BILAL SHAHID MAHMOOD

bmood1@gmail.com | linkedin.com/in/coderbilal | https://github.com/GhostInTheToast

KEY SKILLS

Languages & Backend: Java 8–22 (Spring Boot, Maven), Python 2–3.12 (Django, FastAPI), C# (ASP.NET), Node.js, Kafka
Frontend & Web: JavaScript, TypeScript, React, Angular, HTML/CSS, Bootstrap, Flask
AI/ML & Data: Detectron2, Segment Anything (SAM), GroundingDino, PyTorch, COCO, CVAT, HuggingFace
DevOps/MLOps: Docker (Compose, Traefik), GitLab CI/CD, Jenkins, Bash, Kubernetes, CVAT-CLI, Multi-processing, CUDA
Databases: MySQL, PostgreSQL, H2, Snowflake; NoSQL: MongoDB, Couchbase, Cassandra
Cloud/Infra: AWS (EC2, ECS, EKS, S3, RDS, Lambda, SageMaker, CloudWatch), GCP (GKE, OAuth, LB), IAM, CyberArk
Tools/Misc: Linux, Swagger, Postman, Jira, Splunk, SonarQube, Selenium, Elasticsearch, Distributed Systems

EDUCATION

Post Graduate Program in AI & ML: Business Applications

University of Texas at Austin || Austin, TX
Master of Science (M.S.) in Computer Science (4.00 GPA Rank 1 with published Thesis) | (B.S.) in Computer Science
Lamar University || Beaumont, TX

PROFESSIONAL EXPERIENCE

Moara.io | · Remote **August 2025 – Present**
Head of AI/ML

- Built the retrieval + summarization stack (Elasticsearch + transformers/Hugging Face), improving **Top-3 precision ~28%** and cutting **P95 search latency ~37%**.
- Designed **LLM guardrails** (RAG with citation binding, prompt orchestration, self-checks) to curb hallucinations; drove factuality errors down drastically and raised "answers with citations" coverage to ~99.9**%**.
- Scaled a multi-tenant AWS stack (**Cognito, DynamoDB, EC2/EBS**) with isolated dev/staging/prod and CI/CD; moved from **weekly** to **daily zero-downtime** releases.
- Delivered **data-dense UIs** in **React/Next.js (TypeScript),** streaming results, faceting, relevance/confidence visuals, cutting **time-to-answer ~34%** in usability tests.
- Optimized LLM spend via prompt compression, batching, caching, and token budgeting. **~41% lower** cost per literature review with stable quality.

AutoHDR | Austin, TX · On-site **April 2025 – August 2025**
Senior AI/ML Fullstack Engineer

- Built a secure, full-stack annotation platform on AWS using CVAT (Docker - 19 containers at 1 time, EC2, Traefik) with React + MySQL controls.
- Fine-tuned/trained transformer models for Computer Vision related needs, and Generative AI.
- Led the redesign of image-segmentation pipelines using Meta's SAM v2, Detectron2, and GroundingDino; integrated masks into GenAI image enhancement workflows and AI upscalers.
- Designed Python + FastAPI backend with HTML/JS + Bootstrap frontend; maintained in TypeScript + Next.js.
- Wrote CUDA-safe, multi-process training code for fast inference across GPUs.
- Integrated Hugging Face models (self-hosted + endpoints); optimized pipelines for fast iteration and modular reuse.
- Delivered custom features directly to VIP customers to ensure tailored integration with their platforms, resolving urgent production issues—including those triggered by external infra outages (e.g. Google's 2024 GCP downtime).

Signature Performance, Inc. | · Remote **July 2024 – April 2025**
Full Stack Software Developer

- Enhanced the VA's audit review system, improving operational efficiency for over 2,000 daily audits using **Java Spring Boot, AngularJS, JSP, Junit, and SQL,** while implementing new features and bug fixes to improve experience.
- Led many projects such as creating of new **RESTful API's in microservice architecture**.
- Developed a dispute notification system, integrating backend and frontend technologies to notify auditors of disputed audits in real-time, increasing auditor productivity by 30% through streamlined workflows.

The Western Union | Austin, TX · Hybrid **December 2022 – June 2024**
Full Stack Software Engineer

- Developed and demoed to the CEO the Money Delivery Tracking Journey API and UI that secured funding for WU's Austin Branch. Used **Spring** and **Couchbase** to send the order status using **Kafka** streamed data (async processing), mapped to path and locale using **Camunda**, and displayed **JSON** as progress accordion using **ReactJS**.
- Lead the design and development of enterprise **microservice** applications at Western Union, driving $1B+ annual revenue, utilizing technologies such as **AWS ECS**, **CloudWatch**, **Java Spring, React, Docker, and Jenkins.**
- **Designed and deployed new React based Money Tracker service on frontend for all customers.**

The Reynolds and Reynolds Company | Houston, TX **June 2022 – December 2022**
Full Stack System Software Engineer

- Managed end-to-end application development and delivery for clients, leveraging cutting-edge industry-standard tools.
- Streamlined data extraction from local government websites and Google search results.
- Optimized web scrapers using multithreading tools, reducing application runtime by 88%.
- Ensured accurate custom application development by working closely with clients through PMD to collect and document product requirements; collaborated with a team of development SMEs.

Sobah Systems | · Remote **December 2020 – June 2022**
Software Developer

- Developed custom software for private clients, focusing on scraping, cleaning, and storing large data from over 45,000 wiki URLs per day (Python, Beautiful Soup 4, PostgreSQL, Multithreading).
- Full Stack development (C#, JavaScript, ASP.NET Core, and SQL) for a web application and product used by thousands of customers.
- Worked on a main company product which is used in over 80% of car dealerships in North America.
- Worked alone on said bots alone, demonstrated and taught existing team how to use the bots.

GhostTech Enterprise | · Remote **January 2018 – Present**
Founder and Software Developer

- Founded GhostTech Enterprise as a boutique freelancing venture to build creative, custom-tailored software solutions.
- Specialize in using Python to solve unique technical challenges, from automating workflows to developing data pipelines and backend systems.
- Designed and deployed AI-powered chatbots for platforms like Discord, tailored to client-specific use cases such as moderation, community engagement, and custom APIs.
- Provide one-on-one consulting to individuals and companies, delivering scalable solutions in areas like database architecture, API integration, automation, and intelligent agents.
- Emphasis on rapid prototyping, clear client communication, and delivering clean, maintainable code.