Henrik Sachdeva

Vancouver, BC | sachdevahenrik2002@gmail.com | Linkedin | Github | www.henriksachdeva.dev

Summary

Computer Science undergraduate specializing in LLM fine-tuning, alignment, and retrieval-augmented generation. Built and evaluated end-to-end NLP pipelines using PyTorch, Hugging Face, and FAISS, achieving measurable gains on benchmarks such as HotpotQA and E2E Table-to-Text.

Education

Simon Fraser University Burnaby, BC
Bachelor of Computer Science - Concentration in Artificial Intelligence & Machine Learning *Graduating April 2026*
Relevant Coursework: Machine Learning (410), Deep Learning (420), Natural Language Processing(413), Artificial Intelligence (310), Data Structures & Algorithms (130, 135, 307), Linear Algebra, Probability

Technical Skills

DL/ML Frameworks: PyTorch, Hugging Face (Transformers, TRL, Datasets), Scikit-learn, numpy, pandas.
LLM Techniques: RAG (Retrieval-Augmented Generation), ORPO (Preference Optimization), PEFT (LoRA, Prefix Tuning), Instruct Tuning
NLP/Retrieval: FAISS (Vector Indexing), Sentence Embeddings (BGE, MiniLM), Cross-Encoders, GloVe.
Languages & Tools: Python, SQL, Git, Bash, Docker (Conceptual/Project-based)
Models: T5/FLAN-T5, BERT, GPT-2, Qwen 2.5

Projects

Retrieval-Augmented Generation (RAG) for Multi-Hop QA
Technologies: Python, PyTorch, FAISS, HuggingFace, FLAN-T5 December 2025

- Achieved Joint F1 of 33.91% by architecting an end-to-end RAG system to minimize hallucination on the HotpotQA multi-hop reasoning benchmark.
- Engineered a two-stage Retriever pipeline using the BGE Sentence Embedder for initial dense retrieval (FAISS index) and a Cross-Encoder Reranker to select the top 2 supporting facts, boosting Support F1 to 61.95%.
- Integrated and optimized a FLAN-T5-base Generator that processed the retrieved context and question, demonstrating system-level proficiency in combining information retrieval with sequence-to-sequence generation.

Preference Optimization (Instruct Tuning with ORPO)
Technologies: Hugging Face TRL (ORPO Trainer), PyTorch, PEFT (LoRA), Qwen 2.5 November 2025

- Achieved an Instruct-Tuning Score of 52 by applying the state-of-the-art Odds Ratio Preference Optimization (ORPO) algorithm to the Qwen 2.5 0.5B Instruct model.
- Reduced model training parameters by 90%+ by implementing LoRA (Parameter-Efficient Fine-Tuning) with a rank 16, maximizing training efficiency while achieving high alignment.
- Demonstrated mastery in LLM alignment and safety, successfully training the model to follow complex constraints and instructions defined in the Unnatural Instructions dataset.

Prompt Tuning for Text Generation October 2025
Technologies: PyTorch, Hugging Face Transformers, PEFT (Prefix Tuning), distilgpt2, BLEU Score

- Achieved a BLEU score of 30 on the E2E Table-to-Text task by utilizing Prefix Tuning, a Parameter-Efficient Fine-Tuning (PEFT) approach, on the distilgpt2 language model.
- Optimized LLM generation efficiency by tuning virtual token count and prefix projection, adapting the causal model to generate structured data descriptions without modifying the core GPT-2 parameters.

Hackathons & Activities

Competetive Leetcode - Realtime Coding Platform Feb 2025
Journeyhacks by SFU Surge *Burnaby, BC*

- Created a real-time coding game with team-based API integration and live leaderboard.
- Managed backend updates and debugging using terminal tools and automation scripts.
- Practiced debugging and issue resolution using Linux terminal tools.