Hui Fang

Work Experience：5 | Location: Chicago | Phone: 312-285-6838 | Email: fanghui954@gmail.com

Personal Profile

有多模态LLM / Agents/ SFT/ RLHF/ MOE/ RWKV GPT /YOLO 等LLM相关前沿技术、计算机视觉、自然语言处理经验，跨模态理解和生 成学习经验，在LLM等领域具有多年以上的行业经验，如图像分类与识别、目标检测、AI生成模型框架，

Education Background

Illinois Institute of Technology	Computer Science and Technology	**Master's Degree**	**May 2023 - Dec 2024**
Southwest Jiaotong University	**Computer Science and Technology**	**Bachelor's Degree**	**Sep 2013 - May2017**

Core Skills

➢熟练使用TensorFlow、PyTorch、Keras、Scikit-learn, RNN 熟悉 **Python**， **C++** 等编程语言， **PyTorch**，具备构建、训练和优化 NLP 模型的能力。

➢ 运用SFT、RLHF、大规模模型微调，整合优化多模态AI模型（MoE混合专家模型架构、Transfermer，RAG、RWKV,GPT,Deepseek），处理图像与文本的交互，有大规模 **LLM** 和 **多模态模型** 的设计、训练、调优与部署经验

➢ 分布式训练：DDP，GPU通信，混合精度推理优化，KVCache优化，K8s GPU调度，混合并行策略：混合3D并行（TP+PP+DP）组合优化

➢ 熟悉 Transformer 等大语言模型，具备 **对话生成、语义理解、文本分类，、训练与优化**，特别是在 **多模态对话管理、语义理解、文本生成** 和 **用户行为预测**

Work Experience

Character.ai (Internship) | ChatBot | AI Engineer | Location: California | Mar 2024 - September 2024

➢ 设计并实现了多模态对话管理系统，支持 **文本、语音** 和 **图像** 数据的输入，提升了机器人与用户之间的互动体验，确保chatbot能够准确理解用户意图并生成合理响应，支持多轮对话和上下文理解，使用RAG深度搜索提供专业知识查询，意图识别和 情感分析 能力，快速准确地响应用户提问

➢ KVCache技术经过设计和优化，通过缓存历史键值对数据，显著加快模型推理速度，降低对响应时间，减少计算资源浪费。通过推测抽样技术，优化了生成任务的推理过程，显著提高了多轮对话的响应速度。

➢ 基于 Transformer+RLHF 模型可以结合图像和文本数据进行分析，进行图像-文本交互-RAG深度搜索，**PyTorch**，使用深度学习框架进行语音助手相关模型的训练与优化，提升了系统的响应速度和计算效率

➢ 模型训练-优化：使用RAG，评估（BLEU/ROUGE），多模态LLM修剪，蒸馏，LoRA和端到端数据清理等技术优化模型训练的专业知识，分布式训练：PyTorch（DDP）和评估

➢ 根据用户的意图，语音 或文本 指令执行相应的动作，如查询天气、播放音乐、管理（翻译，语音交互语义分析，多轮对话，情感分析和目标搜索：机器人与用户之间），确保对话的逻辑性和连贯性

Achievements: 显著提高聊天机器人性能，响应准确率提高20%，，客户响应时间减少了 30%，客户满意度提高了 20%

General Motors : | Intelligent Drive System | AI Engineer | Location: Austin | June 2019 - January 2023

- 设计并实现了多模态输入模型，支持语音、文本等多种输入模式的联合处理，提高了语音助手的多场景适应性，通过语音增强和噪声抑制算法优化了嘈杂环境下的语音识别，ASR准确率提高了30%
- 创新应用MoE混合专家模型架构，MoE可以在不同任务之间动态选择专家模型，优化不同任务的执行效率，结合fp16定量技术优化推理过程，系统响应速度提升60%。
- 模型训练-优化：评估（BLEU/ROUGE）、多模态LLM修剪、蒸馏、LoRA和端到端数据清理等技术优化模型训练的专业知识,PyTorch(DDP），以及评估。
- **深度强化学习** 和 **行为决策树** 等方法，优化了自动驾驶车辆的行驶路径选择，确保了高效的行驶路线和安全的驾驶决策。混合精度定量技术优化推理过程，通过实时数据分析，不断优化模型
- 优化的YOLO用于实时目标检测 NLP- LLM技术用于解释道路标志和交通信息，利用激光雷达深度数据检测目标，实时处理和检测行人、车辆、障碍物, LIDAR、摄像头 等传感器数据的融合技术，提升自动驾驶系统的环境感知能力

Achievements: 开发可扩展的语音交互系统，客户满意度95%，日均交互10000 +，通过先进的预处理，ASR效率提升25%。

科大讯飞 | Smart Service Robots | AI Engineer | Location: Shanghai | June 2017 - May 2019

- 通过使用语音识别api和自定义代码集成实现Google语音到文本的转换。利用BERT和GPT模型增强用户意图识别。执行文本分类情感分析和实体识别等任务，以提高对用户请求的理解
- 集成的情绪识别功能使聊天机器人能够根据用户的语音情绪调整语调和回复内容，**深度强化学习（RL）**，开发了智能机器人自主导航系统，能够自动避障并根据环境情况选择最佳路径
- 集成VisualBERT通过使用多模态编码器和注意机制将图像数据与自然语言命令相结合。通过使用深度学习模型和自然语言处理技术，使机器人能够理解和响应视觉和语音命令

Achievements: 成功提高了智能服务机器人的性能和用户满意度。语音识别准确率达到95%，任务完成率提高20%。机器人自主导航效率提高30%