Saif Punjwani

saifpunjwani1230@gmail.com | 516-778-2495 | LinkedIn | GitHub | Portfolio | Google Scholar

Education

Columbia University **August 2025 – May 2029 (Expected)**
Incoming M.S./Ph.D. in Computer Science *New York, NY*

Georgia Institute of Technology **August 2021 – May 2025**
Joint BS/MS in Computer Science, Minor in Mathematics (GPA: 4.0/4.0) *Atlanta, GA*

Threads: System Architecture and Intelligence

Relevant Coursework: Diff Eq, Lin Alg, Dig Sys Design, High Perf Comp Arch, Adv Comp Arch, Data Struct, Algos, Prob Stat, Disc Math, Sys Arch, Compilers & Interp, ML (Grad), Operating Sys, NLP (Grad), Deep Learning (Grad), Sys & Networks, Combo, Robo & Percep, Automata & Complex, Proc Design, Deep Reinforcement Learning (Grad)

Research Experience

Natural Language Processing (NLP) X Lab **October 2024 – Present**
Undergraduate Researcher *Atlanta, GA*

- Investigating new approaches to enhance reasoning in large language models, drawing inspiration from cognitive science to improve multi-step inference capabilities as a joint-lab investigation with AI Voice Assistant Laboratory under Dr. Larry Heck and the NLP X Lab, primary advisor being Dr. Alan Ritter and collaboration with Dr. Wei Xu.
- Developing *Weight of Thought Reasoning*, a novel framework that dynamically prioritizes model attention based on logical depth and contextual relevance to achieve higher reasoning accuracy.
- Creating benchmarks and datasets to evaluate reasoning improvements, focusing on interpretability, consistency, and adaptability across diverse linguistic domains.

Artificial Intelligence Voice Assistant Laboratories **October 2021 – Present**
Undergraduate Researcher *Atlanta, CA*

- Pioneered Large Body Language Models (LBLMs) for real-time gesture generation, developing LBLM-AVA architecture with Transformer-XL and parallelized diffusion achieving 30% reduction in FGD and 25% improvement in inception, working under Dr. Larry Heck.
- Developed probabilistic reasoning frameworks integrating symbolic knowledge with neural architectures, achieving 45% improvement in logical consistency through differentiable theorem proving and latent symbolic manipulation.
- Created Allo-AVA, a large-scale dataset (135B keypoints, 15M words) for allocentric avatar animation, establishing speech-gesture synchronization across 7,500 diverse videos.

Stanford University Artificial Intelligence Lab (SAIL) **May 2023 – August 2023**
Research Scientist Intern *Stanford, CA*

- Researched novel compositional grammar-guided attention mechanisms for autoregressive language models, achieving 37% reduction in perplexity (from 12.4 to 7.8) and 42% improvement in zero-shot generalization across 8 downstream tasks.
- Developed theoretical foundations for high-dimensional representation learning using stochastic differential equations and optimal transport theory, implementing custom PyTorch modules for non-Euclidean geometric deep learning.
- Designed sparse conditional random fields for kinematic trajectory prediction, achieving 98% accuracy (AUC-ROC 0.96) on the BABEL dataset through cross-entropy minimization and manifold regularization from 2.1M sequential samples.

Columbia University Medical Center Neuroengineering **January 2020 – June 2021**
Research/Software Engineering Intern *New York, NY*

- Collaborated with a team of 4 to program 20 sensors in C++ analyzing brain wave data for patients with neurological diseases.
- Created a convolutional neural network model (CNN) using Python and MATLAB to analyze fracture point patterns for sensors.
- Engineered 8 neuro-sensors using microcontrollers, programmed in C to streamline data received from neural impulses.

Structural DNA Nanotechnology Laboratory, New York University (NYU) **March 2019 – June 2021**
Research/Software Engineering Intern *New York, NY*

- Developed a neural net/nearest neighbor net in TensorFlow to accurately predict DNA crystals' thermodynamic properties with 99% accuracy and grouped 10,000 data points for use around the laboratory and scientific community, model developed in Linux.
- Streamlined machine learning regression model data using C++ to use in nanofabrication software and nanorobotics.
- Published as the 2nd author in the *ACS Nano* journal with other post-docs and professors.

Work Experience

General Robotics **May 2025 – August 2025**
Research Scientist Intern *Redmond, WA*

- Built a real-time perception→grasp→motion stack on UR5e+Robotiq; fused RealSense/ZED, SAM2, FoundationStereo; GraspGen+ICP; executed via GRID and wrapped as an *agent* (planner→actor→checker) with on-the-fly object reconstruction for planning.
- Trained hierarchical deep RL: behavior-cloned low-level impedance/velocity controllers, then `SAC+HER` fine-tuning in vectorized sim; option policy (reach/align/grasp/lift/place) with domain+dynamics randomization for sim→real.
- Designed a multimodal reasoning layer: fused language goals, visual detections, depth geometry, and proprioception into a program-of-thought planner; orchestrated tool calls (segmentation, 3D query, grasp scoring) with a verifier that checks predicted effects against state to trigger safe retries/replans.

Scale AI **January 2025 – March 2025**
Research Intern *San Francisco, CA*

- Designed a reasoning-agent evaluation harness spanning text and multimodal tasks with pass@k, step-consistency, latency, and cost metrics to track reliability over long chains of thought.
- Built tool-using agents with planner→actor→checker loops and program-of-thought execution, improving success on compositional, long-horizon benchmarks under tight compute budgets.
- Prototyped adaptive graph-based routing routing that allocates more compute to harder subproblems; integrated PyTorch batching to lower per-sample inference cost without degrading accuracy.

Barometer Inc. **May 2023 – Present**
Lead Research Engineer *New York, NY*

- Engineered our Host Intelligence engine using PyTorch and C++, incorporating our proprietary large-language model and real-time news analysis to enhance brand safety decision-making, achieving an 80% increase in efficiency.
- Led development of Video Analysis Engine using TensorFlow, OpenCV, and AWS Lambda in Linux boosting ROI by 30% and expanding to 20+ media companies and affecting 2 million users.
- Incorporated Random Forests and Gradient Boosting with Kafka and Spark, cutting data processing by 25%.

CS 8803 - Conversational AI **September 2024 – December 2024**
Graduate Teaching Assistant *Atlanta, GA*

- Led weekly recitations for 150+ graduate students covering advanced topics in conversational AI including transformer architectures, dialogue systems, and neural language generation under Dr. Larry Heck.
- Developed and graded programming assignments focused on implementing state-of-the-art language models, achieving 30% improvement in student project completion rates through detailed feedback and guidance.
- Held office hours and created supplementary materials on multi-turn dialogue systems, reinforcement learning for conversation, and multimodal interactions, resulting in 25% increase in assignment scores.

Voloridge Investment Management **May 2024 – August 2024**
Quantitative Research Intern *Jupiter, FL*

- Researched and developed mid-price prediction methods using Python and C++, implementing feature engineering and various statistical models and deep learning frameworks, resulting in a 15% improvement in predictive accuracy.
- Modeled order book data comprising over 10 million data points using advanced stochastic processes and cadlag functions, conducting in-depth statistical/time series analysis that identified 5 key market inefficiencies for mid-price prediction.
- Developed a visualizer to evaluate quantitative trading strategies, aiding in alpha discovery.

CS 2110 - Computer Organization and Programming **January 2023 – May 2023**
Undergraduate Teaching Assistant *Atlanta, GA*

- Instructed 200+ undergraduate students in computer organization fundamentals, including assembly programming, processor architecture, and hardware-software interface concepts.
- Created and evaluated LC-3 assembly programming assignments and projects, providing comprehensive feedback that improved student understanding of low-level programming by 25%.
- Developed tutorial materials for debugging assembly code and understanding computer architecture concepts, reducing average student support response time by 40%.

Siemens Smart Infrastructure, Siemens Inc. **May 2022 – August 2022**
Software Engineering Intern *Peachtree City, GA*

- Utilized a Long Short-Term Memory (LSTM) deep learning model using TensorFlow and MATLAB to predict the type of signal data emitted from a communication circuit breaker, achieving 99% accuracy.
- Developed the backend for a Bluetooth Low Energy (BLE) device communication program using C++ and PyTorch to compile and analyze the changing BLE frequencies nearby and used React for the front-end/Node.js for the back-end.
- Helped create a version control platform (*code.siemens*) for essential projects running Siemens Inc. using Git CLI and Python.

Publications and Journal Entries

1. **S. Punjwani**, L. Heck, "Weight-of-Thought Reasoning: Exploring Neural Network Weights for Enhanced LLM Reasoning," in *arXiv*, *ACL Proceedings*. arXiv:2504.10646, 2025.
2. **S. Punjwani**, L. Heck, "Large Body Language Models," in *arXiv*, *ACL Proceedings*. arXiv:2410.16533, 2024.
3. **S. Punjwani**, L. Heck, "Allo-AVA: A Large-Scale Multimodal Conversational AI Dataset for Allocentric Avatar Gesture Animation," in *arXiv*. arXiv:2410.16503, 2024.
4. **S. Punjwani**, B. Yang, M. Grimes, L. Heck, "Conversational Gestures: Transforming Text into Full-Body Virtual Interactions," in *SouthNLP Proceedings*, 2024.
5. **S. Punjwani**, L. Heck, M. Gombolay, "Developing an End-to-End Method for Training Real-Time Virtual Agent Systems for Text-to-Action Conversion," in *GT Digital Repository*, Georgia Tech Theses & Dissertations Library, 2023.
6. **S. Punjwani**, L. Zhao, A. Ritter, L. Heck, "Integrating Language Models with Symbolic Reasoning for Enhanced Generalization." *(In Review)*, 2025.
7. **S. Punjwani**, T. Ma, "Adaptive Neural Architecture Search Leveraging Evolutionary Algorithms for Dynamic Environments." *(In Review)*, 2025.
8. B. Reichman, A. Sundar, C. Richardson, T. Zubatiy, P. Chowdhury, A. Shah, J. Truxal, M. Grimes, D. Shah, W. Chase, **S. Punjwani**, A. Jain, L. Heck, "Outside Knowledge Visual Question Answering Version 2.0," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
9. S. Vecchioni, **S. Punjwani**, Y. Ohayon, N. Seeman, "Using Nearest-Neighbor Nets for DNA Structure Stability and Thermodynamic Property Prediction," in *ACS Nano*, 2020.

Technical Skills

Languages: Python, C++, C, Java, JavaScript, C#, VHDL, x86, HTML/CSS, SQL, MATLAB, PHP, Go
Frameworks: NumPy, Statsmodels, CUDA, Docker, Linux, PyTorch, TensorFlow, Pandas, Sklearn, SciPy, OpenCV
Awards: ICPC Finalist, Babbage Comp Prog Gold, USAPhO Gold, Outstanding Undergraduate Research, Pythagoras Undergraduate Mathematics Excellence Award, President's Undergraduate Research Award