Yash Sawant

LLM post-training & RL — SFT · reward modeling · GRPO / RLHF · model personalization · memory-native architectures

Irvine, CA | +1-619-983-5500 | ysawant@uci.edu | LinkedIn | GitHub | Scholar

EXPERIENCE

Wethos AI — Irvine, CA
ML Engineer | LLM Post-Training & Personalization — *Jan 2025 – Present*

- Own the full **post-training pipeline (CPT → SFT → GRPO)** for a production coaching model on Qwen — built the reward-modeling and GRPO stages end-to-end, replacing a third-party Gemini API with an owned fine-tuned model.
- Core technical contributor to ATLAS, Wethos's **memory-native foundation model** — an LLM with persistent memory and recall built into the architecture, no per-turn context re-injection (lineage: Memorizing Transformers, kNN-LM, Titans); built the working demo used in investor technical deep-dives.
- Architected an end-to-end **model-personalization pipeline**: PEFT/LoRA fine-tuning across LLaMA, Mistral, and Qwen to replicate an individual's style and domain knowledge from sparse (5–10 doc) data (with a synthetic-data stage), evaluated by a LUAR-based author-style fidelity metric.
- Built **multi-tenant LoRA serving** — one base model serves all users via per-request adapter hot-swapping, holding per-user GPU memory near-constant instead of scaling linearly with users.

Bank of New York Mellon — Pittsburgh, PA
Data Scientist — *Sept 2024 – Dec 2024*

- Built a **Mixtral-8x7b** dual-agent document classifier for regulatory-risk triage over a ChromaDB retrieval layer (web-scraped ingestion), outperforming single-prompt baselines.

JerseySTEM — Florham Park, NJ
Data Science Engineer — *Feb 2024 – Aug 2024*

- Built an LLM report-generation pipeline (Azure OpenAI) over a **SQL/BigQuery** ingestion layer and a Neo4j graph for multi-source retrieval.

PROJECTS

Mechanistic Interpretability of an RL Step — Crosscoder Diffing & Steering — rl-reasoning-diff · rl-crosscoder

- Trained a **TopK crosscoder** to model-diff two OLMo-2-1B checkpoints across their final RL step (DPO → RLVR) — **FVE 0.87**, 5.6% dead, published to the HuggingFace Hub; diagnosed and fixed the decoder-norm L1 gauge-freedom that defeats naive crosscoder sparsity.
- Showed RL's representational edit is diffuse (largest single-feature change +6.5%; weight change ∼500× smaller than the preceding SFT+DPO) yet lands **preferentially on reasoning features** (2.8× baseline).
- **Causally confirmed by steering**: injecting the RL-up-weighted features into the pre-RL model turns reasoning on (0.08 → 0.92) while random and down-weighted directions stay flat — *RL edits **when** a model reasons, not **how***, a mechanistic account of my EMNLP GRPO paper.

embodied-gpt — Foundation Models From Scratch — github.com/yashsawant22/embodied-gpt

- Built a decoder-only **GPT** and a **Vision Transformer** from scratch (attention, residual streams, patch embeddings, training loops — no black boxes), now extending the same next-token framing through VLM → VLA → robot foundation models.

PersonalBench — LLM Personalization Benchmark — github.com/yashsawant22/personalbench

- Open benchmark evaluating LLM personalization across four axes — style fidelity, knowledge retention, preference consistency, temporal stability — companion to the ICML 2026 CTB paper above.

RESEARCH & WRITING

- **Y. Sawant**. "When Gradient Importance Lies: Adaptive LoRA Rank Allocation Fails Under GRPO." *Insights from Negative Results in NLP @ EMNLP 2026* **(Accepted, archival).** — controlled study showing gradient-importance rank allocation that helps under SFT degrades under GRPO. code
- **Y. Sawant**. "Theory-Grounded Evaluation Exposes the Authorship Gap in LLM Personalization." *CTB @ ICML 2026* **(Accepted).** arXiv:2604.26460
- **Technical writing** — blog on training internals, post-training, and building models from scratch. · **Reviewer** — 22 reviews across 6 venues (NeurIPS 2026, 4 ICML 2026 workshops, ACL 2026 SELVA).

EDUCATION

University of California, Irvine — Irvine, CA
Master of Data Science | GPA: 3.7/4.0 — *Dec 2023*

NMIMS University — Mumbai, India
B.Tech in Data Science | GPA: 3.6/4.0 — *May 2022*

TECHNICAL SKILLS

Post-training / RL:	SFT, reward modeling, GRPO, RLHF-style, CPT, PEFT/LoRA, synthetic data generation
ML / Deep Learning:	PyTorch, HuggingFace / transformers, transformer architecture, model evaluation (LUAR)
Interpretability:	Crosscoders, sparse autoencoders (TopK), model-diffing, activation steering, TransformerLens
Systems / Infra:	Multi-tenant LoRA serving, Docker, GCP / AWS / Azure
Languages:	Python, C++, SQL, R, Scala, Java
Models:	LLaMA, Mistral, Qwen, Mixtral, GPT, BERT