Vivek V. Datla, Ph.D.

Sr. Tech Lead, Data Science (AI/NLP) Cambridge, MA • 22 Algonquin Trl, Ashland, MA 01721 901-634-3709 •
vivekvarmadatla@gmail.com • **LinkedIn** • **Portfolio**

Executive Summary

Applied AI leader with **16+ years** spanning research and production delivery across **Information Retrieval (IR)**, **Retrieval-Augmented Generation (RAG)**, **Question Answering**, **Knowledge Graphs**, and **Trustworthy LLMs**. Built and scaled **agent-assist RAG** platforms supporting **21k+ agents** across bank and card LoBs; repeatedly convert research into impact (**>10 technology transfers** at Philips). Inventor on **15 granted** & **35+ filed** patents; **30+ peer-reviewed papers**; **2000+ citations**. Proven track record leading cross-functional teams to ship **robust, high-availability NLP systems** in regulated domains.

Core Strengths

- **IR/RAG Architecture:** Hybrid retrieval (BM25 + dense + SPLADE), **cross-encoder re-ranking**, late-fusion, user-preference boosting, **entropy/nucleus thresholding**, dynamic index routing, **document chunking & windowing**, query rewriting & intent routing.
- **Trust & Safety for LLMs:** Abstention, refusal sensitivity vs. utility trade-offs, **semantic entropy**, **MARS-style meaning-aware scoring**, DetectGPT-style detectors, confidence calibration, provenance tracing for **pretrain vs. context** attribution.
- **Scalable ML Delivery:** Model/product roadmaps, experiment platforms, evaluation harnesses, **guardrails**, CI/CD for models, **observability (drift, latency, quality)**, on-call runbooks.
- **Leadership:** Hiring/mentoring, stakeholder alignment, value stream ownership (~$2M), program management across research → product.

Technical Stack

Languages: Python, SQL, Java, JavaScript, Bash, C, MATLAB **DL/ML:** PyTorch, TensorFlow, JAX (intro), Hugging Face Transformers/PEFT, Sentence-Transformers, scikit-learn, XGBoost **LLMs/Serving:** Llama-2/3, Mistral, GPT-4/4o/4.1, T5/Flan, **vLLM**, **Text Generation Inference (TGI)**, **TensorRT-LLM**, **Triton Inference Server Fine-Tuning:** LoRA/QLoRA, Adapters/IA3, P-tuning, Prefix/Prompt/Instruction tuning; **mixed-precision**, gradient/ZeRO sharding **Optimization:** Quantization (8-/4-bit), **FlashAttention**, speculative decoding, KV-cache management **RAG & IR:** Faiss, Milvus, pgvector, Elasticsearch/OpenSearch, **ColBERT**, SPLADE; monoT5/monoBERT/CE-rerankers; query rewriting, hyde/self-ask, **long-context chunking Graphs:** PyKEEN (RotatE/QuatE/PairRE/HousE), NetworkX, Neo4j **MLOps/Platforms:** MLflow, Weights & Biases, Kubeflow, Airflow, **Docker**, **Kubernetes**, **Ray/Dask**, GitHub Actions/CI **Data/ETL:** Pandas, NumPy, Apache Arrow, Spark, Kafka, Hadoop, RabbitMQ **Monitoring:** Prometheus, Grafana, ELK; **human-in-the-loop** annotation (Label Studio, Prodigy) **Cloud/Infra:** AWS (EC2, EKS, S3, SageMaker), Azure, GCP; serverless patterns

Selected Impact & Achievements

- **Capital One Agent-Assist (IR/RAG):** Led the IR team to a **state-of-the-art RAG** platform for call-center agents; deployed **cross-encoder re-rankers**, **hybrid retrieval**, and **ranked-fusion** with **entropy-gated answerability**, improving top-k precision and reducing handle-time (HT) and AHT variance.
- **Trustworthy Responses:** Shipped abstention & uncertainty pipelines (semantic entropy + nucleus thresholds) that reduced unsafe generations while maintaining task utility; introduced **user-preference re-ranking** and **dynamic profile routing**.
- **Tech Transfer (Philips):** Drove **10+** research-to-product transfers (clinical de-identification, knowledge-graph-assisted diagnosis, DSP assets, ICON semantic search).
- **IP & Publications: 15 granted** patents, **35+ filed**, **64+ invention disclosures**; **30+ publications** (NAACL, COLING, AAAI, WWW, BHI, MLHC, TREC).
- **Awards: Circle of Excellence (2025)**—Capital One's highest honor; **CIO Elite (2023)**; **TechX (2023)**.

Professional Experience

Capital One — Enterprise Data Science, Cambridge, MA Sr. Manager, Data Science (AI & NLP) • *2024–Present*
Manager, Data Science (AI & NLP) • *2022–2024*

- Lead **IR/RAG** for enterprise **Agent-Assist**; scale to thousands of agents across LoBs.
- Built CE/monoT5 re-rankers, hybrid BM25+dense+SPLADE retrieval, **rank-fusion**, **nucleus/entropy gating**, **user preference boosting**, and **context de-duplication**; owned evaluation harnesses (latency, precision@k, groundedness).
- Established **guardrails** (policy prompts, tool gating, abstention) and **calibration** (temperature scaling, MC-dropout ensembles, semantic-entropy).
- Managed end-to-end delivery: data pipelines, training, offline/online eval, A/Bs, **observability & SLOs**, incident playbooks.
- Mentored researchers/engineers; aligned with product, design, risk, and compliance.

Philips Research North America, Cambridge, MA Value Stream Manager & Technology Lead, AI for PMS • *2021–2022*

- Owned AI/NLP for Post-Market Surveillance (~**$2M** scope); led 6+ engineers/researchers; delivered cross-cluster products for PMS & Q&R.
- Delivered knowledge-graph-assisted workflows, clinical NER, and complaint triage models; defined roadmaps and stakeholder alignment.

Senior Scientist • *2017–2021*

- Co-architected a **swarm-based platform** for deploying DL models at scale.
- Built **interpretability-aware QA**, **VQA-Med** systems, and **clinical diagnosis** using KGs; participated in **TREC** (2016–2017).
- Drove IP strategy: 20+ invention disclosures; multiple transfers to business.

Scientist • *2015–2017*

- Led KG-based clinical QA projects; contributed to ImageCLEF/medical captioning;
- Three **TREC'16** tasks; publications at NAACL, COLING, AAAI, **WWW**.

Pacific Northwest National Laboratory(PNNL) — Post-Doctoral Research Associate, Richland, WA • *2014–2015*

- Streaming graph search (Idaho Bailiff initiative), large-scale Bayesian net querying for cyber; co-authored/maintained public releases.
- Contributed to **NOUS** KG construction; DTRA **Chiron**; released **MATEX** ML at scale.

Verified Person Inc., Memphis, TN — Software Engineer • *2008–2009*

- Built national-scale criminal-data warehouse and ETL engines; integrated Bugzilla/SugarCRM.

Academic Research (University of Memphis; ORNL) • *2006–2014*

- Unsupervised structure learning (pattern mining, VOMMs), SRL from induced grammars, LIWC-based forecasting; EHR quality & affect; HMM tutoring interaction models; cloud workflow prediction; network defense via game theory.

Selected Projects & Research Highlights

- **RAG Correctness Estimation (COLING'25, Industry Track):** Co-authored method to estimate RAG answer correctness in production settings; integrated CE ranking signals with confidence features and retrieval stats to improve **groundedness** and **abstention**.
- **Meaning-Aware Uncertainty:** Benchmarked **semantic entropy** and **meaning-aware scoring** for LLM uncertainty; combined with reranker margins and retrieval coverage to minimize harmful outputs without over-refusal.
- **KG-Assisted Clinical Diagnosis:** Joint reasoning over notes, codes, & KGs; deployed de-identification pipelines and semantic search for radiology.
- **VQA-Med & Medical Captioning:** Attention-based captioning with concept mapping; retrieval that returns **text + image** evidence.

Education

Ph.D., Computer Science, University of Memphis, TN — *2009–2014* **M.S., Computer Science**, University of Memphis, TN — *2006–2008* **B.Tech., Electronics & Communication Engg.**, JNTU, India — *2001–2005*

Awards & Honors (Selected)

- **2025** — *Circle of Excellence* (Capital One, highest honor) — Agent-Assist
- **2023** — *CIO Elite* & *TechX* Awards — Agent-Assist
- **2017** — *Breakthrough Innovation Award*, HealthWorks Acceleration
- Best Paper: **ANSS'10**, Overall Best: **SpringSim'10**; multiple student travel awards (IISSO); PNNL *Science as Art* (2015)

Patents (Selected — 15 Granted; 35+ Filed)

- **Hierarchical Self-Attention for Machine Comprehension** (16/916,697)
- **MUDRA: Multi-Domain Real-Time QA System** (16/342,635)
- **Neural Text Simplification via Semantic Alignment** (16/430,676)
- **Open-Domain Real-Time QA** (16/430,788)
- **Question Generation with Fact-based Attentive RNNs** (16/334,135)
- **Clinical Decision Support via Deep RL** (16/491,489)
- **Multimodal Deep Memory Networks for Diagnostic Inferencing** (16/330,174)
- **CRF-based Span Prediction for Fine MRC** (16/681,945)
- **Condensed Memory Networks** (15/707,550)

Complete list available on request.

Publications (Selected)

- **Zhang, Datla, et al.** *An Automatic Method to Estimate Correctness of RAG.* **COLING'25 (Industry Track)**.
- **Ghaeini, Hasan, Datla, et al.** DR-BiLSTM for NLI. **NAACL'18**.
- **Prakash, Hasan, Lee, Datla, et al.** Neural Paraphrase Generation with Stacked Residual LSTMs. **COLING'16**.
- **Hasan, Datla, et al.** Clinical Paraphrase Generation with Attention. **ClinicalNLP@COLING'16**.
- **Lee, Qadir, Hasan, Datla, et al.** Adverse Drug Event Detection in Tweets with Semi-Supervised CNNs. **WWW'17**.
- **Datla, et al.** Open-Domain Real-Time QA (TREC '16/'17).
- **IJCLA'14.** Linguistic features predict truthfulness of short political statements.

Google Scholar: 2200+ citations.

Professional Service

- **TPC/Reviewer:** NAACL (Clinical NLP), ACL, MLHC, AAAI-FLAIRS, IEEE ICDM, SECURECOMM, IPDPS workshops, ICCC, etc.
- **Organizer/Challenge Contributor:** TREC (2016–2017), ImageCLEF (VQA-Med '19).

Keywords for Recruiters/ATS

LLM • RAG • Information Retrieval • Cross-Encoder Re-ranking • Hybrid Search • ColBERT • SPLADE • MonoT5 • Confidence Calibration • Semantic Entropy • Meaning-Aware Scoring • DetectGPT • Hallucination Mitigation • Guardrails • Retrieval Fusion • Query Rewriting • Vector Search (Faiss/Milvus/pgvector) • OpenSearch/Elasticsearch • PEFT (LoRA/QLoRA) • Quantization (8/4-bit) • vLLM • TensorRT-LLM • Triton • FlashAttention • PyTorch • Hugging Face • MLflow/W&B • Kubernetes/Docker • Ray/Dask • Knowledge Graphs (PyKEEN/Neo4j) • Healthcare NLP • De-identification • A/B Testing • Observability • SLOs

---

Optional Addenda (On Request)

- Full patent & invention disclosure list
- Detailed project metrics (precision@k, AHT deltas, coverage/groundedness)
- Teaching/mentoring, invited talks, and outreach
- Extended bibliography (journals, workshops, posters)