Jash Mehta

✉ jashmehta3300@gmail.com jash-mehta-3300 jashmehta3300.github.io jashmehta3300

Summary

I am an Applied Research Scientist at ServiceNow, where I work on the **Apriel** family of reasoning and multimodal LLMs achieving SOTA performance in enterprise AI. I earned my M.S. in CS from Georgia Tech and have published research at top *CL venues, including EMNLP and EACL. I am seeking roles in AI research and applications centered on LLMs and multimodal intelligence.

Experience

ServiceNow, Applied Research Scientist June 2024 - Present

- Core contributor to **Apriel-1.5-15b-Thinker**: [demo] [report] [model]
  - 52 on Artificial Analysis Index; similar to Deepseek R1 0528 and Gemini 2.5 Flash.
  - #1 in Small AI Models ($4B - 40B$) category; best performance on τ^2-Bench (Agentic) & IFBench.
  - At 15B parameters, it's over 10× smaller than comparable models that score >50 on AA.
- Shipped open-weights **Apriel-Nemotron** models: Apriel-Nemotron-15b-Thinker and Apriel-5B-Base
  - SOTA performance on enterprise benchmarks while consuming 40% fewer tokens than QwQ 32B.
  - Pretrained on 4.5T tokens; achieves performance similar to OpenAI o1-mini and QwQ 32B.
  - Recognized by NVIDIA CEO Jensen Huang at ServiceNow Knowledge 2025.
- Multimodal Language Models:
  - AU-Harness: an open-source toolkit for multimodal LLM evaluation with 127% faster processing across 380+ audio-language tasks. [github] [website] [paper]
  - Training latency-aware LALMs for audio understanding, audio generation and enterprise tasks.
- Long-context support upto 1M tokens:
  - Implemented attention mechanisms like DCA for inference-time context expansion.
  - Expand context-window efficiently using techniques like YARN, NTK, etc.
  - Support sequence length and model parallelism for training upto 128k tokens in FastLLM.

Cisco, Software Engineer II (Intern) May 2023 - Aug 2023

- Founder & developer of MartianBank: Cisco's open-source project used by 10+ Outshift teams.
- Architected a microservice app in Python, JavaScript, and Go to enhance software supply chain security.
- Deployed on AWS cluster using Docker and Kubernetes; and set up CI/CD pipelines and test suits.

Research Student, Guide: Prof. Zeerak Talat, University of Edinburgh Jan 2022 - Apr 2023

- Mined $\sim$1M tweets to create a novel low-resource dataset & experimented with bi-LSTM, mBERT, XLM-R, etc. using PyTorch and Huggingface **(EMNLP 2022)**
- Fine-tuned transformers in FL setting to obtain 14.52% improvement in F1-score. **(EACL 2023)**.

Unicode Research, Guide: Dr. Swapneel Mehta, MIT Jan 2022 - Apr 2023

- Served as TA for **Google Research** funded ML Course UMLSC 2021 with 100+ students.
- Worked with the SimPPL team to build better civic integrity AI tools (supported by Google, AWS).

Education

Georgia Institute of Technology Aug 2022 - May 2024
M.S. in Computer Science, Specialization: Machine Learning GPA: 4.0/4.0

- Teaching Assistant: CS 6220 Big Data (Fall 22 & 23); CS 6675 Adv. Computing Systems (Spring 23 & 24)
- M.S. Project (Advisor: Prof. Ling Liu): Privacy & Security in LLMs.

University of Mumbai Aug 2018 - May 2022
B.E. in Computer Engineering GPA: 9.67/10

- Research Assistant (Advisor: Prof. Ram Mangrulkar): NLP & federated learning.

Skills

Languages:	Python, Javascript, Golang, C, HTML
AI / ML:	PyTorch, Hugging Face Transformers, vLLM, Pandas, NumPy, LangChain
Web / Databases:	FastAPI, Flask, Node.js, React.js, REST / gRPC, SQL, Redis, SQL
Tools:	Weights & Biases, Git, Kubernetes, Docker, Jupyter, Bash

Projects

Fine-tuning LLMs using Social Reputation Signals (Guide: Prof. Judy Hoffman) Video ▶

- Utilized content reputation as cost-effective supervision signal to bypass costly human feedback (RLHF).
- Created end-to-end pipeline for fine-tuning of LLaMA models using PyTorch and Huggingface.

Efficient LLM Inference

- Studied current state-of-the-art techniques for efficient LLM inference like **vLLM**, Hydra, Orca, etc.
- Implemented staged (decode draft model) speculative decoding that further improves the performance.

Heterogeneous SuperFed (Guide: Prof. Alexey Tumanov)

- Co-trained a large family of models in FL with weight-shared learning (reducing cost from $O(K)$ to $O(1)$).
- Using knowledge distillation to enable training in truly heterogeneous (compute, model and data) conditions.

Selected Publications [Full List]

[1] S. Radhakrishna, A. Tiwari, A. Shukla, M. Hashemi, R. Maheshwary, S. K. R. Malay, J. Mehta, P. Pattnaik, S. Mittal, K. Slimi, K. Ogueji, A. Oladipo, S. Parikh, O. Bamgbose, T. Liang, A. Masry, K. Mahajan, S. R. Mudumba, V. Yadav, S. T. Madhusudhan, T. Scholak, S. Davasam, S. Sunkara, and N. Chapados, "**Apriel-1.5-15b-Thinker**," 2025 🎓 [technical report].

[2] S. Radhakrishna, S. Parikh, G. Sarda, A. Turkkan, Q. Vohra, R. Li, D. Jhamb, K. Ogueji, A. Shukla, O. Bamgbose, *et al.*, "**Apriel Nemotron 15b Thinker**," 2025 🎓 [technical report].

[3] S. Surapaneni, H. Nguyen, J. Mehta, A. Tiwari, O. Bamgbose, A. Kalkunte, S. Rajeswar, and S. Tejaswi Madhusudhan, "**AU-Harness: An Open-Source Toolkit for Holistic Evaluation of Audio LLMs**," 2025 🎓 [under review - ICML].

[4] S. T. Madhusudhan, S. Radhakrishna, J. Mehta, and T. Liang, "**Millions scale dataset distilled from r1-32b**," 2025.

[5] J. Gala, D. Gandhi, J. Mehta, and Z. Talat, "A federated approach for hate speech detection," in ***EACL 2023***, 🎓 [Impact factor: **8.80**].

[6] D. Gandhi, J. Mehta, N. Parekh, K. Waghela, L. D'Mello, and Z. Talat, "A federated approach to predict emojis in hindi tweets," in ***EMNLP 2022***, 🎓 [Impact factor: **23.1**].

[7] J. Mehta, D. Gandhi, N. Rathod, and S. Bagul, "**IndicFed: A federated approach for sentiment analysis in indic languages**," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pp. 487–492, 2021 🎓.

Co-curricular activities & achievements

- Awarded Inspire Scholarship, **Top 1%** candidates in Higher Secondary Certificate (12th Grade), 2018.
- Part of Shalizi–Stats reading group which focuses on the stats book "Advanced Data Analysis from an Elementary Point of View" by Prof. Cosma Shalizi and Bayesian Machine Learning.
- Attended the Advanced Language Processing Winter School (ALPS) 2022.
- Built a predictive model for automotive component part failure for a Big4 consultancy firm.