

# ABHIJIT BENDALE

+1-617-697-6247 ◇ abhijitbendale@gmail.com ◇ US Citizen ◇ Campbell, CA

Linkedin ◇ Google Scholar ◇ Website

## EXECUTIVE SUMMARY

---

Applied AI leader and hands-on engineer with 10+ years building and shipping large-scale ML systems across LLMs, multimodal AI, and computer vision. Delivered production AI platforms at global scale (400M+ users and devices).

Founder & CTO of Two Platforms (raised 20M seed, 10M ARR), where I built multilingual LLM and multimodal AI platforms deployed at national scale. Former CTO of Samsung STAR Labs, where I led NEON, a real-time generative digital human platform later integrated into Samsung Electronics (\$50M+ internal valuation). Previously at Samsung Electronics, architected on-device AI platforms deployed across 100M+ Galaxy devices.

Deep expertise in large-scale ML systems, distributed training, post-training, RLHF, evals, Agentic AI systems. Expertise in performance optimization across model, runtime, and hardware, with a strong focus on low-latency and high-throughput AI systems. Inventor on 30+ patents with peer-reviewed papers at leading AI conferences (3,500+ citations).

## EXPERTISE & TECHNICAL SKILLS

---

- **Languages & Core Tools:** C++, Python, TypeScript, PyTorch, CUDA/Triton, NCCL, MPI.
- **Foundation Models & Training:** LLM/VLM pre-training and fine-tuning; Transformers, MoE, CLIP-style multimodal models; distributed training with DeepSpeed, Ray, FSDP.
- **Post-Training & Alignment:** SFT, RLHF, preference optimization, reward modeling, LLM-as-judge evaluation, synthetic data, human-in-the-loop pipelines; TRL, PEFT.
- **Inference & Serving:** High-throughput LLM serving with vLLM, TGI, NVIDIA NIM, TorchServe; continuous batching, speculative decoding, KV-cache optimization, quantization.
- **Retrieval & Search Systems:** Billion-scale vector search, FAISS, Vertex Matching Engine, PgVector, RAG pipelines; multimodal retrieval across text, image, audio, and video.
- **AI Agents & Orchestration:** Tool use, function calling, multi-agent workflows, agent evaluation and reliability; LangGraph, LlamaIndex, Google ADK, MCP.
- **Cloud & Infrastructure:** Production AI platforms on GCP/AWS/Kubernetes; Vertex AI, GKE, SageMaker, EKS/ECS; distributed data pipelines, observability, reliability, and secure deployment.

## WORK EXPERIENCE

---

### Two Platforms

2021 – 2025

#### Founder & Chief Technology Officer

Campbell, CA

Founded Two Platforms - AI infrastructure & applications startup (\$20M seed from Jio Platforms & Naver Corp.) focused on multilingual LLMs, multimodal/voice AI, and enterprise applications for the APAC market. Scaled team of 20+ engineers across ML, platform and product.

- Built SUTRA, a multilingual AI platform spanning foundation models, semantic search, voice AI. Owned the full lifecycle from data and Transformer/MoE training to RLHF, evaluation, and deployment; operated 200+ GPU infrastructure for training and serving.
- Partnered with strategic customers including Jio Platforms to design and deploy AI applications for search, customer support, and workflow automation across cloud, VPC, and on-prem environments, contributing to \$10M+ ARR.
- Developed enterprise AI stack including RAG, vector search, agentic workflows, and voice agents, enabling rapid deployment of production AI systems across telecom, finance, healthcare, and retail.
- Deployed multilingual multimodal AI systems serving 400M+ users across text, image, audio, video, and documents. Built consumer products including Zappy (AI chat, 20M+ conversations) and AutoCameo (personalized video messaging), launched nationally in partnership with Jio Platforms.

**Samsung Technology & Advanced Research (STAR) Labs**  
*Chief Technology Officer*

Sept. 2020 – June 2021  
Mountain View, CA

As CTO of STAR Labs, led NEON, a real-time generative digital-human platform, from research prototype to commercialization; later integrated into Samsung Electronics at a \$50M+ internal valuation. Built and led a 40+ person multidisciplinary team across AI research, graphics, platform, and product, managing \$30M+ R&D funding over 2 years.

- Conceived and led NEON, an interactive avatar platform enabling temporally stable 4K photorealistic digital humans with sub-100ms latency, launched at CES 2020.
- Designed the core real-time multimodal generative AI stack, combining motion/audio-conditioned generation, identity personalization, neural rendering, optimized GPU inference, and streaming media delivery.
- Built production ML/media platform spanning data pipelines, model training/evaluation, low-latency serving, observability, SDKs, cloud deployment, and enterprise integration.
- Drove commercialization through NEON View, NEON Frame, and NEON Studio across interactive customer service, digital displays, mobile/web avatars, and enterprise media/content generation.
- Led APAC enterprise partnerships across media, finance, retail, and telecom; supported deployments and pilots with Samsung business units and external partners.
- Generated 10+ patents across core models, data pipelines, and system architecture; led research collaborations with MIT, Stanford, and the University of Washington.

**Samsung Research America**  
*Director / Principal Engineer / Staff Engineer*

Nov 2015 – Aug 2019  
Mountain View, CA

Led architecture and scaling of on-device AI and perception platforms deployed across 100M+ Samsung Galaxy devices. Built and managed a team of 60+ engineers and researchers (20+ PhDs) with a \$15M+/year annual R&D budget.

**Bixby Vision & On-Device AI**

- Built and scaled Bixby Vision, on-device AI deployed across 100M+ Galaxy devices, enabling visual search, translation, shopping, and scene understanding.
- Built Samsung's on-device AI platform from first principles, enabling ML deployment across Android, Tizen, Galaxy devices, Smart TVs, digital displays, and XR systems under strict latency, power, and memory constraints.
- Designed real-time camera and perception pipelines, processing live video frames for object detection, OCR, scene understanding, and AR experiences with low-latency on-device inference.
- Built platform SDK and APIs (Vision, AR, rendering) enabling partner applications and services to integrate with Bixby Vision, creating an extensible ecosystem.
- Led hardware–software co-design with Samsung Semiconductor (Exynos NPU), Qualcomm (Snapdragon), and Google (Android), optimizing inference via quantization, acceleration, and cross-platform deployment.
- Drove transition from applied research to production systems, defining architecture, evaluation, and deployment strategies for large-scale commercialization.

**AR & Mixed Reality Platforms**

- Led AR and Mixed Reality platform initiatives in collaboration with Microsoft (HoloLens) and Google (Android / ARCore), shaping hardware–software integration and platform architecture.
- Directed and contributed hands-on to Samsung's in-house AR/MR stack, building SLAM and sensor fusion systems using RGB, event-based vision (DVS), IMU, and depth sensors for real-time spatial mapping.
- Developed low-latency, power-efficient perception and rendering pipelines for XR, including patented approaches for event-driven vision and sensor fusion.

## RESEARCH EXPERIENCE

---

### **SRI International**

*PhD Intern*

June 2014 – September 2014

*Princeton, NJ*

- Built algorithms to automatically summarize and reconstruct sequential storylines from large-scale photo collections, modeling temporal structure across millions of images **DARPA ALADDIN**.

### **IDIAP Research Institute (EPFL)**

*Visiting Student (with Prof. Sebastien Marcel)*

October 2013 – January 2014

*Switzerland*

- Researched continual learning and online model adaptation for large-scale biometric recognition under distribution shift (EU/NSF Open Vision project).

### **MIT Media Lab**

*Research Scientist (with Prof. Ramesh Raskar)*

August 2010 – July 2011

*Cambridge, MA*

- Developed VisionBlocks, a social computer vision platform (Google App Engine Award), and conducted image super-resolution research for DARPA ARGUS-IS.

### **MIT McGovern Institute & Dept. of Brain & Cognitive Sciences**

*Visiting Student & Research Scientist (with Prof. James DiCarlo)*

July 2009 – July 2010

*Cambridge, MA*

- Developed deep learning and biologically inspired vision models for object recognition and pedestrian detection. Research supported by **DARPA NeoVision**.

## PUBLICATIONS & PATENTS

---

Selected list of publications (cited 3800+ times) is listed below:

1. A. Bendale *et.al.* "SUTRA: Scalable Multilingual Language Model Architecture" 2014 (*arxiv* 2405.06694)
2. A. Bendale, T. Boulton "Towards Open Set Deep Network" CVPR 2016 (oral, 2.5K+ citations)
3. A. Bendale, T. Boulton "Towards Open World Recognition" CVPR 2015 (oral, 900+ citations)
4. A. Bendale, T. Boulton "Reliable Posterior Probab. Estimation for Face Recognition" Biometrics W, CVPR 2014
5. A. Bendale, *et.al.*, R. Raskar "VisionBlocks: Social Computer Vision Platform" SocialCom 2011

Selected list of Patents (from 30+):

1. Generating digital avatar (US 11544886, 2023)
2. Real-time interactive digital agents (US 17686201, 2023)
3. CMOS-assisted inside-out DVS tracking (US 11381741, 2023)
4. Semi-dense depth estimation with DVS (US 11143879, 2021)
5. Semantic mapping for low-power AR ( US 10812711, 2020),

Full list is available via Google Scholar.

## EDUCATION

---

### **University of Colorado**

Ph.D. in Computer Science (Dissertation: Open World Recognition)

Advisor: Prof. Terrance Boulton

2011 – 2015

### **University of Colorado**

M.S. in Computer Science

2007 – 2009

### **University of Pune**

B.E. in Electronics and Telecommunications

2003 – 2007