SAYYOR YUSUPOV

✉ sayyor.yusupov@gmail.com
☎ +33605919675
Tashkent, Uzbekistan (Remote)
linkedin.com/in/sayyory
github.com/SYusupov

Professional Summary

MLOps and AI Application Developer with 5+ years of total programming experience and 2+ years specializing in engineering and deploying scalable, end-to-end AI applications on Google Cloud Platform. Proven expertise in building RAG pipelines, production-level MLOps (Cloud Run, Cloud Functions, Terraform), and full-stack development (FastAPI, Django).

Technical Skills

- **MLOps & Cloud:** GCP (VertexAI, Cloud Run, Cloud Functions, Cloud Tasks, Cloud SQL, Scheduler, Cloud Storage, BigQuery, Firestore, Secret Manager, Terraform), Docker
- **AI & ML:** LLMs (RAG, Agents, Prompt Engineering), Reinforcement Learning (Stable Baselines3), PyTorch, TensorFlow, Scikit-Learn
- **Back-End & Data:** Python, Django, FastAPI, Flask, SQL (PostgreSQL), Apache Spark, Hadoop, Pandas, NumPy
- **Languages:** English (Fluent), Russian (Fluent), Uzbek (Native)

Experience

Artificial Intelligence Engineer (Oct 2024 - Present)

Valeo — *Créteil, France*

- Designed and deployed hybrid LLM pipelines (UI-based and Google Workspace event-triggered) for large-scale document extraction; implemented a **2-stage split-merge strategy** to handle token limits, reducing manual processing by **80%**.
- Managed a scalable MLOps infrastructure on **Google Cloud Platform (GCP)**, using Cloud Run, Cloud Functions, Cloud Tasks, Cloud SQL and others.
- Collaborated on developing and shipping high-performance REST APIs with **FastAPI** and **Flask**.
- Optimized **Retrieval-Augmented Generation (RAG)** pipelines through advanced prompt engineering, boosting response relevance by **40%** and reducing hallucinations.
- Prototyped, tested, and validated classical and deep learning models for text classification, prompt optimization, confidence scores.
- Engineered autonomous agents using **Agent Development Kit** to automate complex internal workflows.
- *Tools: GCP, FastAPI, Flask, Gradio, Streamlit, VertexAI, RAG, Agent Development Kit*

AI Engineering Intern (Mar 2024 - Aug 2024)

Valeo — *Créteil, France*

- Developed and implemented Reinforcement Learning models for Electronic Design Automation.
- Improved existing solutions by making them 25% more stable by maintaining the quality.w
- Incorporated state-of-the-art methods from academic literature.
- Optimized model performance by fine-tuning parameters and reducing execution time.
- *Tools: Stable Baselines3, MATLAB, Python, PyTorch, Docker, GitHub*

Python Developer & Code Reviewer (Apr 2021 - Oct 2021)

Rialtic Inc. — *USA, Remote*

- Converted structured medical/clinical policy specifications into executable Python code.
- Collaborated with a global team and reviewed code for quality assurance.
- *Tools: Python, PyTest, Git*

R&D Work for Combinatorial Optimization (Mar 2020 - May 2021)

Amoeba Energy Co., Ltd. *Japan*

- Programmed heuristics and algorithms on Vehicle Routing and Task Assignment of Vehicles.
- Tested and optimized solutions for computational efficiency and optimality.
- *Tools: Python, Matplotlib, Seaborn, Git*

Coding Projects

UzNews Digest: End-to-End Serverless AI News Pipeline

Live Project: (Link to Telegram Channel) *Code: (Link to GitHub Repo)*

- Designed and deployed a fully automated, serverless news aggregation pipeline on **Google Cloud Platform**.
- Architected a decoupled microservices system using **Cloud Run (asynchronous with FastAPI)**, **Cloud Tasks** for resilient, rate-limited queueing, and **Cloud Scheduler** for 30-minute automated runs.
- Implemented universal, site-agnostic data extraction using **LLMs (VertexAI, `crawl4ai`)**
- Used **VertexAI's Embedding API** for advanced duplicate detection.
- Used **Firestore** (with native TTL policies) for cost-effective data management.
- *Tools: GCP (Cloud Run, Cloud Tasks, Firestore, Scheduler), Python, FastAPI, LLMs (OpenAI), `crawl4ai`, Docker, Telegram Bot API*

Actors and Movies Statistics Website

Live Webpage: MovieStats.Online, *Code: (Link to GitHub Repo)*

- Engineered a full-stack web application using **Django** and deployed on **GCP Cloud Run**.
- Automated a serverless data-ingestion pipeline from the TMDB API using **Cloud Functions** and **Cloud Scheduler**.
- Used **Cloud SQL (PostgreSQL)** for robust data storage and retrieval.
- *Tools: Python, Django, GCP (Cloud Run, Cloud Functions, Scheduler, SQL), TMDB API, HTML/CSS*

Paper Implementation: Fine-tune BERT for Summarization

Links: Report, Code, Slides

- Implemented BERTSUM for extractive text summarization.
- Developed approaches for processing texts exceeding BERT's 512-token limit.
- *Tools: PyTorch, NLTK, Numpy, Pandas, Transformers, Matplotlib*

Peer-to-peer Delivery Service Cargo

Repos: Data Management, Data Analytics

- Utilized Apache Hadoop, Airflow, PostgreSQL for data storage and queried data with Apache Spark.

Education

Erasmus Mundus Master Program (Big Data Management & Analytics) (2022 - 2024)

Awards: Erasmus Mundus Joint Master Scholarship

- **Centrale Supélec, Université Paris-Saclay** (France)
- **Universitat Politècnica de Catalunya** (Spain)
- **Université Libre de Bruxelles** (Belgium)

Bachelor of Arts in Environment and Information Studies (2017 - 2021)

Keio University *Japan*

- *Awards: Masatada Kobayashi Scholarship for International Students (2017-2021)*