AYUSHI CHADHA

Dehradun, India | +91 87557 00158 | ayushichadha48@gmail.com | LinkedIn: www.linkedin.com/in/ayushi-chadha-ai | GitHub: https://github.com/Ayushichadha | X: https://x.com/AyushiChadha24

SUMMARY

AI Research Engineer focused on retrieval-augmented generation, reasoning architectures, and transformer/CUDA optimisation. Currently prototyping subgoal-augmented hierarchical reasoning models and scalable, fluid-intelligence systems for personal AI.

TECHNICAL SKILLS

- **Languages & Frameworks:** Python, C, C++, CUDA, PyTorch, Triton, Selenium
- **LLM, Reasoning & RAG:** LangChain, LlamaIndex, Hugging Face Transformers, RAPTOR, CLIP, Diffusion Models, hierarchical reasoning models (HRM)
- **Optimisation:** Mixed-Precision, Gradient Checkpointing, Memory-Efficient Attention, Kernel Fusion
- **Infrastructure:** AWS (S3, Lambda, EKS), Docker, Vector DBs (Pinecone, Chroma), Git, LangSmith
- **Research Interests:** Reasoning architectures (HRM, cognitive core), Transformer Efficiency, Multimodal Agents, Self-Reflective RAG

RESEARCH EXPERIENCE

Hierarchical Reasoning Model with Subgoal Augmentation Independent Research 2025 – Present

- Designed and implemented subgoal-augmented extension to a 27M-parameter neuro-inspired Hierarchical Reasoning Model (HRM) with fast/slow recurrent pathways, feudal-style latent subgoal head, and adaptive gating.
- Achieved +8% accuracy and 5–6% loss reduction on ARC-mini (identical parameter/data budget vs baseline) with markedly improved training stability and internal reasoning trace legibility.
- Directionally aligned with Karpathy's “small cognitive core” vision and emerging test-time compute paradigms; explicit focus on adaptive, non-static objectives for fluid intelligence in small, personal devices.

PROFESSIONAL EXPERIENCE

Software Developer – Propero Consulting Pune, India Feb 2023 – Dec 2024

- Architected “ShopiBot”: end-to-end domain-specific Retrieval-Augmented Generation (RAG) pipeline (dense retrieval, RAPTOR, corrective and adaptive QA). Cut latency from 53 s→20 s (~62%) via streaming inference, vector-DB sharding, and prompt/embedding caching, increasing top-k precision by 28% and reducing hallucinations by 35%.
- Implemented Active-RAG agentic workflows: leveraged knowledge graphs and LangChain planners for multi-tool reasoning, self-reflection, and autonomous task execution.
- Built multimodal image-generation system: CLIP-guided prompt augmentation and Stable Diffusion fine-tuned with DreamBooth; added controllable diffusion/GAN-style image-to-image translation modules.
- Deployed anomaly-aware checkout-flow tester: combined DOM-tree pattern mining with RL-style feedback, catching 92% regressions pre-release; integrated LangSmith telemetry for continuous evaluation.
- Unified GTmetrix, PSI, and Pingdom APIs in a LangChain workflow; auto-generated PDF audits, saving ~75% manual QA hours; scaled NLP micro-service for review sentiment (2M+ reviews/day, <200 ms p95).

Software Engineering Intern – Propero Consulting Aug 2022 – Jan 2023

- Built Page-Speed Analyzer MVP and automated testing modules using Robocorp RPA and LangChain.
- Earned Robocorp certifications; laid foundation for later agentic-AI migration of QA pipelines.

RESEARCH & PERSONAL PROJECTS

- Andrej Karpathy-Inspired LLM Stack: re-implemented transformer kernels (matmul, GELU, layernorm) in pure C/CUDA; experimenting with Triton fused kernels and memory-efficient attention for billion-parameter models.
- Autograd & GPT From Scratch: authored minimal autograd engine, trained bigram/trigram models, replicated GPT-2-style language model and custom tokenizer end-to-end.

EDUCATION

B.Tech in Electrical Engineering – College of Technology, G.B. Pant University of Agriculture & Technology, Pantnagar 2017 – 2021

ADDITIONAL INFORMATION

- Selected for Unify Contributor Program (Y-Combinator-backed) 2024 – present.
- Speaker on efficient LLM deployment, agentic workflows, and multimodal UX.