

KEVIN JONES

📞 828-508-1414

✉ jone2342@purdue.edu

🌐 [linkedin.com/in/kevin-jones-26y](https://www.linkedin.com/in/kevin-jones-26y)

🐙 github.com/kevinjones557

EDUCATION

Purdue University

West Lafayette, Indiana

M.S. in Computer Science (GPA: 4.00)

May 2026

- **Coursework:** Natural Language Processing, Machine Learning Systems, Statistical Machine Learning, Reinforcement Learning, Computer Vision with Deep Learning, Graduate Algorithms, Parallel Computing, Distributed Systems

B.S. in Computer Science (GPA: 3.99) | Concentration: Machine Intelligence

May 2025

- **Coursework:** Data Mining and Machine Learning, Data Structures and Algorithms, Linear Algebra, Probability

EXPERIENCE

Kensho Technologies

Summer 2025

Machine Learning Engineering Intern

Cambridge, Massachusetts

- Designed **novel table ingestion strategies for RAG Pipeline**, orchestrating experiments via Airflow and driving significant performance gains (**200% higher MRR and 300% better accuracy**). Prepared best-performing strategy for **production**.
- Drove **end-to-end ML experimentation process** from ideation to evaluation, analysis, and deployment. Built a **21,000+ LLM-generated question-table dataset** tailored for table-specific retrieval benchmarking in financial RAG systems.
- Researched properties of embedding models including the effect of explicit prompting, the impact of text placement, and robustness and effectiveness of various templates to optimize effectiveness of text embeddings.

Southwest Power Pool

Summer 2024

Software Engineering Intern

Little Rock, Arkansas

- Built a complete Java Swing app to generate valid/invalid test data, improving testing efficiency across teams.
- Developed tests for 10+ projects using JUnit/Mockito, addressed security issues, and automated static scans for 50+ repos.

Infoscitex Corporation

Summer 2023

Operations Analyst Intern

Dayton, Ohio

- Created a multi-process Python data pipeline with Pandas and Matplotlib visualization, featuring a Tkinter GUI and Numba optimization. Extended a C++ AFSIM plugin to simplify scenario setup for clients.

RESEARCH & OPEN-SOURCE PROJECTS

Baseball Large Language Model | Python, Pytorch (<https://github.com/kevinjones557/Baseball-LLM>) Summer 2024 – Present

- Designed and **trained a 124M-parameter Transformer LLM from scratch** in PyTorch, including a ground up implementation of multi-headed self-attention, without using PyTorch's built-in Transformer module.
- Pretrained on **2B tokens and finetuned for Q&A on 200M tokens**, then domain-adapted using 1,000 baseball Q&A pairs. Baseball finetuning **boosted token prediction accuracy by 50%** on baseball queries vs. the general model.
- Increased training efficiency with FlashAttention (2× speedup) and gradient accumulation to overcome GPU constraint and experimented with **custom BPE tokenizer** to reduce vocabulary size; replaced with GPT-2 tokenizer for scalability.

LLaMA Model Distillation | Python, Pytorch, LLaMA, Distillation (<https://github.com/kevinjones557/Distillation>) Spring 2025

- **Distilled LLaMA 3.2 1B into a 749M-parameter student model** by removing alternate layers and training the remaining pre-initialized layers utilizing a hybrid soft and hard loss objective, encompassing a total of 2.2B tokens.
- Evaluated model performance across six NLP tasks using log-likelihood-based generative reformulations for multiple choice tasks. **Achieved 82% knowledge retention** with the distilled model outperforming the teacher on AGIEval by 1.5%.

Diffusion Model | Python, Pytorch, DDPM, Quantization (<https://github.com/kevinjones557/Diffusion-Model>) Fall 2024

- **Implemented diffusion models (DDPMs) from scratch in PyTorch**, including the full forward/reverse diffusion process.
- Trained six DDPMs on CIFAR-10 across 500 epochs each, totaling 3 million image updates per mode, varying the number of time steps and precision (using gradient scaling for float16). Evaluated models using FID and Inception Score.

Time Capsule | React Native, Node.js, NumPy, ONNX, MongoDB (https://github.com/nathan4sch/Time_Capsule) Spring 2024

- Led design and deployment of a custom ML model for a social app. Collected 4.5K+ samples, engineered feature vectors using Google Vision, trained a neural net in PyTorch achieving 91% classification accuracy, and deployed via ONNX.
- Collaborated with team of 4 to build full-stack app using React Native, Node.js, MongoDB, S3, Google Auth and Spotify APIs

TECHNICAL SKILLS

ML/AI: Transformers, Diffusion Models, Knowledge Distillation, Retrieval Augmented Generation, Embeddings, Mixed-Precision Training, LLM Finetuning, Mixture of Experts, Multi-Head (Latent) Attention

Languages & Libraries: PyTorch, Hugging Face, NumPy, Tiktoken, Airflow; Python, Java, C++, C, SQL, Bash, x86 Assembly