Nicholas Wall

https://www.walln.dev

Email : walln@hey.com
Mobile : +1-214-425-6155

Education

Southern Methodist University — Dallas, TX
- *Master of Science in Computer Science - Lyle Discovery Scholar* — *January 2023 - December 2023*
  Machine Learning and Artificial Intelligence Specialization

Southern Methodist University — Dallas, TX
- *Bachelor of Arts in Computer Science - SMU Distinguished Scholar* — *August 2019 - December 2022*
  Machine Learning and Artificial Intelligence Specialization

Experience

- **Maple AI** — New York, NY
  Engineering & AI Lead — *January 2025 - Present*
  - **Team Leadership & Scaling**: Tripled engineering team while establishing development processes and culture that enhanced team cohesion and productivity.
  - **Full-Stack Architecture & Deployment**: Architected full-stack systems and scaled deployment infrastructure to support production workloads, ensuring high availability and saving thousands of dollars in compute costs.
  - **Online Evaluation Systems**: Built online agentic evaluation framework to continuously assess agent performance in production environments, enabling data-driven improvements.
  - **AI Research Leadership**: Spearheaded research on novel embedding model distillation and compression techniques, achieving 100x reduction in latency while maintaining equivalent performance.
- **Maple AI** — New York, NY
  Member of Technical Staff — *August 2024 - January 2025*
  - **LLM Fine-tuning**: Fine-tuned large language models that improved agent reliability by 20% on internal evaluations while enhancing conversational style and coherence.
  - **Prompt Optimization with RL**: Implemented reinforcement learning techniques for prompt optimization, achieving 5% improvement in instruction adherence by utilizing tailored in-context learning examples over generalized approaches.
  - **Infrastructure Optimization**: Led initiative to redesign orchestration stack, achieving over 70% reduction in latency overhead through architectural optimizations.
- **Independent AI Researcher** — Remote
  Research Engineer — *January 2024 - August 2024*
  - **Mixture of Experts Research & Interpretability**: Trained classifiers on MoE gate logits to probe for emergent concepts and understand internal model representations. Experimented with pruning sparsely activated experts to improve model efficiency while maintaining performance.
  - **Synthetic Data Generation**: Developed and scaled tooling for synthetic data generation, enabling large-scale dataset creation for specialized AI training tasks.
- **IBM** — Dallas, TX
  Intern — *May 2021 - August 2022*
  - **Applied ML Research & Production Systems**: Productionized state-of-the-art NLP models. Built generative chat application serving thousands of users with computer vision integration, retrieval augmented generation, and mobile frontend. Reduced monitoring latency by 10x through real-time React dashboard.
  - **Model Training & Optimization**: Trained and deployed BERT-based models for NLP tasks (NER, co-reference resolution) and Speech-to-Text models (wav2vec) optimized for low-quality phone calls, achieving word error rates under 15. Improved ASR performance by 10% through dataset curation and annotation.
  - **Knowledge Graph Systems**: Designed Python framework for directed cyclic graphs to convert unstructured data into knowledge bases. Developed React visualization tool for exploring knowledge graphs with thousands of nodes, deployed and managed Kubernetes clusters.

Technical Skills

- **Languages**: Python, TypeScript, C++, SQL, Rust • **AI/ML**: PyTorch, JAX, Transformers, Hugging Face, vLLM, W&B • **Infrastructure**: AWS, Kubernetes, Docker, Terraform, SST, Ray • **Fullstack**: Next.js, React, PostgreSQL, Kafka, FastAPI, Serverless, Cloudflare