

OM CHIMURKAR

+91 9860233911 | om.chimurkar2024@nst.rishihood.edu.in | Rishihood University

 GitHub |  LinkedIn |  Google Scholar |  ResearchGate |  ORCID

Education

Newton School of Technology, Rishihood University

B.Tech in Artificial Intelligence

Expected May 2028

GPA: 8.0 / 10.0

Research & Publications

DKV: Anchor + Low-Rank Differential KV-Cache Compression

Independent Research

IEEE TETCI (Under Review) | *Technical Report & Open-Source Runtime* | *Zenodo: 10.5281/zenodo.21539110*

- Architected a training-free KV-cache compression runtime combining exact anchors, rank-32 joint K|V truncated SVD deltas, and multi-signal exact residual tokens to preserve verbatim recall.
- Achieved $1.44\times$ – $2.25\times$ memory reduction and $1.72\times$ faster prefill scaling at 64k context on Apple M3 (8.6 GB) over memory-optimized dense baselines.
- Engineered MLX and CUDA/Triton backends with autotuned fused decode launches, asynchronous H2D prefetching, in-place residual scattering, and runtime reconstruction.
- Validated through long-context ablations that exact residual keys prevent needle-recall collapse across 4k–64k contexts while maintaining distributional fidelity with $\Delta\text{PPL} = 0.00\%$.

CRBench: A Method-Agnostic, Resource-Aware Evaluation Framework for Long-Context LLMs

Independent Research

ASPLOS 2027 (September Review Cycle, Under Review) | *Initial Research Release* | *Open Source* | *Zenodo: 10.5281/zenodo.22085120* | *v0.2.0*

- Designed a query-level, method-agnostic evaluation framework that measures contextual capability retention relative to an identical dense FP16/BF16 reference while explicitly accounting for memory and system resources.
- Introduced normalized quality retention Q , memory efficiency R_{mem} , and resource utility score $S_{\text{res}} = \alpha Q + (1 - \alpha)R_{\text{mem}}$, with a decoupled system track for TTFT, prefill latency, and decode throughput.
- Built an extensible two-track evaluation framework supporting custom context adapters in under 50 lines of code, analytical and physical memory accounting, transparent failure semantics, and reproducible score recomputation.
- Preliminary validation on Qwen2.5-1.5B across five contextual tasks and ten configurations placed DKV High at 98.6% normalized quality with 63.8% memory savings, while DKV Mid achieved 94.2% quality with 73.4% memory savings.
- The current release is an initial research release, and the full version has been submitted to ASPLOS 2027 in the September review cycle.

SpiralState: Emotional State and Reasoning Effort in Language Models

Co-authored with Dhruv Ramani

ICLR 2027 (Submitted)

- Investigated how affective prompt framing influences reasoning effort in language models through controlled evaluations across mathematical and logic tasks.
- Characterized changes in reasoning length and inference behavior across emotional, neutral, and unframed conditions, and tested whether the effect generalizes across distinct reasoning models.
- Analyzed the systems implications of variable reasoning effort, including inference-time compute, memory consumption, and generation workload under different prompt conditions.
- Built a controlled inference evaluation pipeline to measure reasoning workload and resource usage across prompt conditions, linking model behavior to serving-time cost.
- Probed internal model representations to investigate whether changes in reasoning effort are associated with identifiable latent state directions.

Deterministic AGI Benchmarking Under Constructed Cognitive Load

Independent Research

Published Preprint | *Zenodo: 10.5281/zenodo.19652443*

- Designed a deterministic benchmark using constructed-language rule-learning tasks to evaluate reasoning without judge-model scoring, reducing judge inflation and memorization effects.
- Developed 116 original tasks spanning Legal, Science, Social Norms, and Math/Logic domains with deterministic answer verification.
- Evaluated seven frontier models and observed a 23.2 percentage-point performance spread, demonstrating substantial variation in reasoning reliability under constructed cognitive load.

An Ablation Study of Retrieval Strategies for Stock-News RAG Systems

Independent Research

Published Preprint | *Zenodo: 10.5281/zenodo.19086005*

- Conducted controlled empirical ablations comparing semantic, hybrid (BM25 + vector), and Maximal Marginal Relevance (MMR) retrieval on stock-news QA benchmarks.
- Demonstrated that hybrid retrieval improves retrieval coverage over semantic-only search, while MMR increases contextual diversity without proportional grounding gains.
- Established retrieval diversity and factual grounding as distinct, independently optimizable concerns in enterprise RAG architecture design.

Technical Skills

LLM & ML Systems PyTorch, Hugging Face Transformers, LLM Inference Optimization, KV-Cache Compression, RAG, Model Evaluation, Quantization, Low-Rank Factorization (SVD)

GPU & Systems CUDA, Triton, Nsight Systems, Nsight Compute, GPU Profiling, Memory Optimization, Runtime Benchmarking, Linux

Software & Infrastructure Python, FastAPI, Git, Vector Databases, MLX, Reproducible Experimentation

Research Methods

Empirical Ablation Studies | Benchmark Design | Long-Context Evaluation | Memory Accounting | GPU Performance Profiling | Inference Benchmarking | Reproducible Experiments | Model Evaluation

Awards & Experience

Cognizance '25 (SnapSyntax) — IIT Roorkee **1st Place / Zonals Winner** *May 2025*

- Secured first place in the SnapSyntax technical challenge at IIT Roorkee’s annual technical festival.

Google — Gemini API Developer Contest **Contestant** *May–August 2024*

- Built and deployed an asynchronous LLM application integrating the Gemini API with custom prompt pipelines under live competition constraints.

Honors & Certifications

Anthropic: AI Fluency Framework & Foundations **AWS:** Fundamentals of Machine Learning & Artificial Intelligence
Credential IDs available upon request