Harshveer Singh
https://llstringll.github.io/
harshveer321.code@gmail.com • Git: @llStringll • +918427337873

SKILL MATRIX:

Skill	Experience	Companies
C/C++, Python, Shell scripting, Pytorch	5 years	College/IIT Mandi/ Vaultedge/ChartChat/CERN
Statistics, Probability, Linear Algebra, Bayesian Inference	5 years	Research work/College/Vaultedge/Merce-Mettl/Repodox/IIT Mandi/CERN/ToneTag
NLP, NLI, Huggingface,Tensorflow, Kubernetees, Docker, Langchain, LlamaIndex, Unsloth, Weaviate, ChromaDB, Redis, LoRA, PEFT, vLLM	5 years	College/Research work/Vaultedge/Merce-Mettl/Repodox/ChartChat/ ToneTag
VertexAI, Docker, SageMaker	2 years	Vaultedge/Mercer-Mettl, ToneTag

SUMMARY:

- **Data Scientist & ML/AI Engineer** with 5+ years of experience in machine learning, AI, and NLP, specializing in Python, C/C++, and PyTorch.
- Proven track record in **enhancing model performance** by 10-30% and developing **CI/CD pipelines** for efficient model deployment.
- Expertise in designing and deploying agentic, **on-device ASR + SLU** systems (<10 MB) with multi-head pipelines, <200 ms latency, and **production-grade wake-word detection** (<1 MB, 97% recall @ <1% FAR).
- Expertise in **Natural Language Processing (NLP)** techniques, **Bayesian inference**, and **cloud platforms** like VertexAI and AWS SageMaker.
- Research experience at **CERN** and **IIT Mandi**, focusing on innovative AI solutions and deep learning models.
- **Hackathon achievements** include top rankings in national competitions (1st in SATURNALIA, 2nd in PEC-FEST, and 3rd in IITB-TECHFEST).
- Delivered technical presentations on advanced AI topics like **variational inference** and **inductive bias in machine learning**.

TECHNICAL SKILLS:

Programming

- C, C++, Python, Shell scripting
- PyTorch, TensorFlow, Keras
- ScikitLearn, HuggingFace, NLTK
- Pandas, NumPy, Matplotlib

Data Analytics

- Statistics, Probability
- Data visualization with Matplotlib
- Data manipulation with Pandas
- Bayesian Inference

Machine Learning

- Theoretical machine learning and deep learning
- Optimisation algorithms
- Loss Landscape Analysis
- PyTorch, TensorFlow, Keras
- ScikitLearn
- Bayesian Inference

Natural Language Processing (NLP)

- NLP techniques and models
- Natural Language Inference (NLI)
- Conversational Agents
- HuggingFace, NLTK

Cloud Management

- VertexAI, AWS SageMaker
- Docker

TECHNICAL PROJECTS:

ChartChatAI visit here

- An independent in-house multi modal large model (finetuned **Llama 11B-Vision quantized**) with a **RAG laye**r to provide candlestick chart analysis in various formats.
- **Agentic flow for** -> Choose instrument, do analysis, **construct quant algos**, get financial data, pass to LLM for basic-level 1 Algo, and deploy to broker.
- Production stack nodejs, CSS/JS, **weaviate vector db, pytorch, llammacpp, LoRA, PEFT**
- Have a healthy user base with suggestive feedbacks.
- Independently launched on Product Hunt here

Repodox visit here

- An attempt at very large scale and efficient RAG pipeline over public, personalized GitHub repos, to create a system to easily fetch and discuss and generate specifics from a code repo
- Production stack nodejs, CSS/JS, weaviate vector db, pytorch, llammacpp
- Work in progress.

PROFESSIONAL EXPERIENCE:

Tonetag Pvt. Ltd. **Jul'2025 - Present**
Senior ML/AI Engineer (Research)

- Working on **edge-device based ASR and SLU** systems for high accuracy, low latency, low RAM footprint pipelines.
  - Leading research and engineering of edge-device ASR + SLU pipelines with high accuracy, low latency, and constrained RAM footprint.
- Designed **agentic AI for merchant voice payments** on edge, managing the full payment lifecycle—including transactions, confirmations, offers, and customer profiling (**LLM runs in the cloud**).
- Lead, designed and deployed **sub-10MB SLU** models (CNN-GRU/Transformer + contrastive learning) with **>95% intent/slot accuracy** and <200ms inference on ARM CPUs.
- Built <1MB wake-word detector ("Hey ToneTag") with depthwise separable CNN + attention pooling; **reached 97% recall @ <1% FAR** under noisy field conditions.
- Converted PyTorch models to ONNX/TFLite INT8 with dynamic axes; integrated C-level inference hooks for Android/Linux embedded systems.
- Lead, and developed **car command edge mode**l for TATA with on-device ASR + multitask SLU:
  - Heads for open-world entities, Digit-CTC, Retrieval(contact slots)
  - Joint training with **multi-loss on shared latent space.**

Vaultedge Pvt. Ltd. **Oct'22 - Jun'2025**
Applied AI, Data Scientist

- Enhanced production model metrics **by ~11%,** and without regressions for US based NBFC customers
  - **Roberta** based in house trained multilingual model, for newly suggested **heirarchial classification**
  - Custom archotecture allowing **longer context lengths** than off-the-shelf RoBerta models.
- Implemented model interpretability techniques
  - Custom model interpretability checks in **Pytorch, tSNE, Latent Space visualizations**
  - Model regression test suite deveoped and run in **Sagemaker** with model registry in S3 ECR
- Developed and integrated CI/CD pipelines for seamless model deployment, reducing the need for manual human intervention and accelerating the model update cycle.
  - **Weights&Biases** registry to our internal S3 model versioning with Sagemaker inference endpoints deployed using **torchserve**
  - Wrote custom docker containers for **efficient scale up of instances** based on CPUUtilization metrics to serve heavy user load of 10-20 customers serving GBs of pdfs concurrently.
- End-2-end implemented robust data preprocessing pipelines to clean and improve noisy real-world data for language models in production.

    - Data clustering in long range documents to improve on needle-in-haystack kind of scenario, weighted text-segment scoring where weights are learnable and transferable b/w different customers, reducing need of re-training models when onboarding new customers.
    - Latent space visualisations.
    - **Latent space loss functions** to improve and filter out model distribution confusions
- Trained, optimized, and deployed multi-lingual language models in production for Indian NBFC customers ranging from HDFC and Hero Fincorp
    - **LoRA PEFT trained Gemma** model in production.
    - Performed various benchmarking and baseline tests on customer data to improve on KPIs
    - LLM **inference with high latency on cpu**, 75 tokens per second, 20-30 concurrent users
    - **Deployed for production agentic use-case**, flow being -> document ingestion, heirarchical classification, decision based extraction, to checking validitiy of document.

Mercer-Mettl **Jun'21 - Sep'22**
NLP Engineer, AI Team

- Improved inter-sentence and intra-sentence cohesion measuring pipeline precision by 25%, involved feature understanding, reducing the output feature space of BeRT by **putting a posterior on latent space**
    - Stack used: Pytorch, Huggingface, Numpy, Flask, Datasets, Kafka
- Implemented production ready email formality checking pipeline for business environments, involved topic modeling of a raw real-life email set that had noisy lexical structure, through improvised LDA
    - Pytorch, ScikitLearn, Numpy, C++ for super **fast high dimensional LDA** and topic modelling
- Analyzed on-prod spell-check pipeline and suggested and upgrade with specific fine-tuning, that increased recall and precision of the model by ~13% and ~32%.
- Developed an automated interview screening bot using a poly-encoder-based Blender model for chat-style evaluation of candidates; the bot selected from a pre-created question bank and accepted answers in text or voice.
    - Evaluated responses by computing similarity scores against a curated answer bank using RAG-like mechanisms with smaller decoder models; used score thresholds for selection.
    - Stack used: Pytorch, Transformers (BlenderBot, Poly-encoders), SpeechRecognition, Librosa, Flask

Mercer-Mettl **Jan'21 - Jun'21**
NLP Engineer Intern, AI Team

- Worked on improving NLI models, improved cohesion detection accuracy on English text by ~16%

CMS Experiment, CERN **Jan'21 - May'21**
Deep Learning Research Intern

- Worked on building a quasi-linear attention model to isolate 'interesting' events from the background
- during the collision of protons with Low-Z targets.

IIT-Mandi **Dec'19 - Jan'20**
Research Intern, Department of Mathematics

- Provided an analytical study on the success of Batch Normalization [Blog]

RESEARCH EXPERIENCE:

Cross-layer residual connection transformer **Oct'20 - Nov'20**

- Developed a novel architecture, which has a recursively "smooth" loss surface, allowing the possibility
- of reaching more generalized minima, even in the absence of good parameter initialization.

Adversarial Training for Facebook's Blender **June'20 - Aug'20**

- Created a self-play regime for conversational agents, and extending that to a competitive conversation where an agent discriminates the output distribution of the other agent against human dialogue distribution.

Poly encoder regime for fine-tuning decoder-only model (GPT-2) **May' 20**

- Showed that a decoder model fine-tuned like such on language modeling apparently is more robust to inductive bias than encoder model even though encoder reached better recall@k/C score

Analytical study of the success of Batch Norm **Nov'19 - Dec'19**

- Showed that batch normalization **smooths the loss surface** and how it brings that effect, through the study of eigenvalues of the hessian of weight matrix. [Blog]

Beta2 variation regime for Adam Optimizer **May'18 - Jul'18**

- Developed a regime for varying beta2 hyper-parameter of Adam, preventing Adam from getting stuck
- in sub-optimal minima.
- A similar result was also shown in a subsection of Sashank J. Reddi et al.

HACKATHONS AND TECHNICAL PRESENTATIONS:

- **Hackathons, 2017 2018**
  - SATURNALIA Hackathon '17 ranked 1st PEC-FEST Hackathon '17 ranked 2nd
  - IITB-TECHFEST Hackathon'18 ranked3rd
- Causality and its importance in variational inference and EM, TIET Jan '20
- Inductive bias in machine learning models, TIET Oct '19
- Effect of constraining the posterior to Gaussian in VAEs, IITB-TECHFEST Nov '17