Shubham Gupta

(717) 590-0924 | skg5962@psu.edu | linkedin.com/in/shubhgupta7049 | github.com/Shubh3005 | shubham.us

Summary

Computer Science senior (Schreyer Honors College, GPA 3.83) with end-to-end ML deployment experience and a published paper at AIES-26. Fine-tuned and deployed EfficientNet-B3 to production (Railway + HuggingFace Hub), built LLM-based annotation pipelines on HPC clusters, and engineered RAG systems with FAISS.

Education

The Pennsylvania State University — Schreyer Honors College University Park, PA
B.S. in Computer Science, Minor in Artificial Intelligence Engineering; GPA: 3.83/4.00 *Expected Dec. 2026*
- **Relevant Coursework**: Machine Learning, Neural Networks, Database Systems, Operating Systems, Data Structures
- **Honors Thesis**: LLM-based codebook maintenance — definition-aware synthetic QA generation, RAG pipelines, and LLM evaluation metrics (Advisor: Prof. Qunhua Li)

Publications

Do Personalized Evaluation Functions Reflect Human Preferences? | *AIES-26* 2026
- A Study of Weighted Proximity in Algorithmic Recourse. Choudhury, Leicht, Bislig, Gupta, Rekha, **Shubham Gupta**, Anand, Guo, Yadav. AAAI/ACM Conference on AI, Ethics, and Society, 2026.

Experience

AI Innovation Scholar June 2026 – Aug. 2026
Deloitte *Philadelphia, PA*
- Building production AI systems and enterprise LLM applications; researching applied large language model deployment in high-stakes enterprise contexts.

Undergraduate Research Assistant (NLP / LLMs) Jan. 2026 – Present
Penn State Department of Statistics — Prof. Qunhua Li *State College, PA*
- Built **RAG pipeline** with **FAISS** vector indexing and **LLaMA-3** inference for political science codebook annotation — novel over Han et al. baseline with **definition-aware synthetic QA generation**.
- Designed **definition compliance metric** for LLM evaluation — quantifying gap between model accuracy and adherence to structured definitions at scale on **ICDS Roar HPC cluster**.
- Collaborating on NLP workshop/journal submission; thesis work targets EMNLP or ACL venue.

Undergraduate Research Assistant (Explainable AI) Oct. 2024 – Jan. 2026
Penn State Yadav Lab *State College, PA*
- Engineered ML evaluation pipelines in **PyTorch** and **scikit-learn** for counterfactual explanation research — improving preference prediction to **84%** over CFE baselines.
- Contributed to **AIES-26** publication on personalized evaluation functions and weighted proximity in algorithmic recourse.

IT Summer Associate June 2025 – Aug. 2025
University of Pittsburgh Medical Center (UPMC) *Pittsburgh, PA*
- Engineered **Nursing Matrix** application integrating Epic EHR data across four hospital units serving 200+ staff; **Redis** caching reduced API latency by **18%**.

Technical Projects

SkinIQ | *EfficientNet-B3, PyTorch, FastAPI, HuggingFace, Docker, Supabase, React* 2026
- Fine-tuned **EfficientNet-B3** on HAM10000 (10k images, 7 classes) with **transfer learning**, weighted cross-entropy loss, and **WeightedRandomSampler** for 58:1 class imbalance — **63% val accuracy, macro F1 0.66**.
- Built **MLOps** pipeline: model weights versioned on **HuggingFace Hub**, **Dockerized** FastAPI inference server auto-downloads weights on startup and serves predictions at <2s latency.
- Full-stack deployment: Railway (backend) + Vercel (frontend) + Supabase (auth/DB/RLS). Live: skin-iq.vercel.app.

Mosaic-Ward (TartanHacks) | *Python, MediaPipe, FastAPI, WebSockets, Gemini API* Feb. 2026
- Deployed **computer vision** inference pipeline at the edge using **MediaPipe** pose estimation — real-time skeletal tracking at 30fps with <100ms WebSocket streaming latency.
- Integrated **Gemini API** for multimodal clinical report generation from fall physics data (velocity, impact angle) — reducing incident documentation time by an estimated 20 minutes per event.

Distributed Storage System (JBOD) | *C, Pthreads, Sockets* Sept.–Dec. 2024
- Implemented high-throughput storage backend with mutex synchronization and write-ahead logging — **10**× throughput improvement, 1,000+ concurrent connections.

Leadership

Founding Member / Mechanical Lead, Robot in 3 Days (Ri3D) Aug. 2023 – Present
- Led rapid prototyping under 72-hour hard constraints — system architecture decisions, component integration, and cross-functional team coordination.

Mentor, FIRST Global Teams (Belize, Niger, Rwanda, Bolivia) June 2021 – Present
- Mentored 20+ students across four countries in programming, robotics systems, and engineering problem-solving for international competition.

Technical Skills

Languages: Python, C++, C, JavaScript/TypeScript, SQL, Java
AI/ML: PyTorch, scikit-learn, EfficientNet, Transfer Learning, RAG, FAISS, HuggingFace, Gemini API, MediaPipe
Infra/Web: Docker, FastAPI, Railway, Vercel, Supabase, Redis, Git, Linux, SLURM/HPC
Concepts: Model Deployment, LLM Evaluation, Distributed Systems, Algorithmic Recourse