

LUCA PALONCA | Software Engineer

+39 346 4743 229 | luicapalonca@gmail.com | Como, Italy (Flexible across US time zones)

github.com/liuuuk311/ | linkedin.com/in/luca-palonca/ | luicapalonca.com

Software Engineer & Technical Lead with 7+ years of experience building production-grade AI systems, including real-time conversational agents and large-scale LLM pipelines. Specialized in Python backend development, Google Gemini/Vertex AI, agent architectures, prompt engineering, structured outputs, and multi-modal (audio) integrations. Proven track record delivering high-impact features in healthcare SaaS that processed millions of AI interactions and drove significant business growth. Experienced working remotely with US-based engineering teams across multiple time zones, ensuring clear communication, collaboration, and reliable delivery in distributed settings.

Work Experience

Backend SWE & Tech Lead - Barti Software

San Francisco, US (Remote) - **02/2024 - Present**

Tech stack: **Python (Flask, FastAPI), PostgreSQL, React/TypeScript, Docker, Kubernetes (GKE), GCP (Vertex AI, Pub/Sub, Cloud Run), Twilio**

- **Led backend architecture and team of 5 engineers as Tech Lead** (promoted P1 → P2 → P3), driving architecture decisions, code reviews, standups, onboarding, and delivery of high-impact features in a multi-service healthcare SaaS platform.
- **Designed and led development of Quinn AI Agent Platform** — custom real-time conversational agent framework built on FastAPI + Google Gemini Live (Vertex AI). Supported multi-modal audio streaming (Twilio Media Streams + PCM), agent architectures, tool/function calling, YAML-configured behaviors, and structured Pydantic outputs. Delivered AI Phone Receptionist and Silent Exam Scribe features (3,000+ exams documented).
- **Built production AI Document Summarization pipeline** using Gemini 2.5 Flash: implemented prompt engineering with clinical-safety guardrails, structured extractive outputs (Pydantic), event-driven architecture (Cloud Pub/Sub), and comprehensive monitoring/evaluation of 2M+ patient documents.
- **Developed scalable AI infrastructure** handling 3.5M+ LLM requests with monitoring, cost controls, rate-limit handling, and audit logging; incorporated lightweight RAG patterns for context injection (org catalog, patient data).
- **Built and scaled automated data migration pipeline** (batched, multi-threaded GKE workload): migrated 51M+ records across 276 practices (23.7M files, 9.5M appointments, etc.) over 2,100+ runs, replacing manual weeks-long processes with near-zero-touch automation.
- **Delivered 10+ external integrations** and key platform features (DoseSpot e-prescribing, FramesData/Vision Web optical ordering, CapitalOS financing, Family Billing, Online Scheduler with availability engine, Machine Integration Electron app, Self-Service Onboarding), contributing to \$1M ARR milestone and enabling CS to onboard more practices weekly than prior quarters.
- **Standardized backend architecture** across services (multi-tenancy, OpenAPI automation, CI/CD improvements) and supported full-stack delivery in React/TypeScript, improving maintainability, scalability, and cross-team velocity.
- **Mentored team and supported Engineering Manager** with Jira prioritization, consistently rated "Exceeds Expectations" for reliability, technical mentorship, and impact.

Backend SWE - Cedar Software

New York City, US (Remote) - **11/2022 - 02/2024**

Tech stack: **Python, Flask, PostgreSQL, Redis, Docker, Kubernetes, GCP**

- Improved API response time by 95.6% from 8s to 350ms through meticulous data analysis and query optimization.
- Slashed infrastructure costs by \$42k annually by optimizing GCP geocoding requests
- Developed a data migration pipeline for onboarding customers' data, reducing import time by 40% through streamlining processes and optimizing data transfer efficiency, migrated more than 300k records
- Implemented 50+ cypress end to end tests, ensuring a 85% increase in test coverage and bolstering system

reliability

- Provided essential support to the frontend team by contributing to various tasks using React and Typescript, leveraging technical expertise to enhance application functionality and user experience.

Backend SWE - *Sailogy SA*

Chiasso, Switzerland - **03/2020 - 11/2022**

Tech stack: **Python, Django, DRF, PostgreSQL, Docker, Redis, RabbitMQ**

- Achieved 97% test coverage on business logic by adding 100+ unit tests using pytest
- Successfully led the decomposition of the Django monolith into multiple microservices using FastAPI, enhancing system scalability and maintainability

ETL Engineer - *Avvale (formerly Techedge Group)*

Milan, Italy - **01/2019 - 03/2020**

Education & Research

- Co-author of peer-reviewed paper: "Online anomaly search in time series: significant online discords" (Knowledge and Information Systems, Springer, 2020). Developed statistical approach for real-time anomaly detection in data streams, minimizing false positives.

Master's degree in Computer Science - *Università degli Studi di Milano-Bicocca*

10/2019

Bachelor's degree in Computer Science - *Università degli Studi di Milano-Bicocca*

10/2017

Skills

Prompt Engineering, Structured Outputs (Pydantic), Tool/Function Calling, Agent Architectures, Real-time Audio Streaming, Multi-modal AI, Guardrails & Safety Prompts, LLM Evaluation & Monitoring, Vertex AI / Gemini, Lightweight RAG. Python, FastAPI, Flask, PostgreSQL, Docker, Kubernetes (GKE), GCP (Pub/Sub, Cloud Run), CI/CD, React/TypeScript, System Design, Performance Optimization, Git, API Architecture, System Design