Premith Kumar Chilukuri

premithchilukuri.com chpk cpremithkumar ch.premith.k@gmail.com +1 (540) 558-3599

Work Experience

ABB
March 2025 – Present

AI Engineer — *United States*

- **Architected an enterprise agent orchestration platform** that translates natural-language requests into multi-agent automations, using supervisor-specialist delegation, parallel DAG execution, stateless container workers, and Kubernetes scaling.
- **Built and deployed multimodal RAG at industrial scale:** a **400K+ document** corpus, **419K-entity** knowledge graph, and **1.66M relations**, integrating PostgreSQL, Apache AGE, pgvector, HyDE query expansion, and relevance grading. Combined graph relationships, vector similarity, and structured retrieval to connect industrial evidence across sources.
- **Engineered reliable long-running agents** with persisted human approval checkpoints, context compression, token budgets, dynamic skill loading, sandboxed Python execution, and end-to-end tool-call tracing.
- **Integrated enterprise tools and data** through MCP, OAuth connectors, and schema-aware access; delivered SAP analytics and versioned document generation through FastAPI, Docker, and multi-provider LLM inference.
- **Built a dynamic skills framework** that lets agents author, version, and selectively load task-specific capabilities; enforced structured input/output contracts and sandboxed code execution to support reusable, controlled enterprise automation.
- **Developed cost-conscious, multi-provider inference** and API delivery for enterprise agents, supporting OpenAI-compatible access and streamed responses through FastAPI while keeping execution containerized and observable.

Engineering Design Tools Analyst Intern — *Jan 2025 – March 2025*

- Built conversational motor search across **500+ product specifications**, reducing search time by **70%**; developed LLM/RAG evaluation across **500+ queries** using **10+ metrics**.
- Engineered **LLM-based product filtering and recommendations** with attribute extraction and parameter matching; assessed fine-tuned Qwen, LLaMA, and LLaVA models with automated and human-aligned evaluation.

Robotics Intern — *May 2024 – Dec 2024*

- Developed a **6D pose-estimation pipeline** reporting **5 mm MSSD at 0.6 s inference**; investigated diffusion/NeRF reconstruction, language-guided detection, and vision-language approaches to 2D-to-3D CAD.
- Prototyped **vision-language 2D-to-3D CAD modeling** with GPT-4o and Qwen2-VL; built language-guided object isolation with **Phi-3 and SAM2** for pose-estimation workflows.

Virginia Tech
June 2023 – May 2024

Graduate Research Assistant — *Blacksburg, VA, USA*

- **Co-authored manufacturing-guided 3D diffusion research** using a 3D U-Net and ResNet surrogate to steer $\mathbf{64^3}$ **voxel** porous anode designs. Integrated a learned minimum-feature-size surrogate into the later denoising stages to guide generation toward manufacturing constraints. The study used **2,735 training/validation designs** and reported normalized minimum feature size of **0.63 versus 0.54** for unguided generation.
- Researched **single-view NeRF and diffusion-based 3D reconstruction**; applied **LoRA adaptation** to SDXL and explored language-model-guided generation for zero-shot 3D object workflows, including quantized language models and pose-estimation applications.
- Built **LangGraph multi-agent workflows** with ReAct tool use and PostgreSQL retrieval; developed **hybrid/HyDE RAG** for business and process optimization, with a profile-reported **55% improvement in context retrieval accuracy**.
- Developed the **AVA language-model chatbot**, combining LoRA topic fine-tuning and retrieval-augmented generation to support business queries with current evidence.
- Created an augmented-reality visualization of **186 neural-network results across 85 UCR datasets** for the World Map of AI research project.

Supervue AI
Jan 2020 – Jan 2023

Lead Machine Learning Engineer — *India*

- **Led the AI application team** and developed **Supervue FR V2.0**, a ViT-based facial-recognition API adopted by **50+ clients**; reduced API and application deployment costs by **10%**.
- Built **real-time zero-shot person detection and tracking** using a CNN–Transformer encoder–decoder with TensorRT/ONNX inference, tracking **1.3M people** with **99.1% uptime**.
- Designed **auto-scaling microservices** for facial recognition, ASR, and monitoring apps serving **30K users**; developed distributed OCR that improved parsing performance by **50%** for **100M+ documents**.
- Fine-tuned **GPT-2 with Megatron-LM and DeepSpeed** for multi-task chatbot NLU, improving intent-classification score by **10%** over Bi-LSTM/GRU models. Developed social-distance monitoring using patented **one-shot object-flow regression** (**Indian patent 601573**; application 202041034063).

Samsung R&D Institute
Oct 2021 – Aug 2022

Research Collaborator (Volunteer, part-time) — *Remote, India*

- Developed lightweight, multi-threaded **mobile video inpainting at 31 FPS**, with feature transformation, search, and caching; achieved a reported **15% reduction in distortion artifacts**.
- Contributed the **nN-TFT** edge-deployment approach associated with **Indian application 202241043300**.

Zebi Data
June 2019 – Dec 2019

Machine Learning Engineer — *India*

- Helped build a **blockchain-assisted hotel check-in system** with biometric verification and ID-card OCR APIs hosted on AWS, reducing check-in time by **40%**.

Education

Virginia Tech — **Jan 2023 – Dec 2024**
Master's in Computer Science (CGPA: 4.0) — *Blacksburg, Virginia, USA*

BML Munjal University — **Aug 2015 – June 2019**
B.Tech., Computer Science and Engineering (CGPA: 8.8/10) — *Delhi, India*

Technical Skills

Programming: Python, C++, Go, Java, SQL, Shell/Bash
Production AI: FastAPI, Docker, Kubernetes, AWS, Azure, CI/CD, TensorRT, ONNX
Agentic AI & Retrieval: MCP, LangGraph, LangChain, multi-agent orchestration, context engineering, multimodal RAG, PostgreSQL, Apache AGE, pgvector, Redis, Kafka
Research & Modeling: PyTorch, TensorFlow, diffusion models, LoRA, LLM inference and evaluation, quantization, NeRF, 3D reconstruction, OpenCV

Peer Review, Judging & Professional Service

- **Journal reviewer:** Reviewed **20+ research articles** for IEEE Access, Engineering Applications of Artificial Intelligence, and Springer Nature's Signal, Image and Video Processing.
- **TechConnect World Innovation Conference and Expo 2026:** Served on the review panel for the AI for Grid Resilience and Security Challenge innovation spotlight sessions.
- **ABB US Accelerator, 2026:** Served as a key member of the jury panel, contributing technical judgment on emerging innovations.
- **INFORMS QSR Challenge, 2024:** Remotely developed a platform to host the university hackathon and automatically evaluate participant submissions for the Quality, Statistics and Reliability challenge.
- **Virginia Tech outreach, 2023:** Presented Concepts of Manufacturing & Industrial Internet for the Pathways for Future Engineers initiative.

Publications

- **Power Circuit AI: Designing Power Electronic Circuits for Motor Drives with Generative Artificial Intelligence**. IEEE IDCD, 2026.
- **MiRAGE: A Multiagent Framework for Generating Multimodal Multihop Question-Answer Dataset for RAG Evaluation**. arXiv · 2601.15487, 2026. Preprint. **Equal-contribution co-author.**
- **High-Quality Dataset-Sharing and Trade Based on A Performance-Oriented Directed Graph Neural Network**. IEEE Transactions on Automation Science and Engineering, 2025.
- **High-Quality Privacy Preserving Text Data Sharing for Manufacturing Machine Learning Data Market**. IISE Annual Conference and Expo, 2025.
- **Generative Design for Manufacturing: Integrating Generation with Optimization Using a Guided Voxel Diffusion Model**. OpenReview · manuscript record, 2024. Research manuscript.
- **Generating Optimized 3D Designs for Manufacturing Using a Guided Voxel Diffusion Model**. ASME MSEC, 2024.
- **Attention Based Multi-Patched 3D-CNNs with Hybrid Fusion Architecture for Reducing False Positives during Lung Nodule Detection**. Journal of Computer and Communications, 2021.
- **l, r-Stitch Unit: Encoder-Decoder-CNN Based Image-Mosaicing Mechanism for Stitching Non-Homogeneous Image Sequences**. IEEE Access, 2021.
- **AMMDAS: Multi-Modular Generative Masks Processing Architecture With Adaptive Wide Field-of-View Modeling Strategy**. IEEE Access, 2020.

Patents & Invention Records

- **A Novel Architecture To Perform Fast (Real-Time) Video Inpainting**. Application **202241043300**. *Filed*
- **ARCHITECT: Automated Robust CAD-based Hierarchical Intelligent Engineering, Neural Evaluation, Testing, and Assembly Toolkit**. **P250309US01**. *Filed*
- **Agentic Architecture for Automated Assembly of Parametric CAD Models Using Advanced Reasoning Paradigms with Generative AI**. **P250244US01**. *Filed*
- **Agentic Swarm-Driven Autonomous Data Science System for Real-Time Industrial IoT Analytics**. **P250351US01**. *Filed*
- **Autonomous Generation and Advisory of 3D CAD Assemblies from Multimodal Engineering Data**. **P250513US01**. *Filed*
- **Dynamic Auto-calibration System for Spatial Depth Analysis at Crowded Places**. Application **202041034063**. *Granted: 7 Sep 2026; Patent No. 601573*
- **Generative Adversarial Agentic Architecture for Automated Parametric CAD Modeling and Multi-Physics Simulation Optimization Using Advanced Reasoning Paradigms**. **P240832US01**. *Filed*
- **Hypothetical Product Embedding for Industrial Devices**. **P250510US01**. *Filed*
- **Methods for Building Domain Knowledge Databases from Industrial Data Sources**. **P250509US01**. *Filed*
- **Multi-Stage Visual Language Model Enabled Agentic Redline Interpretation and Parametric CAD System for Automated Design and Manufacturing Integration**. **P250273US01**. *Filed*
- **OPTIC: Orchestrated Perceptual Thought Iteration for Cognitive visual analysis**. **P240831US01**. *Filed*
- **VLM-based Agentic Process for Creation of Parametric 3D CAD Models**. **P240846US01**. *Filed*