Utkarsh Upadhyay

+91 9125299916 | btoshine774@gmail.com | linkedin.com/in/utk2103 | github.com/utk2103

Professional Summary

AI Engineer delivering production-grade LLM agent systems, multimodal image pipelines, and end-to-end ML solutions for enterprise clients across jewellery, PR, and EdTech — owning full pipelines from prototyping to production.

Experience

M37 Labs **Jan 2025 – Present**
AI Engineer *Gurugram, India*

- Architected a **multi-tenant AI platform** (EBIC.ai, PR domain) on Python/FastAPI with tenant-aware model routing, dynamic secret resolution from **AWS Secrets Manager**, and isolated runtime configuration across OpenAI, Gemini, SERP, Data365, MongoDB, and S3 — owned the full system from MVP to production.
- Built a production **LLM copilot** using OpenAI Agents SDK with function-calling workflows for brand intelligence, competitor analysis, social insights, task automation, and multimodal generation; persisted threads and tenant-isolated context in MongoDB.
- Designed **async intelligence pipelines** fusing web search, media ingestion, and social signals into executive narrative reports with share-of-voice, sentiment analysis, and predictive insights; built **WebSocket-based voice AI relay**, cron-driven multi-tenant refresh jobs, and one-click export to PDF/DOCX/slides.
- Implemented **retrieval and inference optimization** layers including batch embeddings, vector-search context assembly, cache/TTL controls, retry/fallback logic, audit logging, and signed S3 delivery with automated testing of core pipelines.
- Architected an **AI jewellery image pipeline** (Gemini 2.5 Flash) with section-aware prompt-engineering, classifier pre-pass for ground-truth injection (stone cut, bezel shape, metal tone), and multi-view generation (Uploaded and creative variants per SKU).
- Built a **hybrid BG removal stack** layering Gemini semantic segmentation → rembg (U2Net) → Otsu thresholding → morphological open/close → OpenCV post-processing (HSV correction, elliptical shadow, white enforcement at ≥245px); resolved cross-view geometric drift via classifier-injected constraints.
- Engineered a **three-tier pipeline state system** (in-process dict, Redis TTL cache, PostgreSQL) with semaphore-limited asyncio, 180s hard timeout, exponential-backoff retry (4 attempts), WebSocket progress at 300ms throttle; built a **LangGraph reflection agent** (critique → regen, up to 3 cycles) with APScheduler cron and full audit logs in PostgreSQL JSONB.
- Built an **automated multi-view clothing image pipeline** for a leading fashion retail group using state-of-the-art image generation models, delivering renders directly to retail visual merchandisers and reducing manual image review effort by **30%**.

EasyEd Pvt Ltd **Jun 2024 – Aug 2024**
Machine Learning Engineer *Remote*

- Built an AI ed-tech chatbot for **3 user personas** serving **1K+ test users**; implemented intent-based NLP routing, improving response relevance by **20–25%** and reducing counselor intervention by **30%** via adaptive conversational flows.
- Reduced counselor intervention and manual guidance effort by **30%** through adaptive conversational flows and feedback-driven refinement.

Projects

twain.ai Link | *Next.js 14, FastAPI, Python, Gemini 2.5 Flash, Multimodal AI* **Feb 2026**

- AI jewellery design workspace compressing months of design iteration into hours — integrated Gemini 2.5 Flash for multimodal image generation (text-to-image & image-to-image redesign); built a **multimodal FastAPI inference pipeline** with conditional prompt construction based on input modality and a **custom canvas UI** with real-time style/stone/metal parameter control feeding dynamic prompt templates.

Prompt Studio Link | *FastAPI, Next.js 14, PostgreSQL, pgvector, Docker, Python* **Mar 2026**

- AI prompt analysis platform with a **7-dimensional scoring engine** (clarity, specificity, token efficiency, mode alignment, etc.) — fully deterministic, no LLM dependency, sub-10ms response; pgvector IVFFlat cosine index on 1536-dim embeddings for semantic search; full stack containerized via Docker Compose with Alembic auto-migrations.

Text Summarizer Link | *PyTorch, HuggingFace Transformers, PEGASUS, Python* **Oct 2023**

- Fine-tuned **Google PEGASUS** on SAMSum dialogue corpus for abstractive summarization; evaluated with ROUGE-1/2/L via 8-beam decoding; modular config-driven pipeline (ingestion → tokenization → training → eval) using PyTorch Trainer API with gradient accumulation.

Skills

- **Languages:** Python, SQL, TypeScript
- **Web & Database:** FastAPI, Flask, REST APIs, PostgreSQL, MongoDB, Redis, pgvector, Vector Databases, Streamlit
- **Frameworks & Libraries:** NumPy, Pandas, scikit-learn, TensorFlow, LangChain, LangGraph, PyTorch, Pydantic, OpenAI, LlamaIndex, HuggingFace
- **Cloud & Tools:** AWS (S3, Secrets Manager, Textract), GCP, Git, Linux, Docker, VS Code
- **ML & AI:** LLMs, Generative AI, Prompt Engineering, RAG, Agentic AI, Multi-Agent Systems, Model Fine-tuning, Inference Optimization, Embedding Models, OCR

Achievements & Recognition

- **Publication:** Co-authored & presented "IntelliTask: An AI-Driven Enterprise Task Management System" at ICTIS 2026; selected for **Springer** conference publication.
- **Core Member, Django India** — one of India's most active Python/Django communities; led events at devXsphere; mentored in GSSoC, SSoC, GDSC-KIET, DSDL-KIET; participated in Smart India Hackathon (Nexsafar.ai).

Education

KIET Group of Institutions, Ghaziabad **Nov 2022 – Jun 2026**
Bachelor of Technology, Information Technology; GPA: 7.9/10 *Delhi-NCR, India*