K Satya Sai Nischal

Hyderabad, India | +91-89776-12004 | satyasainischal@gmail.com | linkedin.com/in/satyasainischal | github.com/cattolatte
Machine Learning / NLP — from-scratch transformers, retrieval & RAG, shipped to PyPI: zenith-nlp, polaris-nlp, meridian-rag

Education

Geetanjali College of Engineering and Technology (GCET) **Sep 2024 – May 2028**
B.Tech., Computer Science *Hyderabad, India*

- **Coursework**: Data Structures & Algorithms, Operating Systems, Database Systems, Computer Networks, Machine Learning
- **IEEE Computer Society**: **Secretary**, Student Branch Chapter; competed in IEEEXtreme 19.0, IEEE's global 24-hour contest (19,000+ participants)

Projects

Meridian — From-Scratch Grounded RAG Engine | *Python, PyTorch, Polaris, Zenith, FastAPI, Docker* | GitHub | PyPI **2026 – Present**

- Built a grounded RAG engine over PubMed literature with **no third-party pre-trained models**: from-scratch BPE tokenizer, InfoNCE dense bi-encoder, **IVF/HNSW ANN indexes (no FAISS)**, cross-encoder reranker, and NLI verifier — every answer **cited, verified, or refused**, with calibrated abstention at **80.1% coverage, 0.000 error rate**.
- Trained the from-scratch NLI verifier **41.5%** → **78.3%** on SNLI dev (942K pairs, MLM pretraining over 1.15M sentences) by isolating three bottlenecks: tokenizer mismatch (**+6.0 pts**), data scale (**+22.0 pts**), and learning rate at identical architecture (**+17.6 pts**).
- Diagnosed a reranker overfitting collapse and restored **R@5 0.029** → **0.983** via rank fusion; built **HNSW** reaching **recall@10 0.996 at 0.262 ms**, faster than exact search. Shipped **v1.0** with **254 tests (95.7% coverage)**, seed-averaged ablations, and **published negative results**.

Zenith — From-Scratch Generative NLP Library | *Python, PyTorch, FastAPI, Hydra, MLflow* | GitHub | PyPI **2025 – Present**

- Built and published to PyPI a decoder-only transformer LM library written directly on PyTorch tensor primitives, implementing the modern LLM stack: causal self-attention, RoPE, RMSNorm, SwiGLU, KV-cache, weight tying. **Matches the nanoGPT baseline at 2.08 bits/char**, trained to parity in ∼**10 min on a laptop GPU**.
- Implemented **speculative decoding** with KV-cache rollback (output provably identical to greedy, **up to 3.7**× **fewer target forward passes**) and from-scratch **weight-only int8 quantization** (**4**× smaller weights, output unchanged). Shipped **v1.0** with **103-test** CI and **22 tagged releases**.

Polaris — From-Scratch Encoder NLP Platform | *Python, PyTorch, FastAPI, Docker* | GitHub | PyPI **2026 – Present**

- Architected the encoder-side complement to Zenith: an end-to-end NLP platform published to PyPI across **17 versioned releases** — from-scratch **BPE tokenizer**, transformer encoder with **no** `nn.Transformer`, training, evaluation, and Dockerized FastAPI serving.
- Implemented **BERT-style masked-language-model pretraining** from scratch. In a controlled ablation against an identical randomly-initialized baseline, pretraining lifted first-epoch validation accuracy **73.6%** → **81.0%** and reached a higher best validation accuracy (**86.4%**) in fewer epochs.

ANTIDOTE — AI Agent Knowledge Recall | **1st Place, IndiaCodex '26 Hackathon** | *TypeScript, Aiken, React* | Demo | GitHub **Jul 2026**

- Won the Masumi "Monetize AI Agents" track (Cardano; team of 4): forged data is traced to every agent that ingested it, directly or downstream, and revoked — with decontamination and auditing exposed as paid, hireable services across a **5-agent MIP-003 economy**.
- Authored **3 Aiken (Plutus V3) validators** enforcing quarantine in consensus logic; built SHA-256 sharding with **Merkle non-membership proofs** for provable deletion, plus an LLM pipeline with multi-provider failover (Groq → Gemini → deterministic). **82 tests** (68 unit + 14 on-chain).

Experience

GenAI Intern **Jun 2024 – Jul 2024**
IIT Hyderabad & RemarkSkill *Hyderabad, India*

- Built a multimodal audiobook pipeline (Whisper ASR → GPT-4 → ElevenLabs TTS) deployed to **500+ students**; fine-tuned **GPT-2 (124M)** on ∼**500K tokens**, cutting validation perplexity **28.5** → **24.2**, served behind a FastAPI endpoint.
- Cut hallucination rate **42%** → **27%** on an internal 6-scenario evaluation via structured prompt engineering; added a GitHub Actions CI suite (**90+ tests**), cutting iteration time **45** → **18 min**.

Technical Skills

Languages	Python, C++, TypeScript, SQL, Bash
Machine Learning	PyTorch, Transformers (from scratch), self-supervised pretraining (MLM/CLM), LoRA, mixed precision (AMP), distributed training (DDP)
NLP & Retrieval	RAG, BM25, dense retrieval (bi-/cross-encoders), ANN search (HNSW, IVF), contrastive learning (InfoNCE), NLI, BPE tokenization, speculative decoding, KV-cache, int8 quantization, evaluation & ablation design
MLOps & Infra	FastAPI, Docker, GitHub Actions (CI/CD), MLflow, Hydra, pytest, mypy, PyPI packaging, AWS (EC2, S3, Lambda), Linux, PostgreSQL, Redis