Ram Tewari

+1 845-650-3626 | tewari10@purdue.edu | linkedin.com/in/ram-tewari | github.com/ram-tewari

Education

Purdue University | Junior (3rd Year) — West Lafayette, IN

B.S. Industrial Engineering | Minor in Computer Engineering — *Expected May 2028*

- Relevant Coursework: Data Structures & Algorithm Analysis, Digital System Design (Verilog), Linear Algebra, Advanced C Programming

Experience

Machine Learning Engineer Intern — Jun. 2026 – Aug. 2026

Ultraviolette Automotive — *Bengaluru, India*

- Built an edge-to-cloud AI coaching pilot for in-person EV showrooms, transforming recorded customer interactions into evidence-linked feedback; evaluated conversations from 2 showrooms.
- Benchmarked on-device faster-whisper (INT8) transcription and pyannote diarization; achieved **96.5%** speaker attribution on 50 hand-annotated recordings of real showroom speech.
- Implemented variant-aware retrieval over versioned motorcycle specifications and a deterministic claim verifier with numeric tolerance bounds; code, not the LLM, determined factual correctness.
- Designed a two-pass, turn-level evaluation pipeline linking transcript evidence to product-accuracy findings and coaching signals for customer discovery, objection handling, and test-ride follow-up.
- Built a 100-session model evaluation harness; in a 40-session A/B test, top-k retrieval cut prompt tokens **24%** while preserving detection of 11/11 planted factual errors.

Software Intern — Jun. 2025 – Aug. 2025

Magna International – New Mobility — *Lowell, MA*

- Developed a React teleoperation interface for the Arrival WeMo industrial warehouse robot, integrating live multi-camera feeds, keyboard/joystick navigation, and lift controls for material movement.
- Engineered a Python/FastAPI backend for SSH-based motion commands; measured command latencies of **58 ms (keyboard)** and **62 ms (joystick)**.
- Implemented omnidirectional driving and operator-initiated autonomous-to-manual control transitions, supporting remote intervention during warehouse material-handling operations.

Selected Projects

PHAROS – Context & Memory Infrastructure for Coding Agents — 2025 – Present

Python, FastAPI, MCP

- Built a repository-aware code-intelligence service with Tree-sitter AST indexing of symbols, imports, and call relationships; verified cited source against commit-pinned Git revisions.
- Implemented dense, SPLADE sparse, and lexical retrieval with reciprocal-rank fusion; delivered source-linked code and dependency context to AI coding agents through REST and MCP APIs.
- Evaluated 101 labeled queries across 7 repositories; improved held-out **nDCG@10 from 0.394 to 0.697** and **0.463 to 0.726** across two test sets versus prior PHAROS configurations.
- Separated indexing and embedding workloads from the cloud API through asynchronous workers, with GitHub-referenced source storage to limit duplicated repository data.
- Prototyped Git-history-derived agent memory; in an 80-task evaluation, improved file-selection precision from **63.4% to 77.5%**, while auditing unsupported memory claims.

R&Bot – 33M-Parameter R&B Transformer | *PyTorch, NLP* — 2026

- Implemented and trained a **33M-parameter** autoregressive Transformer from scratch over 4 epochs using an approximately 50M-token mixed-domain corpus.
- Built a staged text-training pipeline using general narrative text and poetry before genre-focused training on alternative R&B, neo-soul, and related lyrics.
- Introduced section-conditioning tokens (`[VERSE]`, `[CHORUS]`, `[BRIDGE]`) to generate structured song sections rather than unconstrained continuous text.
- Implemented configurable decoding and repetition controls; analyzed long-range coherence and recurring-phrase failures to guide subsequent training experiments.

Technical Skills

Languages: Python, C++, C, SQL, JavaScript, Java, SystemVerilog
ML / AI: PyTorch, Transformers, Whisper, pyannote, Retrieval Evaluation, Vector Search, RAG, Edge AI
Systems / Tools: FastAPI, PostgreSQL, pgvector, Redis, Docker, Git, Linux, Tree-sitter, NVIDIA Jetson