MANTHAN TEOTIA

Generative AI Engineer | LLM Agents · RAG · Multimodal AI Video & Image Generation
Ghaziabad, Delhi NCR, India | +91 8826090292 | choudhary.manthan01@gmail.com
linkedin.com/in/manthan-teotia-43641222a | manthan-teotia.pages.dev

Don't read it, watch it: 23 AI-made ads, 3 AI presenter reels and a 90-second film → **manthan-teotia.pages.dev**

SUMMARY

Generative AI Engineer who cut the cost of a video ad by ~98% with AI. 3+ years building production AI: LLM agents, RAG, multimodal text-to-video pipelines and ML. At Brightfold, core engineer of the Creative Studio built for client Vibelets, an AI video-ad platform demoed to 12+ agencies and D2C brands across India, the US and Canada: a 15-second ad in ~5-10 minutes for ~$2, replacing influencer videos. Worked directly with the BytePlus (ByteDance Seedance) and Google Gemini teams to take new video models to production. Earlier built a multi-agent RAG chatbot over 10,000+ contracts and cut supplier selection from 30 minutes to 2 seconds.

EXPERIENCE

Generative AI Engineer | Brightfold | Remote **Dec 2025 - Present**

Client: Vibelets. Built their Creative Studio: a product photo or a winning ad in, a ready-to-run ad out.

- **Cut cost per video ad by ~98% (from ~$120-300 influencer videos to ~$2 per 15-second AI ad) and production time from days to ~5-10 minutes** by taking the Creative Studio from early build to production and a V2 canvas, replacing paid influencer content; helped Vibelets win new clients.
- **Powered pilots with 12+ agencies, D2C brands and an affiliate network across India, the US and Canada (3 onboarded, 2 on trial)** on a multimodal pipeline I built as scalable microservices (Python, FastAPI, Redis, PostgreSQL, AWS): LLM script writers (Gemini, OpenAI) and text-to-video on Seedance and Veo with automatic provider fallback; worked directly with the BytePlus and Google Gemini teams.
- **Built the image-ad pipeline and a 12-model catalog across 6 providers** (Google Nano Banana, OpenAI GPT Image, ByteDance Seedream, Alibaba Qwen, xAI Grok, Flux), matching each creative to the right model by ratio, resolution and cost; plus image Clone and brand-kit-aware scripts.
- **Made Clone ~5x faster (77 s to 16 s)** by rebuilding its reference-ad analysis on Gemini Flash; Clone remakes any winning ad with the customer's own product and words.
- **Turned customer demo requests into shipped features:** agencies asked for Hinglish ads, more hooks to A/B test and better captions; delivered Hinglish video ads, Hook Variations (5 openings per ad) and a 51-style caption library, and kept Hinglish live through a vendor model shutdown with a production hotfix (Veo 3.1 to Gemini Omni).
- **Raised output quality until media buyers said customers would not feel it was AI,** with evaluation and guardrails for generated video (transcript word-accuracy checks, product-drift detection and re-filming, OCR-aware captions), tuned on ~3,000 generated videos.
- **Supported Vibelets' early in-house video-model effort:** prepared training data and ran small-scale fine-tuning experiments on GPUs (LoRA, PyTorch, Hugging Face); worked with the deployment engineer on Docker and Kubernetes rollouts.
- **Fixed 20 customer-reported issues across 3 services in 3 days** during a pre-launch UAT round, test-first with pytest, Playwright and CI; added automatic retries on a second AI provider, failure monitoring, one charge at the quoted price and automatic refunds.

Junior Machine Learning Engineer | EKSUM COMTECH (formerly Aksum Trademart) | Noida **Jan 2024 - Nov 2025**

- **Built ProcureIQ, a multi-agent RAG chatbot** (LangGraph, LangChain, FastAPI, FAISS vector search, embeddings) with QA, summarizer and analyzer agents for document processing, answering in under 2 seconds across 10,000+ contracts and cutting manual review by 75%.
- **Built an RFQ-to-vendor matching engine** (NLTK entity extraction, 384-dim FAISS semantic search, multi-factor scoring): vendor selection from 30 minutes to 2 seconds, 94% match accuracy, 500+ requests a day.
- **Built a hybrid B2B credit-risk engine:** XGBoost with SHAP explanations, Isolation Forest fraud checks, FAISS case retrieval and Gemini-written credit memos, served in real time through FastAPI with dual risk thresholds for policy A/B testing.
- **Forecasting and sales ML:** price forecasting (ARIMA) cut procurement cost 5-8%, demand forecasting (SARIMA + XGBoost) cut stockouts 10%, lead scoring lifted conversion 12%.

Operations & Data Analyst | Valeur Fabtex | Delhi **Feb 2023 - Dec 2023**

- Built a spaCy NLP email classifier for 2,000+ vendor emails (triage time -50%), an ML dropout-prediction model (retention +25%) and Power BI dashboards (reporting 40% faster).

SKILLS

Generative AI & LLMs: Large Language Models (LLMs), AI agents and agentic workflows, multi-agent systems, RAG, prompt engineering, fine-tuning (LoRA with Hugging Face PEFT, PyTorch), LangGraph, LangChain, LlamaIndex, embeddings, vector databases (FAISS, Pinecone), semantic search, chatbots, LLM evaluation and guardrails, OpenAI, Anthropic (Claude), Google Gemini APIs
Multimodal & Video: text-to-video (ByteDance Seedance, Google Veo 3.1, Gemini Omni), image generation and multi-model routing (Nano Banana, GPT Image, Seedream, Qwen, Flux), computer vision (OCR, video QA), speech-to-text, BytePlus Ark, ffmpeg
Machine Learning: XGBoost, scikit-learn, SHAP, anomaly detection, time series (ARIMA, SARIMA), NLP (spaCy, NLTK), feature engineering, A/B testing
Engineering & MLOps: Claude Code (agentic engineering), Python, FastAPI, Flask, REST APIs, microservices, React, PostgreSQL, SQL, Redis, Docker, Kubernetes, AWS (S3), CI/CD, Git, pytest, Playwright, LLMOps, observability and monitoring, latency and cost optimization, production deployment

EDUCATION

Bachelor of Computer Applications (BCA) | Jiwaji University, Gwalior | CGPA 8.5/10 **2021 - 2024**