Ali Asghari

Senior Software Engineer

asghari.ali10@gmail.com | +989147146122 | linkedin | github
Open to Relocating to Barcelona, Spain (Visa/Relocation Support Required)

PROFESSIONAL SUMMARY

Senior Software Engineer with 6+ years of experience specializing in AI-driven applications and high-scale Python backends. Expert in architecting RAG pipelines, fine-tuning LLM workflows (Qwen, GPT, Gemini), and deploying production-grade AI services. Proven track record of handling 3M+ daily requests and maintaining high system reliability in various industries.

TECHNICAL SKILLS

- **AI/ML:** RAG, LLM Guardrails, LLM Orchestration (Ollama, Groq), Vector Databases (ChromaDB), Whisper (STT), Transformers, Prompt Engineering.
- **Backend:** Python (FastAPI, Django, Flask), Celery, RabbitMQ, RESTful APIs, System Design.
- **Data & DevOps:** PostgreSQL, Redis, S3, Docker, Docker Swarm, CI/CD, Grafana, Prometheus.
- **Languages:** Python, JavaScript (React.js), SQL, Java.

WORK EXPERIENCE

AI Software Engineer | Careberry *Dec 2025 – Present*

- Designed a high-reliability RAG orchestration layer for healthcare data, implementing LLM response guardrails and validation checks to mitigate hallucinations and ensure 100% adherence to clinical safety protocols.
- Optimizing LLM inference pipelines to improve response latency and accuracy for internal healthcare workflows.
- Integrating vector-based search to streamline access to large-scale medical knowledge bases.

AI Software Engineer | G2Tech *Nov 2024 – Nov 2025*

- Developed an automated Q&A generation pipeline using Whisper and multi-model LLMs, processing video content into structured educational assets.
- Engineered a real-time, voice-enabled Q&A RAG system using ChromaDB and Ollama models, allowing users to query lecture content via natural speech.
- Implemented a Django-based "Chat with PDF" application using Qwen models, enabling context-aware conversations on user-uploaded files.

Senior Software Engineer | Nobitex *Apr 2021 – Aug 2023*

- Maintained and optimized high-throughput APIs via Flask, handling 3 million daily requests for a user base of 6+ million.
- Redesigned a Django reporting engine, achieving a 66% reduction in generation time for critical financial reports.
- Enhanced system responsiveness by implementing advanced Redis caching strategies and PostgreSQL query tuning.

Software Engineer | Snappfood *Oct 2019 – Feb 2021*

- Refactored the order-matching engine using the Hungarian Algorithm, resulting in a 6x speed increase for 3,000+ daily deliveries.
- Built an autonomous financial microservice using Celery/RabbitMQ to manage weekly payouts for 1,000+ fleet members.
- Deployed dynamic auto-scaling solutions using Docker Swarm and monitoring dashboard with Grafana and Prometheus, ensuring 99.9% uptime during peak traffic hours.

EDUCATION

- **MSc in Computer Science** | Tarbiat Modares University | Graduated 2022
- **BSc in Computer Science** | University of Tabriz | Graduated 2019