SAHITH BODLA

bodlasahith@gmail.com | sbodla2@illinois.edu | 669-251-8325 |
https://www.linkedin.com/in/sahith-bodla | https://github.com/bodlasahith | https://bodlasahith.github.io/personal-website/

EDUCATION

University of Illinois at Urbana-Champaign *August 2022 – May 2026*
- BS-MCS in Computer Science with Minor in Economics
- Relevant Courses: Data Structures, Computer Architecture, Applied Machine Learning, System Programming, Algorithms, Database Systems, Adv. Distributed Systems, Wireless Networks, App. Parallel Programming, Computer Security I, Programming Languages & Compilers, Game Development

WORK EXPERIENCE

Amazon, *Software Development Engineer Intern* *May 2025 – August 2025*
- Developed on internal web tool for Buy with Prime's on-call engineers, enabling efficient root cause analysis of system failures and operational alerts
- Integrated 5 internal APIs to aggregate real-time order/return event data and audit trails from AWS services including CloudWatch and EC2
- Used a Vue.js-based frontend and Lambda backend (Java/CDK with TypeScript), streamlining cross-service timeline visualization
- Improved incident resolution speed by providing comprehensive insights into fulfillment, merchant, and consumer interactions within Buy with Prime
- Designed and prototyped a GenAI-powered timeline summarizer using Amazon's proprietary MCP scripting, highlighting anomalies and critical severities

Gen Digital, *Software Engineer Intern* *June 2024 – August 2024*
- Developed and deployed an AI-driven scam detection email scanner, leveraging Norton's proprietary NLP model to analyze phishing patterns to achieve an 80% accuracy rate in identifying fraudulent emails, and acquired expertise in optimizing ML models for real-world applications
- Configured models to balance accuracy, costs, and Lambda execution time, optimizing response speeds and minimizing cloud computing expenses
- Implemented scalable AWS solutions, utilizing Lambda, Step Functions, IAM, and Key Stores to streamline DevOps efficiency and system reliability
- Collaborated closely with infrastructure and UX teams for integration, ensuring seamless real-time detection without disrupting end-user workflows

OneSpace, *Cofounder, Full-Stack Engineer* *February 2023 – present*
- Developed LangChain and GPT-3 document assistant that analyzes PDFs, enabling context-aware annotations, summarization, and note creation
- Built an AI-driven autocomplete feature for the word editor, enhancing writing efficiency by suggesting relevant content and streamlining workflows
- Designed a payment structure for premium AI services, offering tiered access to more powerful models via a subscription-based monetization system
- Implemented a full-stack Electron-based desktop app for document annotation and text editing, optimizing UI/UX for seamless interaction
- Secured $50,000+ in funding through the Cozad New Venture Challenge and beta testing phase with several hundred early adopters

CS 222: Software Design Lab, *Course Assistant* *January 2024 – Dec 2024*
- Created course content, including lecture slides and robust web infrastructure, to support the educational experience for 400 students
- Managed and mentored 6 student teams on software development projects, fostering collaboration, workflow methodologies, and code-testing standards

ACTIVITIES

Aether: A Peer-to-Peer Cloud Gaming Platform *January 2025 – May 2025*
- Novelized a decentralized gaming infrastructure leveraging idle consumer GPUs, CPUs, and memory, enabling low-latency streaming
- Developed a C++ backend integrating Windows Desktop Duplication API for real-time frame capture, FFmpeg H.264 encoding, and UDP transmission
- Built a Python-based signaling server and a React frontend to coordinate peer discovery, session negotiation, and client interactions
- Created adaptive scheduling algorithms combining GPU, VRAM, and latency metrics to assign rendering and encoding across heterogeneous peers
- Benchmarked performance against Steam Remote Play and GeForce Now, with 30% lower CPU usage and <60ms latency under typical conditions

Aerial WiFi Network with Drones *August 2024 – December 2024*
- Implemented an aerial WiFi network using drones equipped with Raspberry Pi 4s to extend connectivity across large outdoor and indoor spaces
- Configured secure, real-time ESP32 transmission and optimized signal strength, SNR, throughput, and latency across variable distances and altitudes
- Addressed signal attenuation, interference, and power limitations through directional antennas, beamforming, data encryption, and power management

Founders - Illinois Entrepreneurs, *Project Lead* *February 2023 – May 2025*
- SeriesFarm: AI-Powered Agricultural Loan Assistance Platform
  - Led the development of a personalized funding platform to help farmers access relevant loans and grants
  - Built a fine-tuned chatbot using LangChain and LLaMA, enabling intelligent filtering of funding opportunities based on user queries
  - Implemented a RAG pipeline with Pinecone vector storage for efficient document retrieval and deployed over AWS SageMaker
  - Developed full platform with React and Django REST, storing user data in a scalable SQLite database, and successfully deployed site
- Hanger: AI-Enhanced Fashion Social Media Platform
  - Led a team to develop a fashion-focused social media app, integrating a Pinterest-style feed using React Native and NestJS
  - Designed an AI-powered image detection system with CLIP, BERT, and OpenCV to recognize clothing types and detect brand logos
  - Developed an Hugging Face-based personalized recommendation engine, improving outfit suggestions and user engagement
  - Secured funding from fashion industry leaders, including ZARA and H&M, and collaborated with a client to refine platform features

Disruption Lab, *Technical Lead* *February 2024 – May 2025*
- Developed RAG system to enhance information retrieval for ACCY 200 course content, leveraging a vector store with 10,000+ indexed course materials
- Deployed a scalable NLP pipeline that integrates the model with Pinecone's vector database, optimizing retrieval efficiency for diverse user queries
- Designed and implemented a Streamlit frontend, enabling seamless interaction and improving response accuracy by 30% with conditioned fine-tuning
- Deployed the RAG on AWS SageMaker, overcoming model inefficiencies and kernel crashes with thorough experimentation to ensure minimized latency
- Engineered adaptive text extraction to handle heterogeneous file formats (HTML, XML, PDFs), expediting preprocessing efficiency for AI model inputs

Quant Illinois, *Trading Division* *February 2023 – December 2023*
- Research current strategies for momentum trading algorithms e.g. MAC-D, RSI divergence, Ichimoku Cloud
- Simulated portfolios with backtesting to project stock progression and optimize trading strategies
- Engineered and tested pairs trading and statistical arbitrage algorithms, achieving a 130% gain over the S&P 500 benchmark

TECHNICAL SKILLS

Languages: Python, Java, C++, C, CUDA, JS, TS, Terraform, HTML/CSS, SQL, MongoDB, Neo4j, Swift, MIPS Assembly, x86, Verilog, Golang, Haskell
Developer Tools: Git, Github, VS Code, Docker, AWS Console, Postman, Firebase, GCP, Google Colab, SQLite, MongoDB Compass, Neo4j Browser
Frameworks: React, React Native, Streamlit, THREE.js, Bootstrap, TailwindCSS, Django, Electron, Express.js, Node.js, Next.js, NestJS, Vue, Jest, Sentry, Prisma
Python Libraries/Models: TensorFlow, Keras, PyTorch, OpenCV, MediaPipe, Pandas, Scikit-learn, NumPy, SymPy, SciPy, Matplotlib, BeautifulSoup, Selenium, Stable Diffusion, Transformers, HuggingFace, OpenAI, Langchain, FAISS, CLIP, BERT, LLaMA, GPT, Langchain