Shubhangi Upasani

Sunnyvale, CA | LinkedIn | Google Scholar

Work Experience

- **SambaNova Systems** — **Palo Alto, CA**
  Principal Engineer (Machine Learning) — ***Jul 2021 - Present***
  - Agentic Context Engineering (ACE) for LLMs - Context, Memory and Post-Training
    - Core contributor to ACE that treats context as an evolving playbook.
    - Engineered agentic systems to develop context playbooks through continuous generation, reflection, and curation.
    - Developed rejection-based sampling + SFT framework to improve context quality.
    - Implemented DPO-style post-training for reflector models using reward-guided data selection.
    - ACE outperformed agent and domain baselines (8–10% gains); accepted ICLR 2026.
  - Multimodal LLM pretraining and finetuning
    - Contributed to multimodal LLM pretraining (e.g., LLaVA, Llama-3.2), including data curation, training and evaluation.
    - Designed and optimized SFT/LoRA post-training pipelines for 7-11B multimodal LLMs with 550k image-text pairs.
    - Improved downstream performance and image grounding on biomedical VLMs and Japanese document QA.
  - LLM inference and long context optimization
    - Designed SAGE-KV, a KV-cache eviction method for enabling 128K-token LLM inference.
    - Achieved 2x-4x higher memory efficiency across multiple LLM backbones; 3.1x speedup over full KV caching.
    - Accepted at SLLM Workshop, ICLR 2025.
  - Memory and latency optimizations for long context inference
    - Reduced Time-to-First-Token (TTFT) using distributed prefill with Star attention for long context inputs.
    - Enabled low-TTFT decoding (2.5× speedup) under fixed 8K KV budgets via Star Attention + SnapKV.
    - Shipped DeepSeek-R1 (128K context) on SambaNova Cloud.
  - Self-speculative prefill for text compression
    - Designed training-free cross-model self-speculative prefill methods for prompt compression; accepted at RSI, ICLR 2026.
    - Achieved 90-100% accuracy retention when compared to full prompt baseline with significant TTFT reduction.
- **Amazon** — **Seattle, Washington**
  Software Development Engineer Intern — ***May 2020 - July 2020***
  - Led end-to-end development (design, coding, testing, deployment) of visualizer tool for optimizing inventory plans.
  - Improved plan load times by ∼50%; Reduced cost of backend services to $10 (3-fold reduction).

Education

- **Georgia Institute of Technology** — Atlanta, Georgia
  Master of Science in Computer Science – Machine Learning Specialization ; GPA: 4.00 — *Aug 2019 - May 2021*
  Coursework: Machine Learning, Computer Vision, Natural Language Processing, Deep Learning, Artificial Intelligence, Big Data
- **Delhi Technological University** — New Delhi, India
  Bachelor of Technology – Electronics and Communication Engineering ; GPA: 9.21/10 — *Aug 2015 - May 2019*

Research Experience

- Zhang, Qizheng, Changran Hu, Shubhangi Upasani et al. (2025) Agentic context engineering: Evolving contexts for self-improving language models. ICLR 2026
- S. Upasani, R. Raju et al. (2026) Cross-Family Speculative Prefill: Training-Free Long-Context Compression with Small Draft Models. RSI Workshop, ICLR 2026
- S.Upasani et al. (2026) Test-Time Adaptation via Many-Shot Prompting: Benefits and Pitfalls. RSI Workshop, ICLR 2026
- G. Wang, S. Upasani et al. (2025) LLMs Know What to Drop: Self-Attention Guided KV Cache Eviction for Efficient Long-Context Inference. SLLM Workshop, ICLR 2025
- Li, Jonathan, Nasim Farahini, Evgenii Iuliugin, Magnus Vesterlund, Christian Häggström, Guangtao Wang, Shubhangi Upasani et al. (2025) SnapStream: Efficient Long Sequence Decoding on Dataflow Accelerators. ISC 2026
- Kumar A., Nayyar A., Upasani S. et al. (2020) Empirical Study of Soft Clustering Technique for Determining Click Through Rate in Online Advertising. Advances in Intelligent Systems and Computing. Springer 2020

Skills Summary

- **Languages**: Python, C/C++, MATLAB, HTML, CSS, JavaScript
- **Tools**: PyTorch, HuggingFace, vLLM, Scikit-learn, Git, Slurm, GPU, Distributed Training (Data Parallelism, Tensor Parallelism, FSDP)