

JOHN BOLOGNINO

+1(203) 644-0805 ◊ jcbolo72012@gmail.com ◊ github.com/jcbolo72012 ◊ [linkedin.com/in/john-bolognino](https://www.linkedin.com/in/john-bolognino)

EDUCATION

Stanford University Artificial Intelligence Professional Program (Non-Degree)	June 2026
Boston University, Kilachand Honors College	May 2023
<i>Bachelor of Arts, Computer Science, Psychology Minor</i>	GPA: 3.73
Honors: Upsilon Pi Epsilon Computing Honor Society, Dean's List all semesters, Undergraduate Research Award (3x)	

SKILLS

Languages: Python, Go, Ruby, SQL, JS; **Familiar:** C, R, Matlab
ML/Data: PyTorch, HuggingFace Transformers, FastAI, CUDA, NumPy, Pandas, Model Distillation, Real-Time Inference
MLOps/Infra: AWS SageMaker, S3, Lambda, EC2, Docker, Colima, Datadog, Git

PROFESSIONAL EXPERIENCE

Savvas Learning Company	Boston, MA
Software Developer	<i>December 2024 - Present</i>
Associate Software Developer	<i>June 2023 - December 2024</i>

- Designed and trained transformer-based NLP models with HuggingFace and PyTorch, including a custom multi-headed architecture for automated essay scoring in a real-time production system, resulting in a 75% reduction in AWS inference costs.
- Led model and infrastructure upgrades that reduced average real-time inference latency by 50% (800ms to 400ms per batched request), migrating AWS SageMaker endpoints from Inf1 to Inf2 and upgrading base models from BERT-base to DeBERTa-v3.
- Distilled behaviors from frontier large language models (Claude Haiku and Sonnet) into smaller transformer models to balance scoring quality, latency, and deployment cost.
- Deployed and served containerized models via AWS SageMaker real-time endpoints, supporting low-latency inference for an educational platform used by over 10,000 school districts.
- Established production monitoring and alerting using Datadog, tracking endpoint health, latency, and model performance
- Managed versioning, evaluation, and iterative improvement of all production ML models, incorporating feedback from subject-matter expert (teacher) raters into continuous model refinement.
- Spearheaded and organized monthly paper club for team of 12, discussing recently published academic literature in machine learning and natural language processing as they relate to our development objectives.

Potoo Solutions	Norwalk, CT
Python Developer/IT Team Member	<i>May 2019 - June 2023</i>

- Developed NLP pipeline using Python libraries NLTK and Spacy to predict consumer satisfaction with a product and the likelihood that it is being sold by an illicit third party using eCommerce data including product reviews. Utilized Tableau to create interactive dashboards to communicate these insights to clients.
- Refactored core ETL processes in Python and Golang with asynchronous design to speed database update jobs by 300% and scale to handle a 175% throughput increase as well as bespoke ETL tools for clients such as Adidas and VF. Leveraged AWS Lambda and S3
- Developed automation to gather e-commerce metrics for over 600 clients with accompanying GUI made with tkinter for account managers to access data from vendor APIs and internal PostGRES database, streamlining workflow by over 1,500 hours yearly

RESEARCH EXPERIENCE

Learning, Intelligence & Signal Processing lab at Dartmouth College, Safra Center for Brain Science	Hanover, NH
Researcher	<i>January 2022 - Present</i>

- Applied transformer-based models to decode handwritten language from recorded motor cortex neural activity incorporating signal preprocessing via filtering.
- Built a novel three-dimensional neural activity dataset and used Grad-CAM to identify cortical regions critical for character-level classification.
- Presented results at *International Conference for Machine Vision (ICMV)* 2024.

Visual Information Processing Lab at Boston University	Boston, MA
Undergraduate Researcher	<i>June 2021 - January 2022</i>

- Co-authored a [paper](#) documenting a benchmark fisheye Re-Identification dataset with annotations accepted to *IEEE Conference on Advanced Video and Signal-Based Surveillance (AVSS)* November 2022
- Three-times awarded Undergraduate Research Opportunity Program funding to research adaptation of person re-identification algorithms for use on fisheye images
- Gathered benchmark performance statistics for a novel indoor, multi-camera fisheye dataset for person re-identification on eight state-of-the-art deep learning re-identification algorithms designed using Pytorch, TensorFlow, and FastAI

PROJECTS

Electrode-Specific Masking in EEG-to-Text Translation: November 2025 - Designed model pretraining methodology leveraging region-specific features of EEG data and modern transformer architectures to achieve a 39% increase in training speed over existing methods for EEG-to-Text Translation on the ZuCo dataset. Whitepaper currently in review through Stanford Professional Program