

Simon Benigeri

✉ simon.benigeri@northwestern.edu simonbenigeri.com [in simon-benigeri](#) [simon-benigeri](#)

Summary

Ph.D. Candidate in Artificial Intelligence at Northwestern University, specializing in Natural Language Processing (NLP) and conversational AI. I bring 3 years of AI research experience in cardiology and 3 years of AI industry experience at IBM. Most recently, I interned on the **Siri Speech, Innovation** team at **Apple** (Summer 2026), where I added active learning workflows to an automatic prompt-learning framework used by ML practitioners across language and speech tasks. Previously, I interned at **Samsung Research America, AI Center** on the **Language & Personal Intelligence Team** (Summer 2025), where I developed a novel method to improve the reasoning capabilities of small LLMs.

Education

Northwestern University *PhD in Artificial Intelligence* *Sept 2023 – present*

- Advisor: [Larry Birnbaum](#) [✉](#)
- **Research focus:** Natural Language Processing, Conversational AI

Northwestern University *MS in Artificial Intelligence* *Sept 2020 – Dec 2021*

- GPA: 3.98/4.0

University of Pennsylvania *BA in Mathematics* *Sept 2010 – June 2016*

Experience

Graduate Research Assistant - PhD in AI *Evanston, IL*
Northwestern University *Sept 2023 – present*

- Conducting research to develop practical conversational AI systems, focusing on enhancing reliability in real-world applications and robustness to complex but natural human conversational behavior.
- Published a paper at EMNLP 2026 Findings, focusing on robustness and reasoning in conversational AI.
- Teaching Assistant for CS 337: Natural Language Processing (Fall 2024, Fall 2025), guiding students in foundational and advanced concepts of NLP.

PhD Intern - Siri Speech, Innovation *Cupertino, CA*
Apple *June 2026 – Sept 2026*

- Added active learning workflows to an automatic prompt-learning framework used by ML practitioners across language and speech tasks.
- Contributed to training text-to-speech quality judges used to evaluate Siri voice customization.

NLP and AI Research Intern *Mountain View, CA*
Samsung Research America, AI Center, Language & Personal Intelligence Team *June 2025 – Sept 2025*

- Developed a novel method to improve the reasoning capabilities of small large language models (LLMs).
- Research paper and patent application are in progress.

AI Research Engineer - Cardiology *Chicago, IL*
Bluhm Cardiovascular Institute Center for AI, Northwestern University *Feb 2022 – present*
(Transitioned to part-time in Sep 2023)

- Designed end-to-end agentic cardiovascular care architecture (clinical, safety, deployment layers) culminating in ARPA-H ADVOCATE grant submission.
- Collaborated with physician-scientists to develop AI/ML solutions for cardiovascular disease, focusing on heart failure and multimodal data integration.
- Led the creation of high-quality cardiac amyloidosis datasets, increasing data quality from 3,000 weakly labeled to 700 gold-standard cases, leading to multiple publications and grant funding.
- Built predictive models for heart failure subtypes and mortality risk using clinical text and structured data, contributing to the team's published research.

- Automated echocardiogram processing with startup Ultromics to enhance diagnostics for heart failure and cardiac amyloidosis, with plans for publication.
- Developed HIPAA-compliant data de-identification tools for DICOM images and clinical notes, ensuring data security and reducing processing time by 75% compared to previous de-identification solutions.

Senior Technical Consultant - AI

IBM

Paris, France

Feb 2018 – Sep 2020

- Developed NLP-driven AI systems for enterprise applications (banking compliance, HR, customer service, IT support, construction), integrating machine learning with business logic in collaboration with clients.
- Led the end-to-end model development lifecycle, from data collection and annotation to evaluation, deployment, and monitoring; developed application back-ends.
- Achieved significant impact, including reducing manual effort in analyzing annual reports and adverse news for KYC compliance by 50%, as highlighted in [Global Trade Review](#).
- Supported customer complaints for the French National Railway Company (SNCF) during strikes via the chatbot [Tout Oui](#), as referenced in [Ouest-France](#).

Publications

Characterizing and assessing robust conversational interaction: Amendments in order specification (EMNLP 2026 Findings)	2026
Automated identification of heart failure with reduced ejection fraction using deep learning-based natural language processing (JACC: Heart Failure, 2025)	2025
A Flash in the Pan: Better Prompting Strategies to Deploy Out-of-the-Box LLMs as Conversational Recommendation Systems (ACL: COLING 2025)	2025
Evaluating the performance and potential bias of predictive models for the detection of transthyretin cardiac amyloidosis (JACC: Advances, 2025)	2025
Multi-domain Summarization from Leaderboards to Practice: Re-examining Automatic and Human Evaluation (ACL: GEM 2023, GitHub)	2023

Projects

Probing LLM planning and reasoning by playing 20 questions	GitHub
◦ Built a backend to evaluate LLM reasoning for the game 20 questions; implemented budget forcing (S1: Simple test-time scaling) to improve reasoning performance.	
CalendarBot	GitHub
◦ Built a Retrieval-Augmented Generation (RAG) chatbot for calendar event retrieval using a hybrid approach: date-based retrieval followed by vector-based retrieval.	
Protecting forest elephants through audio monitoring	Blog post
Deep reinforcement learning agent for stock trading	GitHub
LSTM language model in PyTorch	GitHub

Skills

Languages: Python, SQL, Bash

Technologies: PyTorch, TensorFlow, Hugging Face Transformers, NLTK, SpaCy, NumPy, pandas, scikit-learn, FastAPI, Docker, Git, AWS, Google Cloud Platform (GCP), Azure, Relational Databases, Vector Databases

Research and collaboration: Research writing (ACL, JACC), effective communication with cross-functional teams, presenting research findings, communicating technical concepts to diverse audiences, solving open problems in collaborative environments