Shrey Pandey

https://www.linkedin.com/in/shrey2076
https://shreypandey.github.io/

Email : shreypandey1509@gmail.com
Mobile : +91-905-726-1430

Experience

- **Microsoft Research** — Bangalore, India
  Research Fellow — *Jan 2024 - Present*
  - **CoPilot Agents**:
    - Leading the development of LLM-powered agents aimed at eliminating technical debt in large-scale codebases by automating software development tasks.
    - Designing and implementing a dynamic AgentBuilder framework that generates agents based on customizable recipes, facilitating scalable automation of routine and complex coding operations.
  - **PwR - Programming with Representations**:
    - Developed a system of conversational programming where natural language conversations are translated into a custom domain-specific language with built-in guardrails using LLMs, allowing users to build software easily.
    - The DSL is then translated into Python code to run on JugalBandi-Manager, enabling seamless execution of user instructions.
    - Check PwR-Studio here and PwR-NL2DSL here.
  - **Jugalbandi Manager (JBManager)**:
    - Leading the development and maintainence of the open source project - JB Manager, which aims to release and manage conversation bots built using PwR for non-profit usecases.
    - JB Manager is responsible for providing multi tenant support, language and interaction medium agnostic conversation flow chatbots using finite state machines.
    - Check JugalBandi Manager here.
  - **Multi-Turn Conversational Information Retrieval**:
    - Developed a conversational system to streamline the form-filling process using LLMs resulting in reducing inefficiency and user experience.
    - The system dynamically guides users through follow-up questions, automatically selecting and filling the correct form from a pool of 15,000 options, significantly reducing manual effort.

- **Myntra (Walmart Global Tech)** — Bangalore, India
  Senior Software Engineer, Data Science Engineering — *Apr 2023 - Dec 2023*
  - **NNFetch - In-house Vector Database**
    - Created a dynamic recommendation platform, enhancing product discovery by blending metadata filters and partitioning with HNSW and ANNOY indexing for lightning-fast nearest product retrieval.
    - Empowered multiple data science projects with a pivotal recall set generator, adapting effortlessly to evolving metadata and delivering remarkable **throughput of 1M RPM** at a peak **P99 latency of 20ms**.
    - Delivered adaptive clustering logic, optimizing recall score and latency by extending existing indexes, enabling precise adaptation to diverse business use cases with rapidly changing metadata.

- **Myntra (Walmart Global Tech)** — Bangalore, India
  Software Engineer, Data Science Engineering — *Jul 2021 - Mar 2023*
  - **Personalised Search Re-Ranking**
    - Designed and developed a real-time, user-personalised search-ranking service, using long-term and short-term user embeddings, query relevance, product attributes in an XGBoost model optimised for CTR.
    - Engineered a high-performance solution for millions of users and products, leveraging Aerospike data store, SIMD-based cosine similarity processing, and ONNX Runtime for XGBoost model inferencing, delivering seamless operation at **500K RPM** with a remarkable **P99 latency of 30ms**.
    - Proven positive impact on business metrics through AB tests, enhancing user experience and relevance at scale.
  - **Recommendations**

        - Engineered a versatile framework for real-time product recommendations, leveraging embedding-based features and MCDA to optimize for relevance, personalization, and real-time business metrics, and using Determinantal Point Processes (DPP) to incorporate diversity in the recommendations.
        - Scaled the system to handle throughput of **5M RPM** with a **P99 latency of 40ms**, achieving a **19x latency reduction** and an **88% annual hardware cost savings** compared to the previous system.
    - **Machine Learning Scaling**: Built reusable components for scaling data science models by optimizing for latency and throughput. Maintaining compatibility by python, aligning with preferences of data scientists and easy to use with minimal code changes. Proposed solutions have been adopted to improve multiple data science services at Myntra.
        - **Effective Vector Store**: Engineered a memory-efficient solution, consolidating embedding vectors in shared memory to **reduce latency by 4.5x** and **reduce memory footprint by 6x**. Enhanced application throughput per node by enabling more worker processes per node.
        - **Cosine Similarity and DPP**: Introduced a high-performance approach using Single Instruction Multiple Data (SIMD) for calculating similarity scores and re-ranking, drastically cutting computation cycles and latency. The method handles up to **15x more traffic** while improving latency.
    - **Image Pose Detector**: Created an optimised online service for pose detection from an image using a ResNet-18 model with custom layers. Able to handle **27x more throughput** with a **6x reduction in latency** from its predecessor on a single node.

- **Myntra (Walmart Global Tech)** Bangalore, India
  Software Engineer Intern *Summer 2020*
    - **Image Generation**: Designed and implemented a two-step framework for topwear product swapping in images. Developed an encoder-decoder-based generative model for preliminary image generation, followed by an image refinement process to incorporate garment textures, ensuring accurate top-wear product replacement while retaining pose and body shape.

Education

- **Motilal Nehru National Institute of Technology** Allahabad, India
  Bachelor of Technology in Computer Science and Engineering; CGPA: 8.43/10.0 *July. 2017 – June. 2021*

Publications

- **Shrey Pandey**, Saikat Kumar Das, Hrishikesh V. Ganu, and Satyajeet Singh. 2024. Rethinking 'Complement' Recommendations at Scale with SIMD. In Proceedings of the 15th ACM/SPEC International Conference on Performance Engineering (ICPE '24). Association for Computing Machinery, New York, NY, USA, 25–36.
- Patel, Dhruv, **Shrey Pandey**, and Abhishek Sharma. "Efficient Vector Store System for Python using Shared Memory." In Proceedings of the Second International Conference on AI-ML Systems, pp. 1-6. 2022.
- **Pandey, Shrey**, Yash Srivastava, Yukta Meena, and Rupesh Kumar Dewang."CLOTON: A GAN based approach for Clothing Try-On." In 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 595-601. IEEE, 2021.

Achievements

- Awarded **Best Long Paper Award** at Convergence 2023 (Flipkart Internal Conference) for the presentation on 'Rethinking complement product recommendations at scale using SIMD'.
- Presented the work on 'Large Scale Recommender Systems in Fashion E-commerce' at BAICONF 2022 hosted by DCAL@IIM-B.
- Awarded **Employee of the Year 2022** at Myntra for delivering substantial cost savings through optimization efforts in scaling machine learning components and implementing the same in recommendation services.

Skills

- **Research Interest**: LLM, RAG, Recommendation Systems, Information Retrieval, Transformers
- **Languages**: Python, GoLang, Java, C++
- **Technologies**: Numba, Numpy, Flask, Gunicorn, Aerospike, PyTorch, ONNX, CUDA, NVIDIA Triton