Toheed Akhtar

toheedak333@gmail.com | +91 ********** | toheedakhtar.github.io
github.com/toheedakhtar | linkedin.com/in/toheedakhtar01

Research Interests

AI observability and monitoring; interpretability; epistemic uncertainty and model reliability; agent behavior and failure modes.

Research Experience

Metacognitive Monitoring and Control in LLM Agents **2026–Present**
Lossfunk · Ongoing experimental research

- Built two complementary evaluations: **candidate-specific monitoring** on GPQA Diamond and **prospective regulation** on ProofWriter, distinguishing answer quality estimates from their use in action selection.
- On a 100-question, single-seed GPQA run, GLM-5.3-Flash's self-aligned scores reached **0.713 within-question AUC**; top-score selection was 72.8% versus 67.9% after shuffling scores. Neutral verification selected 74.2%, so self-reference showed no clear advantage.
- Designed **five shared-candidate conditions** (baseline, reflection, monitor-only, metacognitive control, matched-action control), with a fixed execute/hint/decline utility policy, precomputed hint outcomes, and paired inference on 300 held-out ProofWriter items.
- GLM's control utility was **0.876** versus 0.860 baseline, 0.844 matched control, and **0.937 reflection**; despite primary-monitor AUROC 0.914, it missed all 21 oracle-best declines.
- In an exploratory frozen-answer Llama-3.1-8B comparison, a **ConfTuner** confidence adapter improved policy utility by 0.019 (95% CI [0.009, 0.030]) without changing candidate answers.

Is Factual Correctness Represented in LLMs? **2026**
Cross-dataset probing study · Manuscript in preparation · [See draft]

- Probed **five 7–8B model variants** across **TruthfulQA, HaluEval, and FEVER**, extracting final-token residual, attention-output, MLP-output, and **Jacobian-centroid** features across layers; fit linear/RBF SVM readouts.
- Best within-dataset AUROC reached **0.90–1.00**, but representative mean cross-dataset AUROC was only **0.31–0.63** at IID-selected layers. Transfer was asymmetric: individual dataset pairs worked, but no tested setting worked consistently across all six directed pairs.
- Used **direction alignment, centroid radii, and joint PCA** to examine benchmark-dependent geometry; nonlinear probes, instruction-tuning comparisons, mixed-source training, and normalization did not consistently resolve the transfer gap.
- Fine-tuned explicit correctness-verdict models: AUROC on models' own generated answers was **0.67–0.78** versus 0.93–0.98 on human-written answers to the same questions. The negative claim concerns *tested readouts*, not the existence of all possible truth representations.

Small Language Models Hallucinate Knowing Something's Off **Jan 2026**
Independent interpretability study · Neel Nanda's MATS Exploration task · Public write-up

- Contrasted factual and fictional prompts in Llama-3.2-1B and Qwen2.5-1.5B (base/instruct), measuring a heuristic **uncertainty-token probability mass** from intermediate residual streams.
- Used **logit lens, head ablation, attention maps, and forward/reverse activation patching**; observed model-dependent uncertainty-mass peaks (late in Llama, mid-layer in Qwen) and heads attending to fictional tokens.
- Compared base and instruction-tuned variants to identify layers and attention heads associated with the metric before intervening on their activations.
- Targeted patching changed the metric by **∼14%** in Llama-3.2-1B-Instruct and **∼7%** in Qwen2.5-1.5B-Instruct; small prompt coverage and the proxy metric limit broader conclusions.

Higher-Order Alleys **2026**

Exploratory mechanistic interpretability study · Notes and experiments

- Used **J-Lens/J-space** interventions on **Qwen3.6-27B** to identify causal evaluator directions; matched SELF-versus-OTHER prompts showed these were *not* reliably specific to self-evaluation.
- Designed **Process-Sensitive Replay**: teacher-forced identical visible answers while manipulating hidden answer support through two mechanisms, with same-layer norm-matched random and reset controls.
- On **eight held-out items**, a frozen layer-42 direction responded to both structured process degradations beyond random controls, disappeared after reset, and covaried with later confidence; evidence is exploratory, not a proven higher-order mechanism.
- Candidate-to-judgment mediation remains **unresolved**: BF16 coordinate patches introduced off-direction changes, while an FP32 continuation changed downstream judgments even without the patch.

Work Experience

Lossfunk · Bengaluru **Mar 2026–Present**

Research Fellow (Aug 2026–Present); Research Intern (Mar–Jun 2026)

- Investigating how AI systems represent uncertainty, monitor their own performance, and regulate behavior; developing testable, architecture-agnostic hypotheses about metacognitive functions.
- Conducted cross-benchmark correctness probing with linear and nonlinear models, layer sweeps, and representational geometry analyses; characterized distribution-shift failures for a manuscript in preparation.
- Engineered a reproducible LLM-agent research scaffold with five experimental conditions, local and API model adapters, deterministic seeds, prompt hashes, immutable run artifacts, and explicit inference budgets.
- Built analysis and reporting pipelines for candidate-specific monitoring, task-paired comparisons, controller interventions, failure decomposition, and token/latency/cost accounting.

Technical Work

- **LLM from Scratch:** Implemented GPT-style transformer components, tokenization, pretraining and fine-tuning in PyTorch; trained a 127M-parameter model on CPU.
- **$5 LLM (open-source contribution):** Replaced Newton–Schulz iteration with Polar Express in the MuON optimizer; contribution acknowledged on the project's marathon leaderboard.

Technical Skills

Languages: Python, C/C++, JavaScript/TypeScript, SQL.

Machine Learning: PyTorch, Hugging Face Transformers and Datasets, scikit-learn, PEFT/LoRA, TRL, SentenceTransformers.

Agents and APIs: LangGraph, LangChain, CrewAI, MCP, A2A, OpenAI-compatible APIs, FastAPI.

Research Tooling: JLens, custom evaluation harnesses, NumPy, pandas, Matplotlib, Seaborn, Jupyter.

Developer Tools: Docker, Git, Linux, Codex (proficient, skeptical use with independent verification).

Education

B.Tech., Computer Science & Engineering, Rajasthan Technical University, Kota **2022–2026**
CGPA: 8.0/10