

Aditya Pratap Singh

📍 New Delhi, India ✉ adityasingh.devs@gmail.com ☎ +91 9868599920 in [LinkedIn](#) 📄 [GitHub](#)

Summary

Built a 25M-parameter autoregressive transformer from scratch in PyTorch, tokenizer to training loop, no external wrappers, then mapped all 64 attention heads into a taxonomy of functional circuit types, independently cross-validated through hidden-state geometry. First-author preprint on the findings, currently under submission to arXiv. Final-year AI/Data Science undergrad who also designs and benchmarks ML systems architecture, seeking research and applied ML internships in mechanistic interpretability, LLM reasoning, and the broader NLP/LLM systems space.

Skills

- **Languages:** Python
- **Interpretability & Research:** Mechanistic Interpretability (attention circuit analysis, hidden-state probing, PyTorch Hooks), Attention Visualization, Autoregressive Transformer Design, Model Evaluation & Benchmarking, Reproducible Experiment Design
- **ML & LLM Engineering:** PyTorch, HuggingFace Transformers, LLM Fine-tuning (LoRA/QLoRA), RAG, Qwen, LLaMA
- **Systems & Inference:** FastAPI, vLLM, Docker, Asynchronous Pipelines, REST APIs, Groq
- **Data & Tooling:** NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, PyYAML, Git, Weights & Biases, Jupyter, Streamlit, Postman

Publications

At the Threshold: Tracing Computational Hierarchy through Convergent Evidence in a Small Transformer 2026
[Zenodo \(DOI\)](#) · *under submission to arXiv*

- Addressed an open question in mechanistic interpretability: whether attention routing and hidden-state representations converge on one account of transformer computation. Manually classified all 64 attention heads across 8 layers into a taxonomy of 6 circuit types in a 25M-parameter transformer, formalized through a novel Behavioral Complexity Index alongside Shannon entropy, and independently cross-corroborated with hidden-state geometry. Localized a Layer 3 computational threshold and critically weighed it against competing explanations from recent literature (compression valleys, rogue dimensions, massive activations) instead of treating convergence as confirmation.

Projects

NanoLens: Transformer Mechanistic Interpretability Research Toolkit 2026
[GitHub](#) — *Pretrained Weights Released*

- Implemented a fully modular 25M-parameter decoder-only transformer from scratch in PyTorch (8 layers, 8 heads, 512-dim) — tokenizer, attention, residual stream, training loop, and CLI, all without wrappers; YAML-driven architecture scales 10M–150M+ params on standard hardware, theoretically to 1B+.
- Trained at the character level on a 4.4M-character literary corpus, chosen for long-range narrative structure to let linguistic patterns emerge directly from raw character sequences; trained with cosine LR decay and AdamW (separated decay/no-decay groups) on a 90:10 split, achieving val loss 1.1144.
- Engineered dual-mechanism inspection infrastructure capturing all 64 attention heads and 8 hidden-state tensors in one forward pass — pull-based attention maps (1,8,T,T) and push-based PyTorch hooks (1,T,512) per block, without modifying model code — generating a full circuit and hidden-state visualization suite from one command.
- Publicly released the checkpoint, configs, and inspection tooling for one-command reproduction of the full pipeline; findings published as a first-author preprint (see Publications), toolkit built for extension — e.g. tracing LoRA/QLoRA circuit shifts.

Mimir v0: Incident based LLM Reasoning & Hallucination Analysis 2026
[GitHub](#)

- Found structured diagnostic reasoning has a non-monotonic effect on hallucination: cut hallucination from 33% (freeform) to 0% under low ambiguity, but raised it to 50% under high ambiguity, improving reasoning quality (1.58 vs 1.17/2) at the cost of accuracy (16.7% vs 25.0%); key failure modes were premature hypothesis fixation and over-constrained reasoning.
- Investigated across 12 controlled trials (4 scenarios × 2 prompting conditions) via a reproducible pipeline with Qwen 2.5-3B and deterministic multi-run evaluation; scoped for depth over scale, with scale-up planned in a RAG-extension phase.
- Constructed a synthetic incident dataset from real-world outage patterns (GitHub, Cloudflare, GitLab) across varying causal complexity and ambiguity; defined a 4-metric framework measuring accuracy, hallucination rate, evidence grounding, and reasoning quality.

Tesseract v1.2: Text-to-3D ML Inference System 2025
[GitHub](#) — *Towards AI benchmarking study published in Towards AI*

- Designed and benchmarked a stateless, asynchronous inference system for text-to-3D generation using OpenAI's Shape-E, served via FastAPI with concurrent non-blocking requests, CLI interfaces, and device-aware GPU/CPU fallback.
- Engineered a modular architecture with explicit separation of concerns, YAML-driven config, system-level logging, Docker/Kubernetes compatibility, and mesh post-processing with OBJ/PLY/GLB export.
- Benchmarked system behaviour, finding a **330× CPU-to-GPU slowdown** (6.06s–1662s) and inference accounting for **81–83%** of end-to-end latency; latency scaled linearly with batch size and sampling steps, with prompt complexity having negligible impact.
- Authored a technical analysis of the system's design and benchmarking trade-offs, published in *Towards AI* following editorial review.

Experience

Open-Source Engineering — OpenVINO (Intel) Dec 2024 – Feb 2025

- Restored broken metric-logging pipelines across OpenVINO's evaluation infrastructure by implementing external logger integrations (Weights & Biases, MLflow, TensorBoard, Comet), re-enabling experiment tracking and reproducibility for distributed training workflows.
- Extended Datamaro transformation pipelines to enable memory-efficient frame selection across multi-source video datasets (>100k frames), navigating a large-scale production codebase and iterating through multiple rounds of maintainer review.

Advait (AI Community), VIPS-TC Feb 2025 – Present
President

- Scaled college AI community from ~50 to 520+ members; drove technical direction via cell-based execution, shipping 32+ projects, 16 research breakdowns, and 16 technical blogs biweekly, while personally delivering talks on production ML systems and LLM internals.

Education

Vivekananda Institute of Professional Studies – Technical Campus (GGSIPU) 2023 – 2027
B.Tech in Artificial Intelligence and Data Science CGPA: 8.37