

SHIVAM KUMAR

Berlin, Germany

+49 1575 6479751

shivammehta1924@gmail.com

github.com/nuss171201

shivammehta.dev

EDUCATION

SRH Berlin University of Applied Sciences

Bachelor of Science in Computer Science

Oct. 2024 – Jul. 2027 (expected)

Berlin, Germany

FOCUS AREAS

- Large Language Models
- Transformer Architectures
- LLM Fine-Tuning
- Reasoning Models
- Deep Learning (PyTorch)
- Classical Machine Learning

PROJECTS

Reasoning Model Training & Inference | *Python, PyTorch, Qwen3-0.6B*

Sep. 2026 – Oct. 2026

- Designed streaming inference utilities for Qwen3-0.6B with KV caching, temperature scaling and top-p sampling; implemented inference-time scaling via sampling and self-refinement.
- Implemented GRPO components (verifiable math rewards, group-normalized advantages, clipped policy-ratio loss, KL penalty, format reward, log-probability and entropy tracking) and ran GRPO training on Apple Silicon.
- Prepared a distillation dataset from 12,000 DeepSeek-R1 reasoning traces, using an automated answer checker to verify that 90.6% of the traces were correct, and implemented the distillation training loss.

Transformer & GPT from Scratch | *Python, PyTorch, tiktoken*

Aug. 2026 – Sep. 2026

- Built a character-level language model step by step from random and bigram baselines through an MLP and attention to a 6-block transformer, lowering training loss from 2.35 (MLP) to 1.93; extended it to a word-level model on *The Odyssey* with MPS acceleration.
- Built a GPT-style model with text generation, loaded pretrained GPT-2 124M weights, and prepared a balanced SMS spam dataset with tokenization, padding and reproducible train/validation/test splits.

Classical ML Projects | *Python, scikit-learn, pandas, NumPy*

Jul. 2026

- Built housing regression workflows with stratified splits, preprocessing pipelines, cross-validation and GridSearchCV.
- Trained and compared SGD, SVM, Random Forest and k-NN classifiers on MNIST, evaluating precision, recall, F1, ROC/AUC and confusion matrices.

RAG & Agent Prototypes | *LangChain, LangGraph, FAISS, OpenAI API, Ollama*

May 2026

- Built web and PDF retrieval pipelines with document chunking, embeddings and FAISS vector search, preserving source and page context.
- Developed LangGraph agent experiments with retrieval tools, URL fetching and conversational state, using OpenAI models and local LLMs via Ollama.

Movie Revenue Analysis & Prediction | *Python, pandas, scikit-learn, Matplotlib, Seaborn*

Dec. 2025

- Cleaned and merged TMDB metadata, credits and keywords for 45,000+ movies (Kaggle), parsing nested JSON fields and engineering profit and ROI features.
- Trained a linear regression model explaining box-office revenue from budget, popularity and vote count, reaching $R^2 = 0.76$ on held-out data.

TECHNICAL SKILLS

Languages: Python (Advanced); SQL (Intermediate)

ML / Deep Learning: PyTorch, NumPy, pandas (Advanced); scikit-learn, Matplotlib (Intermediate)

LLMs & RAG: transformers, tokenization, KV caching, RAG, GRPO, LangChain, LangGraph, FAISS, Ollama

Backend & Tools: FastAPI, Pydantic, Git/GitHub, Docker, Jupyter, uv

INDEPENDENT STUDY (TECHNICAL READING)

Hands-On Machine Learning (Géron), *Deep Learning with PyTorch* (Stevens et al.), *Build a Large Language Model (From Scratch)* (Raschka), *Build a Reasoning Model (From Scratch)* (Raschka), *AI Engineering* (Huyen), *Practical SQL* (DeBarros), *Docker Deep Dive* (Poulton)