Sarvesh Baskar

+91 8754583044 | baskarsarvesh@gmail.com | linkedin.com/in/sarvesh-b | github.com/Sarvesh-369 | sarvesh-369.github.io

Education

Birla Institute of Technology and Science (BITS) Pilani — Goa, India
Dual Degree: Bachelor of Engineering in Computer Science & Master of Science in Physics — *Sep. 2020 – July 2025*

Experience

AI Consultant — Aug. 2025 – Present
Avyott — *Remote*

- Designed and deployed multimodal RAG chatbots using LlamaIndex, supporting grounded Q&A over enterprise text, image, and table documents for active clients at IPCA and Taj.
- Built Model Context Protocol (MCP) servers in Python/Node.js to integrate chat workflows with ticketing systems and intent clarification, reducing task completion time by over 30%.

AI Research Assistant — Jul. 2025 – Present
University of Maryland, College Park — *Remote*

- Engineered zero-token video question answering in SmolVLM2 by predicting LoRA weights via a perceiver hypernetwork, reducing query-time token load 1500× and query TTFT by 6×–80×.
- Conducted large-scale OpenQwen2VL training and evaluation using Slurm cluster workflows to mitigate object hallucination and enhance visual grounding, publishing at ACL Findings, ECCV, and EMNLP.

AI Research Assistant — Aug. 2024 – Jul. 2025
University of Maryland, Baltimore County — *Maryland, USA*

- Developed PlanForge, an Architect-Builder-Runner agent system generating verified Python solver scripts from natural language descriptions, reaching 100% success rate on NaturalPlan.
- Led AI modeling and evaluation for low-level LLVM-IR code deobfuscation (85% cyclomatic complexity reduction) and conversational persona-gap resolution (+42% human A/B preference gain).

AI / Software Intern — May 2024 – Jul. 2024
Techisy — *Remote*

- Built a duplicate bill detection system using OCR, OpenAI embeddings, and CLIP to match visual and textual duplicates, achieving 95% identification accuracy.
- Developed a RAG-based candidate-to-job matching pipeline using LangChain, Ollama, and FAISS; deployed interactive Streamlit prototypes, reducing manual screening efforts by 40%.

Selected Projects

Video2LoRA | *SmolVLM2, LoRA, Hypernetworks, PyTorch, Python* — Mar. 2026 – Present

- Designed a perceiver hypernetwork that outputs LoRA adapters in a single forward pass, eliminating visual tokens from query context and enabling composition of independent video chunk adapters in rank space.

PlanForge | *LLM Agents, Planning, Code Generation, Verification* — May 2025 – Present

- Developed code-generation pipelines that convert natural language constraints into execution-grounded solvers, utilizing iterative code-test-fix loops for programmatic verification.

Publications

4 Peer-Reviewed Publications (including *ACL Findings 2026*, *NAACL 2025 SRW*) & **3 Papers Under Review** (including *ECCV 2026*, *EMNLP 2026*) focusing on multimodal reasoning, agentic planning, and LLM alignment.

Technical Skills

Languages: Python, C/C++, Bash, SQL, LaTeX, HTML/CSS
Frameworks & Libraries: PyTorch, Transformers, vLLM, LangChain, LlamaIndex, DSPy, RAG, VLMs, OpenCV, TensorFlow, HuggingFace
Developer Tools & Systems: Git, Docker, Slurm, Redis, ChromaDB, FAISS, Ollama, MCP, Linux