

Balasubramaniam Srinivasan

Senior Applied Scientist & Tech Lead · Agentic AI Science · Amazon AWS

Santa Clara, CA • Email • Website • LinkedIn • Google Scholar

Summary

Senior Applied Scientist and Tech Lead at Amazon AWS, leading the science roadmap for agentic AI AgentCore platform primitives on Amazon AgentCore: agent security and red-teaming, agent customization and optimization, agentic search, and agent evaluations. Work at the intersection of large language models, reinforcement learning, and symmetry aware representation learning, architecting production systems from first prototype to general availability across Amazon Bedrock and Amazon AgentCore at internet scale. Ph.D. in Computer Science from Purdue; 20+ publications and patents and 1,500+ citations at ICLR, NeurIPS, ICML, EMNLP, and *ACL, including multiple oral and spotlight papers.

Experience

Amazon AWS · Senior Applied Scientist & Tech Lead

Apr 2024 – Present

Agentic AI: Amazon AgentCore (current) · Amazon Bedrock (prior). Santa Clara, CA.

- *Agentic Customization and Optimization (AgentCore, Preview May 2026, Tech Lead).* Lead science for continuously improving the context supplied to production LLM agents. Mine OTEL traces from live agent runs to auto-generate and maintain skill files, tighten system prompts and tool descriptions, train RL policies for optimizing task correctness, cost, and other user defined evaluators on multi-step workflows. [Launch news.]
- *AWS Agent Registry and Agentic Security (AgentCore, Preview Mar 2026, Tech Lead).* Lead Science for AgentCore Registry that discovers, indexes, and governs *compound agentic systems* (heterogeneous combinations of LLM agents, MCP servers, tools, and skill files). Built a pre-deployment security scanner and *RL-driven offensive red-teaming* pipeline mapped to multiple agentic security taxonomies, as well as fine-tuned models for agentic search. **Adopted by 1000+ enterprise customers.**
- *AgentCore Agent Evaluations (GA Mar 2026).* Owned the *synthetic-data generation pipeline* for agent eval and the *meta-evaluation strategy* that grades the evaluator (judge calibration, eval quality drift, agreement with human raters). Contributed methodology for compound agents on task success, tool-use and trajectory correctness, and safety / policy adherence. **Adopted by 500+ enterprise customers.**
- *Bedrock Intelligent Prompt Routing (GA Apr 2025, Tech Lead).* Led Science for routing incoming user prompts across LLMs of different capabilities and costs, to maximize performance constrained on cost and other user defined metrics. Designed the lightweight routing model, calibrated quality predictors, the user-controlled quality / cost frontier, and the offline / online evaluation harness. Published as IPR (EMNLP 2025). **1M+ requests / week in production; 40%+ inference cost reduction at quality parity with the top model; <150 ms added routing latency.**
- *Bedrock Automatic Prompt Optimization (GA Apr 2025, Core Contributor).* Built automatic prompt component wise extraction and rewrite strategies and performed per task quality evaluation across Claude, Llama, Nova, Mistral, etc.
- **Leadership:** Set the multi-quarter science roadmap for the org and present executive narratives to Director / VP. Lead senior-scientist hiring loops as a member of the hiring committee. Mentored 15+ junior scientists, applied scientists, and Ph.D. interns over the last four years. Drive cross-org technical direction across AgentCore, Bedrock, and Kiro on shared evaluation standards, post-training infrastructure, and science direction.

Amazon AWS · Applied Scientist & Tech Lead

Jun 2022 – Apr 2024

AWS AI Research and Education, Santa Clara, CA.

- *AutoDoc, Amazon DataZone (GA, Tech Lead).* Led science on LLM-based documentation platform for enterprise data catalogs (summaries, column descriptions, and downstream-use-case generation) for tables *without access to column contents*. Schema-aware routing, iterative refinement, LLM-as-judge evaluation. **Handles tables with 1,000+ columns and collections of 100K+ tables across 25+ enterprise customer domains;** announced at AWS re:Invent 2023.
- *Business Name Generation (GA, Core Contributor)* Designed a small permutation-equivariant model that jointly generates business-friendly column names for a table; beat SOTA LLMs by 10+ absolute percentage points

Purdue University · Graduate Research Assistant (Ph.D.)

Aug 2018 – May 2022

Advisor: Prof. Bruno Ribeiro. Department of Computer Science.

- Dissertation: higher-order graph representation learning. Equivariant and permutation-invariant architectures for sets, graphs, and hypergraphs, with theoretical expressiveness results and practical algorithms for molecular property prediction, combinatorial tasks, and network analysis.

ARM · Software Engineer, Performance Modeling & Analysis

Jul 2015 – Jul 2016

Microarchitectural memory sub-system performance modeling for next-generation ARM Cortex processors; built a memory-subsystem benchmark suite used internally and by external silicon partners.

Education

Purdue University · Ph.D., Computer Science

Aug 2018 – May 2022

Thesis: *On Higher-Order Graph Representation Learning*. Advisor: Prof. Bruno Ribeiro. **GPA 4.00 / 4.00.**

University of California, San Diego · M.S., Computer Science

Sep 2016 – Jun 2018

GPA 3.97 / 4.00. UC San Diego Award for Excellence in Teaching (2018).

BITS Pilani · B.E. (Hons.), Electrical & Electronics Engineering

Aug 2011 – May 2015

GPA 8.97 / 10.0; Top 2% Merit Scholarship. B.E. thesis at the Indian Institute of Science (Advisor: Prof. S.K. Nandy):

Design of a runtime resource manager and a simulator for a massively parallel dynamically reconfigurable accelerator.

Selected Publications

BayesFlow: A Probability Inference Framework for Meta-Agent Assisted Workflow Generation. *EACL 2026*
B. Yuan, Y. Zhou, Z. Xu, K. Ramnath, A. Feng, **B. Srinivasan**.

MEDAL: Diffusion Language Model Inference with Monte Carlo Tree Search. *EACL 2026*
Z. Huang, K. Ramnath, Y. Chen, A. Feng, S. Woo, **B. Srinivasan**, Z. Xu, K. Zhou, S. Wang, H. Ding, L. L. Cheong.

IPR: Intelligent Prompt Routing with User-Controlled Quality / Cost Trade-offs. *EMNLP 2025*
A. Feng*, **B. Srinivasan***, et al. (Amazon AWS). Production system: 1M+ req / week, 40%+ cost reduction.

A Systematic Survey of Automatic Prompt Optimization Techniques. *EMNLP 2025*
K. Ramnath, et al. (incl. **B. Srinivasan**), Amazon AWS.

CoverICL: Selective Annotation for In-Context Learning via Active Graph Coverage. *EMNLP 2024*
C. Mavromatis, **B. Srinivasan**, Z. Shen, J. Zhang, H. Rangwala, C. Faloutsos, G. Karypis.

TabSyn: Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. *ICLR 2024 (Oral)*
H. Zhang, J. Zhang, **B. Srinivasan**, Z. Shen, X. Qin, C. Faloutsos, H. Rangwala, G. Karypis.

BioBridge: Bridging Biomedical Foundation Models via Knowledge Graphs. *ICLR 2024*
Z. Wang, Z. Wang, **B. Srinivasan**, V. N. Ioannidis, H. Rangwala, R. Anubhai.

OpenTab: Advancing Large Language Models as Open-domain Table Reasoners. *ICLR 2024*
K. Kong, J. Zhang, Z. Shen, **B. Srinivasan**, C. Lei, C. Faloutsos, H. Rangwala, G. Karypis.

HYTREL: Hypergraph-enhanced Tabular Data Representation Learning. *NeurIPS 2023 (Spotlight)*
P. Chen, S. Sarkar, L. Lausen, **B. Srinivasan**, S. Zha, R. Huang, G. Karypis.

NameGuess: Column Name Expansion for Tabular Data. *EMNLP 2023*
J. Zhang, Z. Shen, **B. Srinivasan**, S. Wang, H. Rangwala, G. Karypis.

ESAN: Equivariant Subgraph Aggregation Networks. *ICLR 2022 (Spotlight)*
B. Bevilacqua*, F. Frasca*, D. Lim*, **B. Srinivasan**, C. Cai, G. Balamurugan, H. Maron.

On the Equivalence between Positional Node Embeddings and Structural Graph Representations. *ICLR 2020*
B. Srinivasan, B. Ribeiro. (Lead author.)

Relational Pooling for Graph Representations. *ICML 2019*
R. Murphy, **B. Srinivasan**, V. Rao, B. Ribeiro.

Janossy Pooling: Learning Deep Permutation-Invariant Functions for Variable-Size Inputs. *ICLR 2019*
R. Murphy, **B. Srinivasan**, V. Rao, B. Ribeiro.

Selected Honors and Service

Academic: UC San Diego Award for Excellence in Teaching (2018); BITS Pilani Top 2% Merit Scholarship (2011–2015).

Area Chair: NeurIPS, COLM, ARR (2023–present). *Reviewer:* ICML, ICLR, AAAI, LoG (2020–present).

Technical Skills

Agentic AI & LLMs: LLM agents, tool use, function calling, RAG, multi-agent (A2A) systems, planning, skill files, MCP, browser and coding agents, OTEL agent observability, OWASP Agentic Top-10, agent red-teaming and evaluation (Agent-Bench, SWE-bench, τ -bench, GAIA, WebArena); frontier models (Claude, GPT, Llama, Nova, Mistral, Gemini, DeepSeek, Qwen).

Post-training & Alignment: SFT, DPO, PPO, GRPO, DAPO, RLHF / RLAIF, reward modeling, preference-data curation, prompt optimization, LoRA / QLoRA / PEFT, distillation, quantization, speculative decoding, constitutional / debate-style oversight; LLM-as-judge design; HELM, lm-eval-harness, AlpacaEval, MMLU, GPQA, AIME, HumanEval, FrontierMath, ARC-AGI; capability and safety evaluation for compound agentic systems.

Foundations & Systems: Transformer / MoE, graph and hypergraph neural networks, equivariant / permutation-invariant architectures, diffusion (continuous / discrete), long-context and SSMs, dense / sparse / hybrid retrieval. PyTorch, JAX, HF Transformers, DeepSpeed, Megatron-LM, TRL, vLLM, TensorRT-LLM, SGLang, FlashAttention, Ray, DGL, PyG; DDP / FSDP / Tensor / Pipeline parallelism; AWS (AgentCore, Bedrock, SageMaker), FAISS; Python, C++, SQL.