

Vidushi Anand

GitHub | LinkedIn | Portfolio | vidushianand09@gmail.com | Gurugram, India

SKILLS

Domains: LLM Evals, LLM-as-a-Judge, AI Agents, Observability, Guardrails, Red-Teaming, RAG, Agent Memory

Frameworks: Python, PyTorch, Hugging Face Transformers, LangGraph, LangChain, CrewAI, DSPy, LiteLLM, LiveKit, Pipecat

Tools: OpenTelemetry, Redis, FastAPI, Cog, pytest, FAISS, ChromaDB, SQLite, Ollama, Git, CI/CD, SQL, Java, C

EXPERIENCE

AI Engineer - Noveum.ai

May 2026 – Present

Python, OpenTelemetry, Redis, CrewAI, LiveKit, Pipecat, spaCy, Gemini, Ollama, pytest

San Francisco

Build the evaluation, tracing and guardrail layers that customers run on top of their AI agents:

- **NovaGuard** - an **LLM guardrail engine** that blocks over-budget and rate-limited calls *before* they reach the provider, using **atomic Redis budget checks** shared across processes; covered by **165+ tests**.
- **noveum-trace** - an open-source, **OpenTelemetry-compatible tracing SDK** for agents. Shipped **CrewAI auto-instrumentation** (96 tests) and privacy by default: **PII redaction** (spaCy + regex), HTTPS-only transport, no credential logging.
- **NovaEval scorer reliability** - failed judge calls were averaged in as low scores, so one scorer reported **4.38 (33% pass)** instead of **6.69 (48%)**, under-reporting agent quality by **35%**; a timed-out judge in a 4-judge panel also counted as a vote. Separated errors from real scores in every aggregation path.
- **Voice-agent scorers** - judges saw one reply without the call around it, so resumed answers after an interruption read as repetition (**5 of 50** calls, 5+ points each). Added conversation history and an **uncertain** status when two judge runs disagree; validated on **50 production calls** with a rerun as noise control.
- **Judgecheck** (*novaeval judge-check*) - tests an **LLM judge** or NovaEval scorer on **metamorphic variants** of known-correct answers (reworded, hedged, fact removed, wrong value, refusal), reporting false accept/reject rates with **95% Wilson intervals**, no human labels; matched the prototype on **819 answers**.

AI Engineer Intern - APi.market (MagicAPI Inc | Noveum.ai)

Jun 2025 – Oct 2025

Python, FastAPI, Cog, FAISS, ChromaDB

San Francisco

- **NovaEval** - built the **RAG evaluation suite** from scratch: **30+ automated metrics** for retrieval, generation and end-to-end quality (faithfulness, hallucination detection, citation quality, answer completeness).
- Designed a **stage-by-stage RAG pipeline evaluator** (query, retrieval, reranking, generation) that pinpoints which stage failed.

PROJECTS

Judgecheck: How Reliable Are LLM Judges? | Gemini 2.5 Flash / Flash-Lite, Llama 3, Ollama

Oct 2026

- Tested **5 judge setups** (~118 questions × 8 answer variants); Mem's grader passed **82% of refusals** on personal questions.
- Fixing the prompt cut refusals passed from **39% to 7%** but made Flash-Lite fail **56%** of correct, hedged answers; a targeted rubric fixed both (**4%** refusals, **0%** hedges), while "be strict" failed 100% of hedges.
- Llama 3 8B said "correct" to **487 of 520** answers; a 1B judge looked perfect on a one-sided test, so **test both directions**.

Adversarial Attacks on LLM Judges | PyTorch, Transformers, Ollama, Gemini [GitHub]

Oct 2026

- **White-box (GCG)**: a 20-token gradient-optimized suffix passed **107–108 of 108** held-out refusals on a **Qwen3-4B** judge across 3 seeds, yet transferred to **0/108** on Llama 3 and Qwen3-4B-2507.
- **Black-box**: compared random, greedy and **LLM-attacker search** on equal query budgets; a readable LLM-written suffix got **24%** of refusals past Llama 3 8B and **transferred to Gemini** (5% to 18%).

Mem: Persistent Memory Layer for LLM Agents | Python, DSPy, SQLite, Gemini 2.5 Flash-Lite [GitHub]

Sep 2026

- Extracts **atomic, dated facts** from notes, coding sessions and git commits; a nightly **consolidation** pass supersedes stale facts while keeping history, and secrets are redacted before storage.
- Evaluated on **LongMemEval** with per-failure attribution (never stored / not retrieved / misanswered): most losses happened at **write time, not retrieval**, and the grader itself was unreliable, which led to Judgecheck.

VoxLoad: Cognitive Load from Speech | Python, Whisper, spaCy, ffmpeg [GitHub]

Mar 2026

- Multi-agent pipeline (ffmpeg, **Whisper** word timestamps, then pause, filler-word and speech-rate agents and an aggregator) that scores **cognitive load** against each speaker's own baseline.
- Pilot with 3 speakers: raw signals varied **4–5×** across speakers, so **per-speaker z-scores** were essential; fillers tracked load for low-filler speakers and **pauses** for high-filler ones, while speech rate was unreliable.

EDUCATION

Dronacharya College of Engineering, Gurugram

B.Tech in Computer Science (AI & ML)

Expected 2027

CGPA: 8.8/10

AWARDS AND LEADERSHIP

- **Smart India Hackathon 2025, Grand Finalist**: edge IoT controller for power-grid phase imbalance.
- **Kaggle Notebooks Expert**: top 3% (1,569 of 60K+).
- **Co-Lead, Deviators Club** (Aug 2025 – May 2026): organized Debug Decrypt 2.0, a 3-day DSA contest with **110+ teams**.