小明

(+86) 138-0000-0000 · xxx@gmail.com · **推理优化架构工程师** · GitHub @username

个人总结

具备扎实的自然语言处理与深度学习背景，熟练掌握 PyTorch 及相关推理框架，拥有大规模模型部署与优化经验。在多机并联和模型蒸馏方面具备深入理解，致力于推动模型推理性能的提升。

教育背景

清华大学, 计算机科学与技术, *硕士* 2024.09 - 2027.06

- 研究方向聚焦于深度学习与自然语言处理，相关课程包括模型优化与并行计算。

北京理工大学, 信息与计算科学, *学士* 2020.09 - 2024.06

- 学习了 C++ 编程和 CUDA 编程基础，参与课程项目进行并行计算实践。

技术能力

- **编程语言**: Python, C++, CUDA, JavaScript, SQL
- **框架/工具**: PyTorch, TensorFlow, Triton, DeepSpeed, Docker
- **专业技能**: 深度学习, 自然语言处理, 模型蒸馏, 高并发推理优化

工作经历

公司名称 | 部门/团队, 推理优化工程师 YYYY.MM - YYYY.MM

- **项目名称:** 高并发推理框架优化，针对实际业务场景进行了性能调优，提升了系统并发处理能力。
- 负责 C++ 与 CUDA 相关模块开发，实现了推理速度提升 25% 的目标。
- 优化推理流程，减少了模型加载时间，提升了系统整体响应速度。

项目经历

大规模中文情感分析系统, 负责人 2024.10 - 2025.03

- **技术栈:** Python, PyTorch, Triton, vLLM
- 构建基于 RoBERTa-wwm 的情感分类模型，准确率达到 93.2%，比 baseline 提升 7.8%。
- 部署模型于 Triton 环境，实现高并发推理，延迟低于 50ms，支持 5000+ 用户并发访问。
- 进行模型蒸馏，参数压缩 15.3×，推理速度提升 12×，显著降低了资源消耗。

智能客服意图识别与多轮对话系统, 核心成员 2025.04 - 2025.09

- **技术栈:** Python, LLaMA-3, DeepSpeed, Hydra
- 设计意图识别与动作执行统一框架，提升知识问答准确率，系统已服务 50,000+ 用户/日。
- 使用 DeepSpeed 优化显存与通信性能，确保系统在多轮对话中表现稳定。
- 多轮会话成功率从 72.4% 提升至 92.1%，意图识别 F1 提升 13.5%。

竞赛获奖

- 全国人工智能创新竞赛**一等奖** (*官网链接*),2024 年 08 月
- 个人博客/技术文章: *https://blog.example.com*

社区参与/其他

- 参与开源社区贡献, **为多个项目提交 12 个 PR**
- 技术博客累计访问 5 万次，发表技术文章 10 篇
- 参加多次技术分享与交流活动，提升团队协作与跨部门沟通能力。