个人信息

姓名：张伟

电话：(+86) 138-1234-5678

邮箱：zhangwei.llm@email.com

GitHub：github.com/zhangwei-ai（附LLM相关项目链接）

LinkedIn：linkedin.com/in/zhsangwei-llm

现居地：北京 | 可远程/异地办公

技术博客：zhangwei.ai（可选，如有技术文章可增加）

教育背景

清华大学 | 工学硕士 | 计算机科学与技术（人工智能方向）

2020.09 - 2023.06

主修课程：自然语言处理、分布式机器学习、大模型架构设计

毕业论文：《基于LLaMA-2的医疗问答系统低资源微调方法研究》（TOP 10%论文）

北京大学 | 工学学士 | 软件工程

2016.09 - 2020.06

工作经历

字节跳动 | 大模型应用工程师（高级）

2023.07 - 至今

模型优化：主导公司内部大模型（参数量10B+）的微调与推理加速，通过LoRA/P-Tuningv2降低显存占用35%，推理速度提升40%；

落地应用：开发电商场景多模态大模型应用（文本+图像生成），支持日均100万次调用，错误率低于0.5%；

技术攻关：设计动态Prompt框架，结合用户行为数据提升推荐系统CTR 15%；

阿里巴巴 | 算法工程师（实习）

2022.03 - 2022.12

参与千问大模型早期开发，负责数据清洗与SFT阶段实验，显著降低有害输出概率；

搭建基于BERT的客服意图分类系统，准确率提升至92%。

项目经验

项目名称：金融领域大模型智能投顾系统

技术栈：LLaMA-2-13B、LangChain、Faiss向量数据库、FastAPI

成果：支持实时金融政策解读与投资建议生成，客户满意度提升25%，获公司创新奖。

项目名称：开源中文大模型微调工具链

技术栈：PyTorch、Deepspeed、vLLM

成果：GitHub Star 500+，被3家初创企业采用。

技能专长

大模型技术：全流程开发（预训练/微调/部署）、推理优化（量化/FlashAttention）、Agent开发

编程语言：Python（熟练使用asyncio高性能框架）、C++（模型底层优化）

工具链：HuggingFace Transformers、MLflow、Docker/K8s、Prometheus监控

语言能力：英语（CET-6，可熟练阅读AI论文）

证书与荣誉

Kaggle LLM竞赛银牌（2023）

国家奖学金（2021）

发明专利：《一种大模型安全对齐方法》（公开号CNXXXXXX）

自我评价

深耕大模型技术栈，熟悉行业最新进展（如Mixtral、Gemma等），长期贡献开源社区；

兼具算法研究与工程落地能力，擅长将论文成果转化为实际业务价值；

热爱技术分享，主导团队内部LLM技术培训10+场。