Aniket Sakpal

(412) 983 9955 | www.linkedin.com/in/aniket-sakpal/ | aniket17sakpal@gmail.com | USA

Summary

- Machine Learning Scientist with **10 years building production-grade AI for high-stakes environments**—Search, Ranking, and Marketing Bidding systems
- Holder of **2 U.S. patents** in AI Hallucination Detection & Agentic Systems; Consistently rated top performer with **highest-band reviews** at Expedia Group
- Co-founded and built from scratch of **AI learning co-pilot platform** for students -> www.aitado.com

Education

Carnegie Mellon University | Pittsburgh, USA
Master's in Data Science
Key C/W: Deep & Machine Learning, Distributed Systems, Graph Theory, A/B Testing, Regression, ML Deployment

Skills

Full-spectrum expertise spanning LLMs/Agentic AI, causal inference, experimentation, regression modeling, and large-scale ML deployment:

- **AI & ML:** Agentic AI, RAG, Deep Learning, Reinforcement Learning (PPO, GRPO), Multi-Agent Systems, Video LLMs
- **Engineering:** PyTorch, FastAPI, Flask, Next.js, LangGraph, SQL, PySpark
- **Analytics:** A/B Testing, Interleaving, Regression, Predictive Modeling, Optimization, Feature Engineering
- **Tools:** Python, Tableau, R, Git, Distributed Systems, Knowledge Graphs

Experience (10 Years)

Expedia Group | Austin, TX
Machine Learning Scientist - Agentic & Multi-modal AI | 06/2024 - Present

- Built **production-grade AI analytics copilot** with adaptive reasoning loop — intent understanding, autonomous tool selection via ReAct, multimodal execution, and episodic history compaction — mirroring the architecture of production coding agents like Claude Code. The system includes:
  - Designed a **declarative skill system** where analytical workflows (metric diagnosis, root cause analysis etc.) are YAML-manifest-driven and hot-loaded — encoding reproducible causal breakdowns without LLM drift
  - Implemented a high-throughput **state management layer** backed by Redis with key-sharding, TTL policies, and partial-state hydration to support sub-second agent transitions.
  - Built a **tiered memory architecture** — factual, episodic, and preference memories with vector compression, relevance decay, and attention-gated retrieval
  - Built a **context-engineering layer** employing dynamic context pruning, hierarchical chunk selection, and semantic routing to optimize LLM context windows for efficiency and precision.
  - Developed a **knowledge graph based RAG pipeline** that retrieves context from MBR/WBR documents using knowledge-graph expansions, custom embedding debiasing, query rewriting, and MMR-based reranking for business-grounded answers.
  - Deployed a **production async FastAPI orchestration** server with event-driven agents, Redis Streams, background coroutine workers, and partial-state hydration for sub-second agent transitions — on Kubernetes with Datadog APM, Langfuse LLM tracing, and full CoT logging forend-to-end reasoning transparency

- Built a **patented agentic image-to-video system** that enabled Expedia to become the first travel company to create and deploy AI-generated videos at scale, delivering $7M+ uplift in marketing performance across 10,000+ hotel properties. The system includes:
  - **Hallucination**-detection engine using multi-step Vision-LLM reasoning, object-mask tracking, structural-consistency checks, and temporal cross-attention validation to catch object loss, deformation, and scene anomalies.
  - **Camera-motion** detector using RAFT-style optical flow, motion-field smoothness metrics, and custom video-embedding encoders to flag jitter, instability, and prompt-misaligned trajectories.
  - **Autonomous regeneration** loop where a reasoning LLM performs iterative prompt optimization, shot-plan adjustments, and diffusion-model control tuning to meet expert creative standards.

- **Image-selection module** trained on 1M+ Expedia images using fine-tuned YOLOv8 for human/text filtering, CLIP dual-encoder fine-tuning for semantic ranking, and a LangGraph agentic workflow for frame scoring, quality gating, and VO-clip alignment.
- Integrated **fine-tuned video upscalers** (3D-UNet + latent-space SR transformers) with GAN-based perceptual loss, motion-aware temporal consistency, and artifact-suppression heads to enhance resolution and stabilize frames in regenerated outputs.

Aitado (Independent Venture) | Austin, TX
Co-founder & Lead AI Research Scientist | 08/2022 - Present

- **Built an AI learning co-pilot that transforms questions and documents into an interactive notebook environment**—combining explainer videos, Socratic audio dialogues, text-based Q&A, and adaptive quizzes for end-to-end learning.
  - Designed a **multi-agent architecture** coordinating RAG, code generation, visualization, evaluation, and TTS/voiceover agents to create coherent, pedagogically structured learning modules.
  - Led **back-end development and research** on **RL finetuning (GRPO/RLHF)**, multi-agent orchestration, and automated content-assembly pipelines powering Aitado's scalable education platform.

Expedia Group | Austin, TX
Machine Learning Scientist - Search & Discovery | 10/2022 - 06/2024

- Developed a **Search & Ranking Interpretation Framework** using latent-class discrete choice models, causal inference, and ranking-explainability diagnostics to evaluate how effectively Expedia's ML ranking systems satisfy real customer intent.
  - Built **latent-class choice models** via a custom **EM algorithm** to uncover heterogeneous customer segments, estimate segment-level utility functions, and model **substitution patterns** that drive conversion.
  - Applied **structural causal modeling** with instrumental variables, propensity weighting, and counterfactual estimators to isolate true causal drivers of bookings from position bias, supply bias, and popularity loops.
  - Designed a ranking-evaluation layer using **NDCG**, **ERR**, and other industry-standard ranking metrics, combined with counterfactual swap tests and Shapley/IG-style attributions to diagnose where algorithmic ordering diverged from customer utility.

- Developed and owned the **A/B testing infrastructure for Ranking ML models**, designing statistically rigorous and interpretable evaluation methodologies.
  - Built **ranking-specific experiment pipelines** using **interleaving methods** (team draft / probabilistic interleaving) to achieve faster, variance-reduced comparisons of ranking functions, enabling a **10× acceleration** in model iteration cycles.
  - Designed **interpretable ranking metrics** such as **qualified CTR**, position-normalized engagement, and intent-aligned utility gains to capture signal beyond raw clicks and provide deeper diagnostic insight.

Dream11 (India's Biggest Fantasy Sports Platform) | Mumbai, India
Associate Manager - Product Analytics | 12/2020 - 06/2021

- Improved app rating from **3.1 → 3.7** by leading product analytics and UX experimentation. Drove launch decisions for six major features and built funnel-conversion, marketing-effectiveness, and KPI alerting dashboards.

Mu Sigma | Bangalore, India
Lead Data Scientist | 10/2016 - 12/2020

Led a team of five delivering advanced ML solutions for a Fortune 100 CPG company and a Fortune 1 retailer, spanning predictive modeling, optimization, and large-scale analytics pipelines.

- Reduced churn **11% → 8%** using **survival models (Cox PH, Weibull AFT)** and **unsupervised clustering (k-means/GMM)** to identify high-risk segments and intervention levers.
- Improved lead conversion **26% → 29%** via **regularized logistic-regression scoring**, feature engineering, and uplift-based ranking of sales prospects.
- Increased manufacturing uptime **86% → 92%** by deploying **machine-failure detection models** using gradient boosting, sensor-level anomaly detection, and real-time scoring pipelines.
- Performed **marketing-effectiveness modeling** (MMM/ROI models) and **market-basket analysis** using association-rule mining and lift-based pattern discovery for strategic product placement.