

Mashrur Rahman

Full-Stack & AI Engineer

✉ mashrur950@gmail.com | ☎ +880-1402-148869 | 📍 Dhaka, Bangladesh
🌐 [mashrur-rahman-fahim](https://mashrur-rahman-fahim.com) | 🌐 [mashrur-rahman-fahim](https://mashrur-rahman-fahim.com) | 🌐 mashrurrahman.com

SUMMARY

Full-Stack & AI Engineer and final-year Computer Science student specializing in production LLM systems — real-time voice agents, Retrieval-Augmented Generation (RAG), multimodal AI, and agentic tooling built on the Model Context Protocol (MCP). Currently building an AI-powered children's English-learning app at BRIM AI, and previously shipped an AI news-verification platform at Shottify. First author of a peer-reviewed conference paper (ECCT 2026) on AI-moderated social platforms. Experienced across the stack (FastAPI, Next.js/React, .NET, Node, PostgreSQL/pgvector) and the LLM toolchain (Gemini, Whisper, streaming TTS, vector search). ICPC Asia Regional contestant and a 3rd-place MCP Hackathon finalist (6100+ developers). Seeking research-oriented and engineering roles in AI/ML and full-stack systems.

Research Interests: Machine Learning & Deep Learning, Large Language Models & Agentic AI, Retrieval-Augmented Generation (RAG), Multimodal AI, and the Model Context Protocol (MCP).

PUBLICATIONS

Mashrur Rahman, Chowdhury Ajmayaan Adil, Ahnuf Karim Chowdhury, Nahid Asef, Tahsina Muthaki, and Tanjila Broti, "MemeStream: A Dedicated Meme-Sharing Platform with AI-Powered Content Validation and Real-Time Engagement," *International Conference on Electrical, Computer and Communication Technologies (ECCT)*, 2026. (Accepted; peer-reviewed — all authors at Ahsanullah University of Science & Technology.)

- Proposes an MVC platform (React 19 + ASP.NET Core 9 + PostgreSQL + SignalR) applying Google Gemini multimodal AI for automated content validation, a multi-factor feed-ranking model, and a gamified "LaughScore" engagement system with real-time messaging, evaluated in a user study (4.3/5 satisfaction) — the peer-reviewed research behind the MemeStream project.

PROFESSIONAL EXPERIENCE

BRIM AI

Dec 2025 – Present

AI Engineer — Loveland Kids

[Play Store](#)

- Built a production real-time AI voice-call pipeline (Pipercat orchestration, Google Gemini LLM, Whisper STT, streaming TTS) over a low-latency WebSocket/PCM transport with barge-in handling, delivering ~2s end-to-end response latency (speech to spoken reply, spanning STT, LLM, and TTS)
- Designed a Retrieval-Augmented Generation (RAG) system using pgvector embeddings over curated lore and graded textbooks to deliver age-appropriate English teaching grounded in the child's own course material
- Shipped subscription billing, FCM push notifications, and an admin analytics panel hardened with TOTP-based two-factor authentication
- Engineered a FastAPI + PostgreSQL backend with a React/Vite frontend, deployed on Hetzner; published a signed Android app to the **Google Play Store** and built an **iOS** version (Capacitor/WKWebView)

Shottify

Sept 2025 – Nov 2025

Full-Stack Developer

[Live](#) | [LinkedIn](#) | [Facebook](#)

- Developed an AI news-verification platform with credibility scoring and multi-length analysis generation to combat misinformation and promote media literacy
- Engineered RESTful backend APIs in FastAPI with PostgreSQL (Neon) and Redis caching, plus an automated web-scraping pipeline that ingests news and routes it through AI models for credibility assessment
- Built a responsive Next.js frontend from Figma with a TipTap rich-text editor (Markdown persistence) and an admin panel for manual submission and content moderation
- Deployed on GCP (VM, Cloud Storage) and Render; managed sprints with Jira

SELECTED PROJECTS

FleetMind — Agentic AI Fleet Dispatch (MCP)

[Live](#) | [MCP Server](#) | [Agent](#)

Python, Gemini 2.0 Flash, MCP, FastMCP, PostgreSQL, Docker

- Architected an enterprise Model Context Protocol server exposing 29 AI-accessible tools and 2 real-time resources, letting an LLM run end-to-end delivery dispatch (orders, drivers, assignments) through natural conversation
- Built an autonomous agentic loop on Gemini 2.0 Flash that plans and executes multi-step workflows, feeding each tool result back to the model, and an AI assignment engine analyzing 10+ signals (live weather, real-time traffic, vehicle fit) with transparent reasoning
- Designed a 3-layer multi-tenant auth system (SSE proxy, validation middleware, DB-level filtering) with SHA-256 hashed keys; deployed on HuggingFace Spaces via Docker — **3rd place** at the MCP 1st Birthday Hackathon

(Anthropic & Groq), 6100+ participants

Friday — Agentic AI Personal Assistant

[Live](#)

React 19, Node/Express, Gemini 2.5, Groq, PostgreSQL/pgvector + HNSW, Whisper

- Built an LLM agent with **12 function-calling tools** (tasks, memory, search) on a trust-the-model loop (tool-choice auto, multi-step planning), backed by a Gemini-primary / Groq-fallback model cascade and Whisper voice input
- Engineered a **4-way hybrid retrieval** engine (exact + full-text + semantic pgvector/HNSW + fuzzy, fused via reciprocal-rank fusion) on Neon PostgreSQL, plus cross-platform reminders (web Service Worker + native Android AlarmManager); deployed on Render

RomantAI — LLM Companion with Semantic Memory & Voice

[Live](#) | [Code](#)

Flask, PostgreSQL/pgvector, Groq (LLaMA 3.3 70B), Whisper, WebSocket

- Engineered a 3-tier semantic memory system (short-term, long-term, emotional) using pgvector cosine-similarity retrieval and an LLM fact-extraction pipeline that auto-categorizes and deduplicates user facts
- Built a dynamic prompt-assembly engine over 6+ context blocks feeding a multi-model LLM fallback cascade, plus real-time bidirectional voice calls over WebSocket with server-side VAD, Whisper STT, and emotional TTS; deployed on Render via Docker

AI Photo Detector — Deepfake Image Classification

[Code](#) | [Colab](#)

PyTorch, EfficientNet-B0, CUDA, scikit-learn

- Fine-tuned a pretrained EfficientNet-B0 CNN to classify real vs. AI-generated images on the OpenFake dataset (~448K images) to **91.6% test accuracy**, **0.975 ROC-AUC**, and 0.976 precision, using a production training loop (AMP mixed precision, AdamW + OneCycleLR, EMA, label smoothing, discriminative learning rates) on a Tesla T4 GPU
- Built a publication-grade evaluation harness (12+ metrics, ROC/PR, calibration, t-SNE) and quantified robustness under JPEG, blur, and downscaling with GradCAM saliency explainability

MemeStream — AI-Moderated Social Platform

[Live](#) | [Code](#)

ASP.NET Core 9, C#, React 19, PostgreSQL, SignalR, Gemini

- Integrated Google Gemini multimodal AI for automated meme detection and quality scoring at post time, and built a real-time messaging/notification layer with SignalR hubs (1:1 and group chat, read receipts)
- Engineered a multi-signal personalized feed (time decay, diversity interleaving, seen-post tracking) and a gamified 8-tier “LaughScore” ranking; secured with JWT + BCrypt and containerized via Docker Compose; the published evaluation measured **45 ms** median API latency, **100 ms** real-time message delivery, and **99.8% uptime** (basis of the ECCT 2026 paper)

Additional Projects

Python, Flutter, Laravel, MERN, .NET

- **DataHive**: distributed Facebook/Instagram scraping platform — FastAPI backend with a Celery + Redis task queue, cookie-session management, and orphan-task recovery [[Live](#)]. **Uni-Craft**: Flutter/Firebase campus app with real-time chat and FCM push, shipped as a signed Android APK [[APK](#) | [Code](#)]. **VoxKey**: Linux global push-to-talk dictation daemon (raw evdev hotkey + Groq Whisper, Wayland/X11) [[Code](#)]. **Strategic Defense Ops** (Lead Developer): Laravel 9 + Next.js 14 with 4-tier RBAC, Sanctum + Google OAuth, and a tamper-evident audit trail [[Live](#) | [Code](#)]. **Eventify**: MERN club-event platform with a recommendation engine and PDF certificates [[Live](#) | [Code](#)]. **Daintree**: MERN e-commerce/social apps with JWT/OTP auth and Cloudinary media [[Daintree](#)].

TECHNICAL SKILLS

Languages	Python, JavaScript/TypeScript, C/C++, C#, Java, Dart, SQL
AI / LLM	LLM agents, MCP, RAG & vector search (pgvector), Gemini, Groq
ML / DL	PyTorch, TensorFlow, scikit-learn, CNNs/EfficientNet, Whisper, TTS
Frontend	React, Next.js, Vite, Flutter, Gradio, Tailwind CSS
Backend	FastAPI, Node.js/Express, .NET, Laravel, FastMCP, SignalR, REST APIs
Databases	PostgreSQL, MySQL, MongoDB, Redis, Firebase, Neon
Cloud/DevOps	GCP, Hetzner, Render, HuggingFace Spaces, Docker, Celery, Linux
Security	Cryptography, Web Exploitation, Steganography, JWT/OAuth, 2FA
Tools	Git, Postman, Jira

EDUCATION

Ahsanullah University of Science and Technology — CGPA 3.6/4.0
B.Sc. in Computer Science and Engineering

Expected Dec 2026
Dhaka, Bangladesh

- **Relevant Coursework:** Data Structures & Algorithms, Databases, Computer Networks, Artificial Intelligence, Machine Learning, Operating Systems, Software Engineering

ACHIEVEMENTS & ACTIVITIES

- **3rd place** — MCP 1st Birthday Hackathon (Anthropic & Gradio), among 6100+ participants worldwide (Nov 2025)
- **1st place** — CTF (Capture-the-Flag) cybersecurity competition
- **2nd place** — Eventify Hackathon (2025)
- **2nd place** — University Software Exhibition (*Strategic Defense Ops*)
- **Champion** — Shohoj Coding Contest (2024)
- First-author peer-reviewed publication accepted at ECCT 2026 (MemeStream)
- ICPC Asia Dhaka Regional contestant (2023 & 2024); Competitive programming: LeetCode 180+, Codeforces max 1184, CodeChef 2-star, HackerRank 3-star
- Certifications: Supervised Machine Learning (DeepLearning.AI); HackerRank JavaScript & Problem Solving