向

我的研究聚焦于**大语言模型系统**和**智能运维（AIOps）**，致力于构建高效、可靠、智能的大规模AI系统基础设施。凭借扎实的算法功底（NOI 2012金牌、Codeforces Master），研究工作涵盖LLM系统的全栈领域，从推理优化到智能运维：

- **LLM推理优化：** 研发高性能推理加速技术，包括PD分离架构、基于RDMA的KVCache传输、分布式推理系统和内存高效服务策略。作为openFuyao社区架构师，推动DeepSeek等主流大模型在昇腾硬件上的适配部署。
- **智能运维（AIOps）：** 开创无监督异常检测算法（Donut、Buzz、Bagel），成为AIOps领域的基础性方法，支撑云原生系统的自动化故障诊断和根因分析。
- **深度生成模型：** 推进变分自编码器（VAE）及其在时序分析中的应用，代表性工作获得1,200+引用，达到异常检测领域的最优性能。

工作经历

技术专家（天才少年计划），华为技术有限公司 `2022.07 – 至今`

主导智能计算集群系统和LLM基础设施的研发工作。3年内获得15项公司级奖励，包括总裁个人奖-杰出工程师（2024）。

- **openFuyao AI推理基础设施：** 作为openFuyao开源社区架构师，主导昇腾硬件适配的软件栈规划。攻克PD分离架构、分布式KVCache传输、AscendCacheTier等关键技术难题，实现DeepSeek等主流大模型在昇腾910B上的云原生高效部署。
- **LLM推理加速：** 研发基于RDMA的KVCache高速传输技术，显著加速大语言模型推理。已申请4项专利，覆盖KVCache优化和LLM推理加速领域。
- **LogGPT - 运维领域日志大模型：** 打造业界首个免人工标注的运维领域日志分析大模型。在200+服务上实现80%+的异常日志识别准确率，将问题诊断时间从小时级缩短至分钟级。构建4大技术断裂点，涵盖18项算法技术。
- **智能北斗（ADN自运维）：** 主导自动驾驶网络运维平台核心算法研发。开发根因定界算法、拓扑聚类算法和服务地图能力。相比原有系统降低90%误报率，实现分钟级故障发现（原为天级）。因此工作获得总裁个人奖-杰出工程师。
- **北向智能体：** 从0到1打造AI智能体原型，结合API知识图谱与大模型能力。突破场景化API定义和智能API编排等关键技术。与广东移动落地Copilot自然语言接口，将工单处理时长从75分钟降至35分钟。客户评价："It's exactly what I want."
- **大网性能保障：** 研发流式日志采集和异构日志解析技术，处理能力达24GB/小时。实现微服务调用链构建，支撑30万+网络节点，商用后零质量事故。支撑中国电信政企网络2024年H1商用上线。
- **高性能图计算引擎：** 研发轻量级分布式图计算引擎，支持百亿级图存储，10跳遍历<10秒（比HugeGraph快20%）。支撑MAE-CN 20万虚机场景的跨层故障定界。

教育背景

清华大学 `2017.09 - 2022.06`

计算机科学与技术 博士

- 导师：裴丹教授
- 研究方向：智能运维、异常检测、深度生成模型

清华大学 `2013.09 - 2017.06`

计算机科学与技术 学士

- 优秀毕业生
- 学业优秀奖

开源贡献

openFuyao社区 - 架构师 `2024 - 至今`

作为openFuyao社区**架构师**，致力于构建和优化昇腾NPU的AI推理基础设施。通过对前沿技术的深入洞察（Dynamo、DeepSeek AI Infra等），主导昇腾硬件适配的软件栈规划，推动openFuyao AI推理项目的成功建立。

核心技术贡献：

- **PD分离架构：** 攻克昇腾910B上prefill和decode阶段分离的关键难题，实现更优的资源利用率和吞吐优化。
- **分布式KVCache传输：** 设计并实现高性能分布式KVCache传输机制，实现推理节点间的高效内存共享。
- **AscendCacheTier：** 研发AscendCacheTier分层缓存管理系统，显著提升昇腾910B硬件上的LLM推理性能。
- **Mooncake社区贡献：** 与蚂蚁集团合作，基于Mooncake构建下一代分布式KVCache架构（Tiered Caching）。主导设计Client端热点缓存优化，实现mooncake store get/batchget接口**60%-80%性能提升**。贡献被Mooncake社区认可并纳入其Roadmap。
- **DeepSeek模型部署：** 成功实现DeepSeek等主流大模型在昇腾硬件云原生环境的高效部署，打通前沿开源模型与国产AI基础设施的桥梁。

研究项目

基于RDMA的KVCache传输实现LLM推理加速 `2025.01 - 至今`

提出一种基于RDMA（远程直接内存访问）的高速KVCache传输方法，用于加速分布式推理节点间的大语言模型推理。该技术解决了LLM服务中的内存带宽瓶颈，实现高效的KVCache共享，减少多轮对话和长上下文场景中的冗余计算。

作为**技术负责人**，设计系统架构，实现RDMA通信协议，优化生产部署的内存管理策略。

LogAnalyzer：基于LLM的智算集群故障诊断 `2024.11 - 至今`

研发**LogAnalyzer**，一个面向大规模智算集群（万卡以上GPU/NPU）的基于LLM的智能故障诊断系统。系统解决关键痛点：**海量日志**使人工分析不可行，**复杂故障模式**使现有规则系统无法完全覆盖。

核心技术贡献：

- **异常日志识别：** 日志预解析、预标注和异常聚合，过滤噪声、降低数据量，同时保留关键故障信号。
- **异常信息提取：** 多模态提取各节点异常信息，生成LLM友好的结构化故障上下文。
- **故障传播链分析：** 结合作业级和进程级异常事件与领域知识（CANN、NPU、HCCL），构建可解释的故障传播链。
- **DeepResearch智能体工作流：** 诊断过程中的自适应知识检索，具备自纠错和反思能力，避免错误传播。
- **LLM驱动知识库构建：** 自动汇总内部文档、支持案例和排障博客，形成LLM可检索的故障知识库。

部署于**科大讯飞X1**集群，达成**87.5%整体诊断准确率**和**84.3%未知故障准确率**，将诊断时间从**天级缩短至分钟级。**与科大讯飞联合发布**LogAnalyzer**于**华为全联接大会2025**，开创**"以AI管AI"**运维理念。获《华为人》杂志报道，并在**华为全联接大会2025集群运维论坛**展示。

作为**项目负责人**，设计系统架构，主导算法研发，推动生产部署。

盘古基座模型类O1思维链能力增强 `2024.11 - 2025.05`

与盘古大语言模型团队合作，研究基于基座模型发展类O1的思维链推理能力。探索了包括受O1-Journey项目启发的过程奖励模型（PRM）方法和DeepSeek-R1的方法论，显著提升了盘古模型的推理能力。

核心技术贡献：

- **基于PRM的训练：** 实现了遵循O1-Journey的journey learning范式的过程奖励模型训练策略，使模型能够在复杂推理任务中从探索、回溯和自我纠正中学习。
- **DeepSeek-R1方法论：** 采用DeepSeek-R1的强化学习技术，增强长链推理和问题分解能力。
- **推理时计算扩展：** 利用推理阶段的延长思考时间提升模型在复杂推理任务上的性能，验证了测试时计算扩展的有效性。
- **自我纠错机制：** 开发了模型通过迭代优化和反思来检测和纠正自身推理错误的能力。

作为**算法负责人**，设计训练管道，实现核心算法，并在多个复杂推理基准上评估推理增强效果。

Chain-of-Event：微服务可解释根因分析 `2023 - 2024`

论文发表：FSE Companion 2024 | **10引用** | **eBay工业合作**

与eBay公司工业合作： 与eBay上海3位共同作者联合研究，在**eBay全球电商平台**（5,000+微服务、3个数据中心、服务**1.85亿活跃用户**、日均处理**10TB+监控数据**）的生产故障数据上进行评测。

提出Chain-of-Event（CoE），一个面向微服务系统的可解释根因分析模型。CoE将多模态监控数据（指标、日志、链路）整合为统一事件表示，自动学习加权事件因果图以捕获故障传播模式。在Service数据集上达成**79.3% Top-1准确率**和**98.8% Top-3准确率**，在Business数据集上达成**85.3% Top-1准确率**和**96.6% Top-3准确率**（真实生产故障数据：170个服务级故障 + 782个业务级故障）。

作为**核心研究者**，参与算法设计和系统实现。

Donut：基于VAE的无监督异常检测 `2017.09 - 2018.04`

论文发表：WWW 2018 | **1287+引用** | **阿里巴巴工业合作**

与阿里巴巴集团工业合作： 与阿里巴巴集团4位共同作者联合研究，在全球顶级互联网公司的生产KPI数据上进行验证。该算法已支撑**阿里云数据库自治服务（DAS）**，将数据库管理成本降低90%。

提出Donut，一种基于变分自编码器（VAE）的无监督异常检测算法，用于Web应用的季节性KPI。在真实生产数据上达成**F-score 0.75-0.9**，超越最先进的有监督方法。该方法已成为AIOps领域的基础性工作，被微软、阿里巴巴、百度等科技巨头广泛采用。

作为**共同一作**，设计模型架构，提出KDE新解释，发现潜在z空间的时间梯度效应。

Buzz：基于VAE对抗训练的异常检测 `2018.06 - 2019.04`

论文发表：IEEE INFOCOM 2019 | **122引用** | **阿里巴巴工业合作**

与阿里巴巴集团工业合作： 与阿里巴巴集团4位共同作者联合研究，在全球顶级互联网公司的11个生产复杂KPI（机器级指标，具有非高斯噪声模式）上进行验证。

提出Buzz，首个针对复杂KPI的基于深度生成模型的无监督异常检测算法。达成**F-score 0.92-0.99**，显著超越现有方法。理论创新贡献：首个基于分区分析的VAE对抗训练方法，连接贝叶斯网络与最优传输理论。

作为**第一作者**，设计对抗训练框架，提供理论证明，主导全部实验。

代表性论文

* 表示共同一作，# 表示指导学生

第一作者/共同一作论文

[WWW-2018] H Xu*, **W Chen***, N Zhao, Z Li, J Bu, Z Li, Y Liu, Y Zhao, D Pei, Y Feng, J Chen, Z Wang, H Qiao. **Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications.** Proceedings of the 2018 World Wide Web Conference. **[1287引用]**

[INFOCOM-2019] **W Chen**, H Xu, Z Li, D Pei, J Chen, H Qiao, Y Feng, Z Wang. **Unsupervised Anomaly Detection for Intricate KPIs via Adversarial Training of VAE.** IEEE Conference on Computer Communications. **[122引用]**

[IPCCC-2018] Z Li, **W Chen**, D Pei. **Robust and Unsupervised KPI Anomaly Detection Based on Conditional Variational Autoencoder.** IEEE 37th International Performance Computing and Communications Conference. **[114引用]**

[WWW-2023] Z Xie, H Xu, **W Chen**, W Li, H Jiang, L Su, H Wang, D Pei. **Unsupervised Anomaly Detection on Microservice Traces through Graph VAE.** Proceedings of the ACM Web Conference 2023. **[50引用]**

[JSAC-2022] Z Li, Y Zhao, Y Geng, Z Zhao, H Wang, **W Chen**, H Jiang, A Vaidya, L Su, D Pei. **Situation-Aware Multivariate Time Series Anomaly Detection through Active Learning and Contrast VAE-Based Models.** IEEE Journal on Selected Areas in Communications. **[29引用]**

其他代表性论文

[ICASSP-2016] X Li, H Xianyu, J Tian, **W Chen**, F Meng, M Xu, L Cai. **A Deep Bidirectional Long Short-Term Memory Based Multi-Scale Approach for Music Dynamic Emotion Prediction.** IEEE International Conference on Acoustics, Speech and Signal Processing. **[51引用]**

[ICASSP-2016] H Xianyu, X Li, **W Chen**, F Meng, J Tian, M Xu, L Cai. **SVR Based Double-Scale Regression for Dynamic Emotion Prediction in Music.** IEEE ICASSP. **[25引用]**

[arXiv-2019] H Xu, **W Chen**, J Lai, Z Li, Y Zhao, D Pei. **On the Necessity and Effectiveness of Learning the Prior of Variational Auto-Encoder.** arXiv preprint. **[22引用]**

[MediaEval-2015] M Xu, X Li, H Xianyu, J Tian, F Meng, **W Chen**. **Multi-Scale Approaches to the MediaEval 2015 "Emotion in Music" Task.** MediaEval. **[19引用]**

[IJCNN-2020] Z Li, Y Zhao, H Xu, **W Chen**, S Xu, Y Li, D Pei. **Unsupervised Clustering through Gaussian Mixture Variational AutoEncoder with Non-Reparameterized Variational Inference and Std Annealing.** International Joint Conference on Neural Networks. **[16引用]**

[FSE-2024] Z Yao, C Pei, **W Chen**, H Wang, L Su, H Jiang, Z Xie, X Nie, D Pei. **Chain-of-Event: Interpretable Root Cause Analysis for Microservices through Automatically Learning Weighted Event Causal Graph.** FSE Companion 2024. **[10引用]**

[ICONIP-2020] H Xu, **W Chen**, J Lai, Z Li, Y Zhao, D Pei. **Shallow VAEs with RealNVP Prior Can Perform as Well as Deep Hierarchical VAEs.** International Conference on Neural Information Processing. **[5引用]**

[ICONIP-2020] **W Chen**, W Liu, Z Cai, H Xu, D Pei. **VAEPP: Variational Autoencoder with a Pull-Back Prior.** International Conference on Neural Information Processing. **[3引用]**

[CCGrid-2024] Z Yu, Q Ouyang, C Pei, X Wang, **W Chen**, L Su, H Jiang, X Wang, J Li, D Pei. **Causality Enhanced Graph Representation Learning for Alert-Based Root Cause Analysis.** IEEE International Symposium on Cluster, Cloud and Internet Computing. **[2引用]**

[arXiv-2021] **W Chen**, X Nie, M Li, D Pei. **DOI: Divergence-Based Out-of-Distribution Indicators via Deep Generative Models.** arXiv preprint. **[2引用]**

专利

专利（4项已申请）

- **一种基于Client端热点缓存的推理服务加速方法** - 利用客户端热点缓存加速LLM推理服务。
- **一种基于KVCache Endpoint批量聚合的推理集群优化方法** - 基于KVCache端点批量聚合的LLM推理集群优化方法。
- **基于KVCache智能混合传输策略的分布式集群推理计算加速方法** - 基于智能混合KVCache传输策略的分布式LLM推理加速方法。
- **一种基于用户集成需求表达的API自动生成方法** - 基于自然语言需求表达的AI驱动API自动生成方法。

科研项目

- **国家自然科学基金（No. 62072264）：** 基于多模态监控数据的微服务无监督异常检测与定位（2020，负责人：裴丹教授）
- **国家自然科学基金（No. 62202445）：** 基于大规模预训练的人机协同微服务异常根因定位
- **国家自然科学基金（No. 61472214）：** 基于大数据分析的互联网服务性能管理架构研究
- **国家自然科学基金（No. 61472210）：** 基于用户需求的自主无线网络协作模型与优化机制研究
- **大川研究基金（2017）：** 基于深度学习的异常检测（负责人：清华大学裴丹教授）
- **阿里巴巴创新研究计划（AIR）**

荣誉奖项

奖项	年份
华为杰出专家	2025
openFuyao社区架构师	2024
华为总裁个人奖-杰出工程师	2024
华为ICT AI算法大赛奖（基于图的故障根因识别算法）	2023
华为测试技术创新突破奖	2024
华为竞争力突出贡献奖	2024
华为总裁奖（3年15项，含4项个人奖）	2023-2025
华为天才少年计划	2022
华为极客大赛二等奖（约10万人中排名第13）	2025
NOI金牌（全国青少年信息学奥林匹克竞赛）	2012
Codeforces Master（Rating 2299，全球前0.3%）	-
清华大学优秀毕业生	2017
国家奖学金	2014
清华大学人人贷奖学金	2014
清华大学学业优秀奖	2014
搜狗杯第18届智能体大赛优秀奖	2014
省数学奥林匹克竞赛银牌	2011

学术服务

期刊审稿人（完成17+次审稿）：

- IEEE Access

- IEEE Transactions on Intelligent Transportation Systems (T-ITS)
- Neurocomputing (Elsevier)
- International Journal of Networked and Distributed Computing (IJNDC)
- MDPI Algorithms
- MDPI Network
- MDPI AI

竞赛评委：

- 2019 AIOps挑战赛决赛评委（CCF、清华、华为、苏宁联合主办）

受邀报告

- **AiDD 2025 - 第8届AI+研发数字峰会（深圳，2025年11月）：** 受邀演讲 - "智算集群故障诊断算法研究与实践"
- **第4届运筹学与人工智能在业界前沿应用研讨会2025（北大深研院，2025年12月）：** 受邀演讲
- **IEEE INFOCOM 2019（法国巴黎）：** 论文报告
- **WWW 2018（法国里昂）：** 论文报告
- **WWW 2023（美国奥斯汀）：** 论文报告
- **ICONIP 2020：** 论文报告
- **IJCNN 2020：** 论文报告