

ANKIT KUMAR

miankitsingh@gmail.com ◇ Patna, Bihar

ankit.kryvik.com ◇ github.com/ankit-pn ◇ linkedin.com/in/ankit-pn ◇ huggingface.co/ankit-pn

SUMMARY

Software engineer, ~3 years at Publicis Sapient, building Spring Boot microservices for a global investment bank's private wealth platform (99.9%+ availability, ~40% lower latency). Also builds, fine-tunes and evaluates LLM systems end to end.

WORK EXPERIENCE

Associate Software Development Engineer 2 Nov 2023 – Present
Publicis Sapient Bengaluru, India

Client: global investment bank (US) — private wealth platform Aug 2024 – Present

- Led system design (HLD/LLD) for the platform and drove architecture decisions in bar-raiser forums with zero regressions.
- Owned end-to-end delivery of Spring Boot microservices serving private wealth users at **99.9%+ availability**.
- Built property-driven, configurable GraphQL/REST APIs that multiple divisions can adopt without code changes.
- Added Redis and in-memory caching with startup hydration, cutting average latency **~40%** and stabilizing p99.
- Established production observability with Prometheus and Grafana, reducing MTTR **~30%**.
- Mentored engineers and set backend development standards across the team.

Tech: Java, Spring Boot, GraphQL, REST, Redis, Resilience4j, Prometheus, Grafana, Gatling

Research Intern (Remote) May 2023 – May 2024
Rutgers University Remote

- Built automated data-collection and analysis tools (FastAPI, React, SQLite), improving research efficiency.

PROJECTS

FunctionGemma Phone Assistant — on-device function calling Oct 2026

- Built a 4,000-example English/Hinglish function-calling benchmark (20 phone tools; single, parallel and no-call requests) with leakage-safe splits; published dataset and model on Hugging Face (481 model downloads).
- Full-parameter fine-tuned FunctionGemma-270M with DDP on 2× T4 GPUs: held-out exact match **22.3% → 75.4%**, tool selection **31.6% → 93.0%**.

Tech: PyTorch, Hugging Face Transformers, DDP, Kaggle GPUs

Kyoto — AI browser agent for developers May 2026

- Built a Chrome side-panel agent for any OpenAI-compatible provider (OpenRouter, DeepSeek, Ollama, ...) that drives the browser through ~100 tools: DOM, network capture, cookies, debugger, shell.
- Implemented a streaming tool-calling loop (up to 12 round-trips per turn) with reasoning passthrough, MCP servers over JSON-RPC, per-tab sessions, cross-session memory and per-message cost tracking.

Tech: JavaScript, Chrome Extensions (MV3), MCP, OpenAI-compatible APIs

Hide-and-Seek Replication — emergent tool use for \$3 Jun 2026

- Replicated OpenAI's multi-agent hide-and-seek (ICLR 2020) through the chase and fort-building phases for **\$3.10** of GPU time: PufferLib PPO at 1.2M agent-steps/s, 20B steps.

Tech: C, Python, PufferLib, PPO

TECHNICAL SKILLS

Languages:	Java, Python, TypeScript, JavaScript, C, C++, Go, SQL
Backend:	Spring Boot microservices, system design (HLD/LLD), GraphQL, REST, Spring Batch, FastAPI, Node.js, WebSockets, JWT
Reliability:	Redis caching, Resilience4j, Prometheus, Grafana, Gatling load testing
Data:	PostgreSQL, MongoDB, Redis, BigQuery, Cloud SQL, Elasticsearch, Apache Kafka, Parquet
Cloud & DevOps:	GCP (Cloud SQL, BigQuery, Cloud Storage), Docker, Kubernetes, Jenkins, Nginx, Linux
ML / Training:	PyTorch, Hugging Face Transformers, PEFT (LoRA/QLoRA), DDP, mixed precision, PPO (PufferLib)

EDUCATION

Maulana Azad National Institute of Technology (MANIT), Bhopal June 2024
B.Tech, Computer Science and Engineering — CGPA 8.28/10

ACHIEVEMENTS

1st prize, IEEE Ideathon ◇ 1st prize, NITS Hack 5.0