ASHISH SUBEDI

Saarbrücken, Germany | iamashishsubedi@gmail.com | asubedi.com.np | LinkedIn | GitHub
M.Sc. Informatik, Saarland University (Expected **Aug 2027**)

PROFESSIONAL SUMMARY

Backend & AI Systems Engineer and M.Sc. Computer Science student at Saarland University with **5 years** of production experience in **Python** and **Go**. Proven record of building low-latency streaming backends, scalable Document AI / RAG pipelines, and cloud infrastructure on AWS and GCP. Focused on LLM inference internals, distributed backend systems, and trustworthy AI.

EDUCATION

Universität des Saarlandes (Saarland University), Germany *April 2025 – Expected Aug 2027*

Master of Science in Informatik (Computer Science) | GPA: **1.7** (German System)

- Relevant Coursework: Distributed Systems, Data Networks, AI Systems (Seminar), Security, Trustworthy ML, Elements of Machine Learning, Data Science

Kathmandu University, Nepal *August 2017 – April 2022*

Bachelor of Engineering in Computer Engineering | GPA: **3.8 / 4.0**

- ML & AI Club Coordinator; Student Research Assistant in IRPRL Lab (Ne-OCR project)
- Relevant Coursework: Data Structures & Algorithms, Operating Systems, Database Systems, Computer Networks, Machine Learning, Speech & Language Processing (NLP)

TECHNICAL SKILLS

Programming Languages:	Python (Expert), Go, TypeScript, C++, SQL, JavaScript, Bash / Shell Scripting
Cloud & DevOps:	AWS (EKS, ECR, Lambda, SQS, Kinesis, Cognito, S3), GCP (GKE, Cloud Storage), Docker, Kubernetes, Terraform, CI/CD, Linux
Backend & Systems:	Distributed Systems, Real-Time Streaming, TCP / WebSockets, FastAPI, Flask, Django, Celery, REST APIs
AI/ML & Inference:	LLMs, RAG, PyTorch, PagedAttention, Vector Search, Embeddings, MistralOCR, OpenCV, AI Agents
Databases & Messaging:	PostgreSQL, MongoDB, Redis, RabbitMQ, AWS RDS

EXPERIENCE

RTKData.com | **Backend Software Engineer (Werkstudent)** *August 2026 – Present | Germany*

Tech: Go, Python, MongoDB, Redis, WebSockets, Network Sockets / TCP, Docker, Linux

- Developing high-throughput backend server infrastructure in **Go** to ensure smooth, low-latency, and reliable delivery of Real-Time Kinematic (RTK / GNSS) correction data streams.
- Implemented client connection health monitoring, automated rover failover, and dynamic base-station switchover to ensure accurate positioning when rovers travel far from a base station.
- Built end-to-end testing pipelines with multi-rover simulations to benchmark performance, stress test servers under load, and validate delivery reliability against SLO thresholds.

Evana AG | **AI & Backend Software Engineer (Werkstudent)** *June 2025 – July 2026 | Saarbrücken, Germany*

Tech: Go, Python, MongoDB, MistralOCR, RAG, Vector Search, Embeddings, Docker, Linux

- Shipped a production-ready document name generation pipeline and extraction microservices in **Go** and **Python**.
- Built bounding-box retrieval pipelines using MistralOCR to improve layout analysis for complex documents.
- Optimized RAG pipelines with hybrid embeddings, improving document classification accuracy by **10%+**.
- Assessed legacy extraction systems and delivered architectural recommendations based on benchmark results.

Docsumo | **Software Engineer (Backend & ML)** *August 2021 – March 2025 | Singapore (Remote)*

Tech: Python, MongoDB, LLMs, OCR, GCP (Kubernetes), AWS (SQS, S3), Celery, Redis, Docker, CI/CD

- Optimized document processing architecture through smart request batching and optimized DPI conversion to minimize file size without affecting quality, delivering **3x performance** and **9x cost reduction**.
- Built ETL pipelines for LLM training data with **10x speed improvements** and strong data integrity controls.
- Led few-shot learning experiments across GPT-4o, GPT-4o-mini, Mixtral-8x7B, and Llama 3.1, achieving **81% accuracy** and deploying the best model to production.
- Delivered client-specific extraction solutions with **90%+ accuracy** using ensembles of deep learning models and LLMs.
- Created a key-value extraction framework cutting API development time by **70%**, and implemented reliable Salesforce synchronization integrations using custom APIs and Metadata API.
- Built an internal reporting system that reduced report generation time from **hours to minutes**.

Belvy LLC | **Software Engineer (Cloud & Infrastructure, Contract)** *August 2021 – September 2022 | USA (Remote)*

Tech: AWS (EKS, ECR, Lambda, Kinesis, Cognito, S3), Kubernetes, PostgreSQL, Docker, Ruby on Rails

- Migrated a legacy monolithic Ruby on Rails application to containerized cloud infrastructure on AWS using **EKS (Kubernetes)**, **ECR**, and **AWS Lambda**.
- Designed a high-concurrency PostgreSQL usage-metering and telemetry system with stored procedures for real-time customer usage billing.
- Implemented secure authentication with AWS Cognito (MFA) and built scalable real-time streaming ETL pipelines with **AWS Kinesis Firehose** and S3.

NAAMI Research Institute | **Research Intern** *February 2021 – May 2021 | Remote*

Tech: Python, CNNs, Selenium

- Built voice disease classification models with **80%+ accuracy** using CNNs and traditional ML.
- Automated data collection with Selenium, reducing manual effort by **90%**.

SELECTED PROJECTS

ingine: Continuous Batching & PagedAttention LLM Engine | Python, PyTorch, Systems Architecture *GitHub · Blog*

- Built a continuous batching LLM inference engine from scratch in **Python/PyTorch** with **PagedAttention** (16-token logical-to-physical KV cache blocks) for Qwen2.5-1.5B.
- Benchmarked on Apple Silicon (MPS): achieved **1.29x throughput speedup** (19.1 vs. 14.8 tok/s) and **+8.5s faster turnaround** on long requests while profiling latency tradeoffs.

MinClaw: Personal AI Agent | TypeScript, Bun *GitHub*

- Built a self-rewriting personal AI agent daemon in **TypeScript** featuring a CLI/TUI, Web UI dashboard, hot-reloaded config, and WhatsApp, Telegram, and Slack integrations compatible with OpenClaw skills.

Folder Manager for Google Docs, Sheets & Slides | React, Chrome Extension API *Chrome Web Store*

- Engineered and published a Chrome extension adding smart sidebar folders for Google Docs, Sheets, and Slides, serving **1,000+ active users** with a **4.6★** rating.

YFS: Distributed Fault-Tolerant File System | C++, Distributed Systems, RPC, POSIX *Saarland University*

- Built a distributed file storage cluster in **C++** featuring client-server RPC communication, extent caching, directory hierarchy management, and two-phase locking for concurrency control.

SELECTED TECHNICAL WRITING

- **Why Antigravity Eats Weekly Quota Fast**: Disassembled a 193 MB Go binary and empirically probed Gemini's 5-minute implicit prompt-cache TTL cliff.
- **Speculative Decoding** & **I Kinda Understand LLM Serving**: Deep dives into PagedAttention, continuous batching, parallel causal-mask verification, and lossless rejection sampling.
- **Reverse Engineering OnePlus Nord Buds 2**: Sniffed Android HCI logs to decode the RFCOMM battery protocol and build a macOS IOBluetooth menu bar plugin.

LANGUAGES

English (Fluent/C1), German (A2), Nepali (Native), Hindi (Fluent)