Anjan Goswami, Ph.D.

San Francisco Bay Area, CA
VP of AI & Engineering — Agentic Systems
LLM Infrastructure, Applied Science

goswami.anjan@gmail.com
smartinfer.com
linkedin.com/in/goswamianjan

Executive Summary

AI Systems Executive & Builder: 20+ years leading applied AI, multimodal reasoning, agentic workflows, and search/ranking systems across Microsoft, Adobe, Salesforce, Walmart, and Amazon. Scaled global engineering & applied science organizations (80+) and shipped AI capabilities used by tens of millions of users.

Deep Technical Leadership: Led teams training and optimizing production ML models (distributed training, distillation, synthetic data, active learning, on-device models) and built advanced inference tooling including GPU-optimized serving prototypes and early kernel-generation workflows. Creator of **Thinker** (open-source inference router) and designer of the declarative LLM training–inference platform **Boson LLM v2**; explored CSP-based hybrid reasoning models.

Data-Centric AI Leadership: Led data platform development for training data acquisition, synthetic data generation, active learning, privacy-preserving pipelines, multilingual datasets, and labeling/annotation systems, improving model accuracy, robustness, and D-SAT outcomes across multiple product domains.

Search, Ranking & Retrieval: Designed and deployed large-scale search and recommender systems, including ML ranking pipelines, lexical+vector hybrid retrieval engines, semantic and image search models, and personalization algorithms. Built RAG systems for enterprise Q&A, conversational memory for agentic assistants, and synthetic-data & evaluation pipelines for LLM-based search and discovery.

Leadership & Talent Development: Recruited, mentored, and developed **well over a hundred** AI scientists & engineers across Microsoft, Adobe, Salesforce, Walmart, and Amazon; built new AI teams, integrated acquired startup groups, and established global centers of excellence. Extensive experience aligning research, infrastructure, and product organizations to deliver high-velocity execution and platform-scale impact.

Core Technical Competencies

AI Architecture: Agentic workflows, multimodal LLMs, hybrid dense+lexical retrieval, semantic/image search, vector search, RAG.

Model Development: Distributed training, distillation, active learning, compact/low-latency models, post-training evaluation of frontier LLMs.

Inference & Optimization: GPU-optimized serving, kernel-generation workflows, quantization, cost-efficient large-scale inference.

Data & Safety: Synthetic data systems, data-centric AI, evaluation harnesses, RLHF/RLAIF, red-teaming fundamentals, privacy-preserving pipelines.

Leadership: Large-scale org building, cross-functional alignment (Research/Infra/Product), technical due diligence, integrating acquired startup teams.

Professional Experience

Head of AI and Data, Microsoft PowerPoint (Office 365) *Mar 2024 – Present*

- **Copilot Strategy & Execution:** Defined the AI strategy for PowerPoint Copilot; architected the agentic orchestration layer powering slide generation, editing, summarization, Q&A, and rewrite features.
- **Business Impact:** Delivered Copilot capabilities driving **+20% engagement** and measurable productivity gains across global enterprise customers.
- **Scaling AI Org:** Scaled the applied science team from 1 → 16 and influenced cross-functional engineering groups (50+); established synthetic-data pipelines and automated evaluation harnesses to address data scarcity.
- **Safety & Infrastructure:** Collaborated with Microsoft Research on foundation-model alignment; optimized inference latency and cost through Azure GPU cluster tuning, caching strategies, and retrieval improvements.

- **Strategy & Competitive Intelligence:** Authored the internal AI strategy white paper and conducted deep competitive analysis of frontier LLMs, including benchmark-driven capability evaluations and model landscape assessments that informed product investment decisions.

Founder, Stealth Healthcare AI Startup *Sep 2023 – Feb 2024*

- Prototyped a GPT-3.5 powered application for patient education and medical literacy.
- Built full-stack MVP including vector-based retrieval (RAG) on medical journals.

Director of Machine Learning & Engineering, Adobe *Jul 2019 – May 2023*

- **Org Leadership:** Directed a global 60+ person ML organization delivering AI innovations for Acrobat, Sign, Stock, Creative Cloud, and Experience Cloud.
- **Generative & Document AI:** Shipped **Acrobat Assist** (RAG-based Q&A/Summarization) and improved document-structure recognition accuracy by **3**× using data-centric techniques.
- **Multimodal Search:** Launched CLIP-based semantic image search for Adobe Stock, significantly improving recall, precision, and asset discoverability.
- **Infrastructure Optimization:** Built a unified audience-discovery platform for Adobe Experience Cloud; optimized distributed clustering pipelines, reducing cloud infrastructure costs by **75% (4×)**.
- **Customer Engagement:** Led deep customer conversations to translate marketing use cases into technical requirements for the audience-discovery platform.

Director of Machine Learning & Engineering, Salesforce *Jul 2016 – Jul 2019*

- **Einstein Platform:** Led global applied science and engineering teams integrating AI into Service and Community Clouds; delivered ranking, personalization, and conversational AI features achieving **+20% engagement**.
- **Trust & Safety:** Architected recommender systems, feed-ranking algorithms, Q&A retrieval models, and spam-detection systems for large-scale community forums.
- **Governance & Compliance:** Partnered with security, legal, and privacy teams to establish GDPR and HIPAA-aligned data practices for enterprise deployments.
- **Customer Engagement:** Guided product teams in mapping customer challenges to ML-solvable technical problems, accelerating feature adoption and relevance quality.

Principal Consultant, SmartInfer LLC *Oct 2015 – Jun 2016*

- **Etsy:** Advised on search ranking and query understanding architecture.
- **Neurotrack:** Improved Alzheimer's diagnostic computer vision models (pupil tracking) by **20% accuracy**.

Director of Search Science, Walmart E-commerce *Jun 2014 – Oct 2015*

- Led 80+ engineers/scientists in a complete re-architecture of the commerce search engine.
- Built ML-based ranking, query understanding, and type-ahead systems, delivering **+23% revenue uplift** and **+17% sales conversion**.

Earlier Career: Elance-UpWork, eBay, Amazon A9, Microsoft *2005 – 2014*

- **Elance-UpWork (Director):** Built contractor recommendation systems increasing hiring rates by 4%.
- **eBay:** Transitioned Cassini search to ML-based ranking (+6% revenue); architected fashion image search.
- **Amazon A9:** Developed global ranking infrastructure and relevance evaluation systems.

Open Source & Applied Research

Thinker (Open Source): Built a unified LLM inference router with registry-driven model selection, schema-based request language (ThinkerQL), secure credential management, and full observability; supports OpenAI, Anthropic, and Ollama backends.

LLM Systems R&D: Designed early components of **Boson LLM v2**, a declarative framework for LLM training–inference orchestration; explored CSP-based hybrid reasoning (CSP-LM) for structured problem-solving.

NeurIPS 2025 Workshop: H. Kang, E. Bao, A. Goswami. *VLM-SlideEval: Evaluating VLMs on Structured Comprehension and Perturbation Sensitivity in PPT*. arXiv:2510.22045

Patents, Publications & Thought Leadership

- **Patents (10+):** Inventor on patents spanning search relevance, ranking models, personalization, and recommendation systems. Key patents include:
  - Service Agent Personal Recommender System — US 1020369
  - Image-based Popularity Prediction — US 20120303615
  - Query Classification for Improved Search Relevance — US 20120221557

  Full list: smartinfer.com/patents
- **Publications:** Peer-reviewed work in search optimization, document intelligence, evaluation methods, and LLM-based reasoning.
  - *Controlled Experiments for Decision-making in E-commerce Search.* IEEE Big Data, 2015
  - *Towards Optimization of E-Commerce Search and Discovery.* SIGIR ECOM, 2018

  Full list: smartinfer.com/publication/academic_publications.html
- **Leadership & Technical Essays:** Writings on AI systems architecture, reasoning models, applied science leadership, org design, evaluations, and agentic AI. Essays: smartinfer.substack.com
- **Portfolio:** smartinfer.com

Education

Ph.D., Computer Science – University of California, Davis

Executive Education – University of California, Berkeley (Haas School of Business)

M.S., Computer Science – The Ohio State University

M.Tech, Mechanical Engineering – Indian Institute of Technology (IIT) Kanpur