Apurva Gandhi

Curriculum Vitae

(510)-579-7655
apurvasgandhi@gmail.com
apga.github.io/
apurvaga

Education

Aug 2024 - Present **Ph.D. Computer Science**, *Carnegie Mellon University*, Pittsburgh, PA
Advisor: Graham Neubig
Research Areas: Multimodal/Embodied Agents, RL, Program Synthesis, Reasoning/Abstraction

Jan 2019 - May 2020 **M.S. Electrical Engineering**, *University of Southern California*, Los Angeles, CA, GPA: 4.0
Graduated concurrently with B.S.; Focus on Machine Learning.

Aug 2016 - May 2020 **B.S. Computer Engineering and Computer Science (CECS)**, *University of Southern California*, Los Angeles, CA, GPA: 3.98, *Summa Cum Laude*

Selected Coursework
Reinforcement Learning, ML Systems, Game Engines, Numerical Methods, Parallel/Distributed Computing, Mathematics of High-Dimensional Data, Neural & Computational Intelligence

Experience

Microsoft Corporation, Mountain View, CA

Sep 2023 - Aug 2024 **Senior Machine Learning Scientist**, *Office AI/Microsoft Copilot*

- Worked as a cross-org architect and tech lead for Microsoft Copilot, leveraging program synthesis techniques to enable scenarios for the product in the areas of **LLM orchestration/planning, natural language commanding and extensibility (plugins, agents, etc.)**
- Worked on parameter-efficient fine-tuning and constrained decoding to make SLMs (e.g. phi-3) more reliable for tool-use and code generation in Copilot.
- Designed and shipped the grounding/context-passing architecture for tool use in Microsoft Copilot's (Sydney) LLM orchestrator using ideas from programming languages.

Jun 2022 - Sep 2023 **Machine Learning Scientist II**, *Office AI*

- **Created the Office Domain Specific Language (ODSL) framework and the original prototype for Office/M365 Copilot.** The ODSL framework lets apps easily create a strongly-typed DSL for the app's commands and use a RAG-based LLM framework to convert natural language commands into DSL plans that can be verified and run on the ODSL interpreter. See [P1] in Publications and [PC1] for a MSFT Ignite talk. **The ODSL prototype grew into what is now called M365 Copilot at Microsoft.**
- Shipped ODSL in PowerPoint, just 3 months after the original ODSL demo, to limited customers to execute on the M365 Copilot vision and establish Microsoft as the first-mover in the space.
- Was a tech lead on the crew that productionized the ODSL framework into a Copilot platform for M365 apps. With the help of my teammates, leadership and cross-org partnerships, **ODSL has grown to support natural language commanding across 25+ apps including all core office apps (Word, PPT, Excel, OneNote, etc.), many M365 apps (Teams, Outlook, etc.), Azure, PowerApps and more.**
- Prior to work on ODSL, I co-led development of foundation models for user behavior intelligence pretrained on more than 60 billion user interactions with Microsoft Office Applications, powering applications such as search, command prediction, RPA, etc.

Microsoft New England Research & Development, Cambridge, MA

Mar 2022 - Jun 2022 **Applied Scientist II**, *AI Rotation Program*

- Collaborated with the Gray Systems Lab (GSL) on the Tensor Query Processor (TQP): a system that compiles SQL queries to tensor programs that can be executed with PyTorch on GPUs, TPUs, etc. leading to massive speedup in query execution (often $>$ 10X speedup).
- Researched and implemented differentiable approximations to SQL operators such as grouped-aggregation and filtered-aggregation. This lets us introduce a new class of *trainable* SQL queries which use SQL syntax to express end-to-end differentiable, neurosymbolic programs that can be trained from query (input, output) examples to answer a query.
- Our work resulted in a **Best Demonstration Award at VLDB 22** [C3] and a **CIDR 2023 paper** [C1]. I also prepared a demo that was presented at Microsoft Build 2023 [PC2].

Aug 2020 - Mar 2022 **Applied Scientist**, *AI Rotation Program*

- Built a prototype surfing video analysis pipeline (pose estimation, object detection and video action recognition) for the **USA Surfing team** and **US Olympic Committee**. This has now manifested into an official partnership between USA Surfing and Microsoft.
- Applied a sequence labeling approach to the problem of identifying and extracting action items/tasks from digitally handwritten notes on canvases such as Microsoft Whiteboard or OneNote. This resulted in an **EMNLP 22 Industry track publication** [C2].

Amazon AI (Amazon Web Services), Seattle, WA

May 2019 - Aug 2019 **Software Development Engineer Intern**

- Worked on Amazon Lex, a low-code chatbot creation service leveraging natural language understanding and slot filling.
- Implemented neural text normalization to help standardize differently styled content from varying modalities (e.g., spoken, written) to a common style before further processing.

Sandia National Laboratories, Albuquerque, NM

July 2018 - May 2019 **Deep Learning R&D Intern**, *Year-Round*

- Built a framework to facilitate easy testing of defenses against adversarial attacks on ML models.
- Prototyped and reproduced the results of popular attacks and defenses including targeted/untargeted, iterative FGSM attacks and GAN-based defenses.

May 2018 - July 2018 **Mathematics Research Intern for Advanced National Security (MARTIANS)**

- Developed a pruning method for convolutional neural networks achieving at least a 40% parameter reduction on models tested while maintaining accuracy.

Publications

Preprints

[P1] Natural Language Commanding via Program Synthesis, **A. Gandhi***. *T. Nguyen, H. Jiao, R. Steen and A. Bhatawdekar*, **arXiv 2023**

Conference Publications

[C1] The Tensor Data Platform: Towards an AI-Centric Database System, **A. Gandhi***. *Y. Asada, V. Fu, A. Gemawat, L. Zhang, R. Sen, C. Curino, J. Camacho-Rodriguez and M. Interlandi*, **CIDR 2023**

[C2] SLATE: A Sequence Labeling Approach to Task Extraction from Free-form Inked Content, **A. Gandhi***, *R. Serrao, B. Fang, G. Antonius, J. Hong, T. M. Nguyen, S. Yi, E. Nosakhare, I. Shaffer, S. Srinivasan and V. Gupta*, **EMNLP 2022 – Industry Track**

[C3] **Share the Tensor Tea: How Databases can Leverage the Machine Learning Ecosystem**, *Y. Asada*, V. Fu*,* **A. Gandhi***, *A. Gemawat*, L. Zhang*, D. He, V. Gupta, E. Nosakhare, D. Banda, R. Sen and M. Interlandi*, **VLDB 2022 – Best Demo Award**

[C4] **Adversarial Perturbations Fool Deepfake Detectors**, **A. Gandhi*** *and S. Jain*, **IJCNN 2020**

Patents

[PA1] **Extending Query Languages with Differentiable Operators**, *M. Interlandi*,* **A. Gandhi***, *Y. Asada, A. Gemawat, V. Fu, L. Zhang, R. Sen and D. Banda*, October 4, 2022, Patent Pending.

[PA2] **Sequence Labeling Task Extraction From Inked Content**, *J. Hong,* **A. Gandhi***, *G. Antonius, T. M. Nguyen, R. Serrao, B. Fang and S. Yi*, November 12, 2021, Patent Pending.

Magazine Articles

[M1] **Deepfakes: Fooling Humans with Artificial Intelligence**, **A. Gandhi**, USC Viterbi Illumin Magazine 2021

Selected Press Coverage of Work

[PC1] **How M365 Copilot Works**, *David Conger*, Microsoft Ignite, Nov 20, 2023

[PC2] **Inside Azure Innovations – Tensor query processing**, *Mark Russinovich*, Microsoft Build, Jun 1, 2023

[PC3] **Introducing Microsoft 365 Copilot – A whole new way to work**, *Colette Stallbaumer*, Microsoft Blog, Mar 16, 2023

[PC4] **Giving the Sport of Surfing its Next Big Break**, Microsoft In Culture, Jun 22, 2022

[PC5] **How Olympic Surfing Is Trying to Ride the Machine Learning Wave**, *Daniela Hernandez*, The Wall Street Journal, July 27, 2021

[PC6] **Fooling Deepfake Detectors**, *Ben Paul*, USC Viterbi School of Engineering, May 4, 2020

Awards and Honors

Nov 2022 **First Place, Microsoft Hack for Cloud – Executive Challenge**
Won a Microsoft-wide hackathon (over 400 projects competing in the Cloud category) for my work on extending SQL for differentiable and multi-modal query processing, winning an opportunity to pitch to Executive leadership. For details check our CIDR 23 paper [C1].

Sept 2022 **VLDB 22 Best Demonstration Award**
Awarded for our VLDB 22 Demo Paper [C3] where we demonstrate TQP: The world's first Tensor-based SQL Query Processor.

Skills

Languages Python, TypeScript, C#, C++, C, HLSL, MATLAB, Verilog, SQL

Frameworks PyTorch, CUDA, ONNX, NodeJS, PySpark, OpenMPI, DirectX