John Speaks

Personal Website/Portfolio: johnspeaks.com
johnspeaksiv@gmail.com | Pittsburgh, PA | linkedin.com/in/jtsiv | github.com/JTSIV1

Education

Carnegie Mellon University Expected 08/2026
Master of Information Technology (Data Analytics Concentration)
University of Illinois Urbana-Champaign 05/2025
Bachelor of Science in Computer Science and Linguistics (Statistics Minor) *Summa Cum Laude (GPA: 4.0)*

Skills

Technical Skills: Machine learning, natural language processing, vector embeddings, Scikit-Learn, databases, AWS, Azure, Terraform, quantum computing with AWS Braket, Power BI, ggplot, Git, REST API, and HPC.
Coding Languages: Python, Java, Go, TypeScript, JavaScript, HCL, R, Dart, C, C++, Kotlin, Haskell, Bash, C#

Work Experience

Amazon Web Services (AWS) Bellevue, WA
Software Development Engineer Intern (AI/ML) – AWS SageMaker Unified Studio *05/2025 – 08/2025*

- Developed infrastructure to enable persistent monitoring, diagnostics, and resource optimization for thousands of customers' ML workloads by reporting enhanced app-level metrics (CPU, memory, etc.) from SageMaker Studio.
- Engineered and deployed new production services in Go and Java, leveraging CloudFormation for automation and internal pipelines for continuous integration.
- Partnered with key stakeholders, including engineers and PMs, to define and deploy infrastructure solutions that directly addressed customer needs for controlling and monitoring ML workflows.

Common Crawl Corpus Research Champaign, IL
Undergraduate Researcher *09/2024 – 05/2025*

- Partnered with Professor Jonathan Dunn to build a Python pipeline that generated a high-quality, diverse global language corpus from the Common Crawl, filling a critical gap in natural language datasets.
- Applied NLP techniques and linguistic heuristics to filter, clean, and classify hundreds of billions of multilingual text samples for large-scale corpus construction.

CyberGIS Research Center Champaign, IL
Software Developer and Undergraduate Researcher *01/2023 – 01/2025*

- Empowered less technical GIS researchers at multiple universities to perform complex spatial analysis on large datasets by putting powerful HPC resources at their fingertips.
- Maintained and extended a JavaScript backend for CyberGIS-Compute, interfacing with the SLURM job scheduler to enable remote job execution via a Jupyter Notebook interface.
- Authored and published open-source documentation for CyberGISX and its API to ensure researchers could easily interface with our servers and interface with HPC resources.

Phillips 66 Houston, TX
Cloud Innovation Intern *05/2024 – 08/2024*

- Refactored and standardized 23 Terraform modules across AWS and Azure, ensuring adherence to company standards and security policies, which eliminated the need for teams to recreate configurations and improved code reuse.
- Developed a dedicated pipeline in Azure DevOps to periodically validate Terraform configurations, proactively catching infrastructure errors before they impacted our customer teams.
- Spearheaded the company's first quantum computing initiative; built a scheduling optimization proof-of-concept using AWS Braket to evaluate its applicability for optimizing the feed and output of refinery operations.

U.S. Department of State - Embassy New Delhi New Delhi, India
Information and Resource Management Intern *06/2023 – 07/2023*

- Designed a SharePoint app and wrote reports to assist offices across the diplomatic mission in properly maintaining, destroying, and archiving sensitive/classified government records to comply with U.S. law and department policies.
- Performed hardware/software installations, replacement, and disassembly on government computer systems, ensuring security of technology across the mission.

General Services Office Intern *07/2022 – 08/2022*

- Organized and processed sensitive data for over 600 embassy properties across several databases and spreadsheets.
- Coordinated with residents, the purchasing office, and contractors for timely cleaning and maintenance services.

Independent Projects

Comparing and Evaluating Synonymy Representations Across World Englishes 05/2025

- Engineered a research pipeline using Python to evaluate synonym representations from traditional resources and contemporary LLMs like GPT-4o, LLaMA3.3, and Deepseek-V3, demonstrating a systematic bias in both toward Inner Circle English norms.
- Trained regional word embeddings using Gensim's Word2Vec model on web-crawled data from 13 countries, enabling data analysis with Jaccard and Cosine similarity of generated synonym sets.

Scrutinizing LLM Bias Towards Groups Protected Under U.S. Employment Law 12/2024

- Investigated biases in the GPT-4o-mini model using the VADER sentiment analyzer on stories generated from biographical prompts in Python, exposing a statistically significant disparity across different protected classes.
- Identified a secondary skew in how the GPT-4o-mini model would assign biographical labels to stories generated with different sentiments reinforcing the same disparity across protected classes.

Exploring Variation Across Four Languages Using Reality TV Captions and Legal Documents 10/2024

- Explored register variation across English, Finnish, Greek, and Portuguese using methodology extended from Li et al. (2022) with legal documents from the MultiEURLEX corpus, reality TV captions, and web crawl data.
- Used Scikit-Learn to analyze linguistic patterns with both Jaccard and Cosine similarity, uncovering patterns with how different languages communicate on the web compared to formal and casual mediums.

Evaluating Gender Biases Using Embedding Spaces on Social Media and Wikipedia 05/2024

- Compared gender bias across online platforms, including Reddit, Twitter, and Wikipedia with both skip-gram and CBOW embeddings on large corpora of data from each platform.
- Created a custom evaluation metric based on cosine similarity to analyze the average closeness of gendered terms to a corpus of positive and negative sentiment words, revealing biases in the CBOW models.

Publications

Michels, Alexander, Zhang, Ian, Padmanabhan, Anand, **Speaks, John**, Vandewalle, Rebecca, and Wang, Shaowen. 2025. *Expanding Access to CyberGIS-Compute through support for Heterogeneous Workflows.* Presented at the I-GUIDE Forum, Chicago, IL, USA

Michels, Alexander, Kotak, Mit, Padmanabhan, Anand, **Speaks, John**, Wang, Shaowen. 2024. *Providing Accessible Software Environments Across Science Gateways and HPC.* Practice and Experience in Advanced Research Computing 2024: Human Powered Computing, 48

Leadership

TEEL Lab: AI Technicians Bootcamp Pittsburgh, PA
Teaching Assistant *08/2025 – Present*

- Designed hands-on coding exercises to teach students Python programming, data analytics, CloudOps, DevOps, and AI development from scratch within 12 months.
- Lead discussions, presented content, and hosted office hours to support student learning throughout the week.

League of Linguists Champaign, IL
President and Undergraduate Representative *08/2023 – 05/2025*

- Lead and organized club functions like professional workshops, research talks, and meetings with career professionals, growing attendance and engagement by 70% in my time as President.
- Elected representative to the Linguistics Department Diversity and Inclusion Committee and All-Faculty Council where I represented undergraduates' views on program changes and departmental policies.

'Jila' - An Educational App for Q'anjob'al Speakers Champaign, IL
Backend Tech Lead *09/2024 – 12/2024*

- Coordinated with project managers and other teams to create an accessibility tool for hundreds of local Q'anjob'al speakers in the community struggling to access public resources.
- Managed the backend team of 5 to organize data hosting and administrator authentication with Vercel, Clerk, and REST API for our ReactJS app.