SF Bay Area
206 954 3001
linkedin.com/in/abhishekpatnia

Abhishek Patnia

SKILLS

Transformer Fundamentals: Attention mechanisms, Positional Encoding types, MoE, KV Cache; **Agentic AI & LLM Adaptation**: agents, multi-agent architectures, function calling, customer-level optimization via memory, automated LLM-based evaluations, SFT, RFT; **Platforms & Tools**: OpenAI, Huggingface, Predibase, Pytorch, Python, Pyspark on Databricks; **Classical ML**: Deep Learning, Supervised Classification \w Noisy Data; **Working Knowledge:** GPU Fundamentals, CUDA, Triton, Model quantization and distillation, vLLM, AWS, Docker, K8s, Concurrent Programming

EXPERIENCE

Nubank, SF Bay Area — *Senior Staff ML Engineer/Scientist*

March 2024 - PRESENT

- Senior-most technical leader on the Capabilities team, integrating skills into Nubank AI using agents. For example: Pix Transfer Using AI.
- Defined the roadmap and best practices for scalable skill design and integration using agents.
- Partnered with product & vendor teams to reduce inference costs and accelerate launches.
- Leverage synthetic and historic datasets to iterate on agent behavior and de-risk new skill launches

Tinder, Los Angeles — *Staff ML Engineer/Scientist*

April 2019 - Jan 2023

- Founded and scaled the Trust & Safety ML team to 9 engineers & analysts.
- Designed multilingual text/image classification pipelines with Transformers & ConvNeXT.
- Established adversarial pre-processing pipelines and metadata embeddings to improve robustness.
- Built real-time inference pipelines (TFLite, TensorRT, Triton) deployed via Kubernetes.
- Partnered with product and risk analysts to invent KPIs and close coverage gaps across violations.

Amazon, Seattle — *Senior Applied Scientist*

April 2015 - April 2019

- Technical lead for the query understanding science team.
- Designed and launched Amazon's first RNN mapping query → shopping intent, powering shopping experiences at scale.
- Established deep learning training & deployment best practices, introducing GPU-based training infra with AWS Batch.
- Collaborated with infra teams to create a Python-based inference framework, ensuring seamless deployment of models across orgs.

Amazon, Seattle — *Senior Software Development Engineer*

June 2011 - March 2015

- Founding engineer of Kindle X-Ray, building algorithms to identify characters, passages, and images in books.
- Developed entity resolution algorithms to unify aliases (e.g., "Mr. Potter" = "Harry Potter")
- Created large-scale test harnesses: every algorithm change was validated across thousands of books before release.
- Built the Kindle N-gram Corpus (MapReduce + DynamoDB), foundational for topic identification across the Kindle library.

ADVISING

AI Startups, SF Bay Area — *Advisor*

MONTH 2023 - Current

Advisor to early and growth-stage startups through the Tola Capital Advisor Program, meeting with founders and leadership teams to shape their LLM strategy, architecture, and customer-facing features. Provide customer-centric perspectives on how generative AI can drive adoption, reduce costs, and differentiate products. Partnered with companies like Martian and Instawork, helping them design and scale production-ready LLM solutions. Guidance spans agent design and orchestration, fine-tuning open-weight models with LoRA, and inference optimization & evaluation.

EDUCATION

University of Southern California — *MS Computer Science*

2009 - 2011, Los Angeles

Pune Institute of Computer Technology — *B.E. Computer Science*

2003 - 2007, India