Yash Bishnoi

Machine Learning Engineer | PyTorch, Model Reproduction, Evaluation, Open Source

Glasgow, UK (open to relocation) | UK Graduate Visa, no sponsorship required | bishnoiyash274@gmail.com | LinkedIn | GitHub | Portfolio | Hugging Face | Kaggle

PROFESSIONAL SUMMARY

Machine learning engineer with an MSc in Computer Science who reproduces language models from scratch, verifies them bit-exact against official weights and runs controlled experiments. Six pull requests merged upstream in **PyTorch** core, **PyTorch Helion**, **bitsandbytes** and **Liger-Kernel**, each after maintainer review. Research co-author on the revision of an SAE steering side-effects paper (arXiv:2606.08365) after an independent reproduction and a new control.

WORK EXPERIENCE

Open Source Contributor | PyTorch, PyTorch Helion, bitsandbytes, TRL, Liger-Kernel 07/2026 - Present

Tech: PyTorch, TorchDynamo, C++/CUDA, Triton, GEMM dispatch heuristics, GRPO, pytest, GB10 (sm121)

- **Compiler:** landed two fixes in **PyTorch** core (**merged**, commits 59f3b80 and 075742a) splitting FrozensetVariable and OrderedSetVariable off SetVariable in the **dynamo** VariableTracker hierarchy to match CPython.
- **Kernels, C++/CUDA:** merged two changes to the 4-bit GEMM dispatch heuristics in **bitsandbytes** (csrc/gemm_4bit.cu), adding **sm121 (GB10)** tiers chosen on measured tile timings (**median 2.52x**).
- **Frameworks:** merged a fix in **PyTorch Helion** skipping pretuned scaled_mm configs that exceed device shared memory (**147KB** required vs **101KB** available on GB10), plus an sm121 heuristic file.
- **Triage:** reproduced and triaged **11** upstream issues and PRs across **vLLM**, **FlashInfer**, **PyTorch**, **TRL** and **Liger-Kernel**, mostly silent-corruption and kernel-cache failures on sm120/sm121.

Research Co-author | SAE steering side-effects paper (arXiv:2606.08365), with Evan Duan 08/2026 - Present

Tech: Sparse autoencoders, GPT-2-small, activation steering, Spearman and partial correlation, Hugging Face, Kaggle

- Independently reproduced the paper's central result on **GPT-2-small**: decoder crowding predicts steering collateral at Spearman **0.482** vs the published **0.466**.
- Ran a control the paper had not: partialled out firing frequency and activation magnitude, showing crowding survives but frequency is **not inert**, which narrows the reported margin.
- Published the reproduction as a **Hugging Face** blog post and an Apache-2.0 **Kaggle** notebook that re-runs end to end in **~4 min** on a free T4.

Machine Learning Engineer (Independent Research) | BuildFromScratch, self-directed 10/2025 - Present

Tech: PyTorch, Transformers, RoPE, SmolLM2, Qwen3, lm-evaluation-harness, CUDA memory management, pytest

- Reproduced **SmolLM2-135M** and **Qwen3-0.6B** from scratch in single-file **PyTorch**, each verified bit-exact against official Hugging Face weights (**max absolute logit difference = 0.0**).
- Ran a matched-compute experiment on Qwen3-0.6B: a modernised architecture bundle beat the faithful baseline by **17.9%** (**23.52** vs **28.65** perplexity); a second method lost and was reported as a negative result.
- Ran single-variable, 3-seed ablations with 95% confidence intervals (NorMuon vs AdamW: **+0.474 bpb** on WikiText, 95% CI **[+0.443, +0.505]**).

Team Leader | Co-op | Dundee, UK 08/2025 - Present

- Led a team of **8+** across shifts, allocating work and coordinating cover through peak trade.
- Trained and mentored new starters on process and standards, owning consistency across the team.

Market Research Analyst | Nidhi Herbal | India 07/2021 - 09/2024

- Built **Power BI** dashboards in **DAX** translating sales, supplier ROI and demand for non-technical stakeholders.
- Automated **SQL** and Excel ETL in **Python**, cutting monthly report prep by **35%**.

TECHNICAL SKILLS

Machine Learning: PyTorch | Transformers | model reproduction | evaluation (lm-evaluation-harness) | sparse autoencoders | scikit-learn | XGBoost | MLflow
Systems and Kernels: C++ | CUDA | Triton | TorchDynamo | Linux | Docker | Kubernetes
Languages: Python | SQL | TypeScript | JavaScript | C++
Data and Databases: PostgreSQL | MongoDB | MySQL | SQLite | Redis | Qdrant | Neo4j | ETL pipelines | Apache Kafka | Airflow | Spark | pandas | NumPy
Engineering Practice: pytest | GitHub Actions (CI/CD) | Git | Agile and Scrum | Jira | technical writing

PROJECTS

BuildFromScratch: language models reproduced from scratch and verified bit-exact | github.com/yashb98/BuildFromScratch

Tech: Python, PyTorch, Transformers, RoPE, SmolLM2, Qwen3, lm-evaluation-harness, pytest

- Reproduced **SmolLM2-135M** and **Qwen3-0.6B** as single-file PyTorch models, each bit-exact against the official Hugging Face weights (**max absolute logit difference = 0.0**).
- Beat the faithful Qwen3-0.6B baseline by **~18%** at matched compute with a modernised build (**23.52** vs **28.65** perplexity); a partial-RoPE build lost and is reported as a negative result.
- Reproduced the faithful baseline within **2.14x** of the original using **~30,000x** less data, gated by a pytest architecture-parity suite.

Velox AI: production voice-agent platform, design through delivery | github.com/yashb98/Velox_AI

Tech: TypeScript, React, Node.js, FastAPI, PostgreSQL, Docker, GCP, Terraform, Twilio, Deepgram, Gemini

- Shipped a **6-service** Docker stack on GCP with Terraform and Prometheus, targeting **sub-2-second** end-to-end voice responses.
- Cut cost per request by routing **~70%** of simple turns to a local Phi-3-mini ahead of the frontier model, behind a **5-layer** anti-hallucination RAG system.

DataMind: AI-native data analytics and visualisation platform | github.com/yashb98/DataMind

Tech: Next.js, React, FastAPI, PostgreSQL, Qdrant, Neo4j, Kafka, Flink, Apache Iceberg, Kubernetes, MLflow

- Designed and delivered a full-stack analytics platform in a **28-day**, **6-phase** build: a Next.js and React front end with **25+** chart types over FastAPI and **PostgreSQL**.
- Engineered the data layer as a lakehouse (Apache Iceberg, Kafka, Flink) with Qdrant and Neo4j stores, returning full RAG queries in under **600ms** at p50.

OPEN SOURCE AND ACHIEVEMENTS

Merged upstream: six code pull requests, one co-authored commit and one documentation release

- **PyTorch** #193580 (commit 59f3b80): dynamo FrozensetVariable split off SetVariable to match CPython.
- **PyTorch** #196167 (commit 075742a): dynamo OrderedSetVariable split off SetVariable to match CPython.
- **PyTorch Helion** #3415: skip pretuned scaled_mm configs that exceed device shared memory; sm121 heuristic.
- **PyTorch Helion** #3963 (commit ff1a45c, co-authored): epilogue-subtiling store-dimension fix.
- **bitsandbytes** #2030: sm121 (GB10) tier in the 4-bit GEMM dispatch heuristic.
- **bitsandbytes** #2039: sm121 (GB10) in the wide_n_simt 4-bit GEMM dispatch rule.
- **Liger-Kernel** #1332: requires-python 3.10+ in pyproject.toml, closing issue #1302.
- **Project-HAMi website** #865: v2.10.0 documentation version.

Open pull requests

- **Triton** #11363: reject TMA scatter on consumer Blackwell (sm_12x) at compile time.
- **PyTorch Helion** #4045: autotuner fills short per-loop lists in user seed configs.
- **PyTorch Helion** #4097: autotuner fits configs that exceed GPU limits at their first launch.
- **torchao** #4869: make Int8 and Float8 weight-only reapplication safe.
- **TRL** #7085: fix FP8 weight sync for colocated vLLM.
- **JAX** #40644: include jax_high_dynamic_range_gumbel in the JIT cache key.
- **Liger-Kernel** #1328: fix FLCE torch.compile failure from the aten.addmm.dtype_out overload.
- **compressed-tensors** #844: read attention quantization shapes from each layer's config.

Achievements

- Kaggle RSNA knee MRI: **18th of 785** on the public leaderboard, Aug 2026.
- HackerRank: global rank **#198** in the Artificial Intelligence domain.

EDUCATION

MSc Computer Science (Merit) | University of Dundee — 01/2025 - 01/2026
Dissertation: Deep Learning for Facial 3D Reconstruction - Simulator

MBA (Finance) | JECRC University — 2019 - 2021

CERTIFICATIONS

- **IBM Machine Learning** (Verify, 07/2023)
- **Deep Learning and Reinforcement Learning** (Verify, 07/2023)
- **Introduction to Model Context Protocol** (Verify, 02/2026)