André Natal

Address: San Jose, CA, USA
Email address: anatal@gmail.com
Web: andrenatal.com/projects
Github: github.com/andrenatal/
Linkedin: linkedin.com/in/andrenatal/

Work Experience

04/2025 – present
San Jose, USA

Independent consultant

- Building real-time edge-based computer vision systems for recognition, tracking, detection, and pattern estimation by using multiple techniques like embeddings, faceprint, segmentation, and custom deep learning based classification models using Rust and ONNX runtimes by utilizing customized fine-tuned YOLO models.
- Developed a high-performance interruptible speech-to-speech multilingual conversational system in Rust using customized SLMs distilled from LLMs as a backbone for dialog management, turn, and intent detection. Tied with speech interfaces using encoder-decoder transformer models for speech recognition, like Whisper and Conformers, and generative TTS using diffusion and adversarial training with large speech language models, it delivered real-time conversational capabilities over IVRs.
- Developed a Rust agentic system using LLMs over data buckets in order to search for patterns on unstructured data (text, audio, and video) to be used as an aid in the threat intelligence process on cybersecurity teams and also for market-moving news on finance and alpha seekers.

12/2023 – 04/2025
San Francisco Bay Area

Principal AI and ML Engineer
MinIO

- Designed, developed, and implemented a reactive agentic LLM-based system that reasons and operates over data lake changes upon open-ended instructions given in natural language.
- Developed a speech enabled conversational system using custom whisper models and retrieval augmented generation to allow users to chat with their buckets of unstructured data using Rust as the backend system and Webrtc as its communication stack.
- Developed a sentiment analysis system with the intent of determining and scoring the temperature of thousands of chat conversations between technical support and customers, aiming to classify the customers with a higher risk of attrition.

07/2023 – 12/2023
Palo Alto, California, United States

Principal AI and Machine Learning Engineer
Gather

- Built RAG pipelines utilizing neural information retrieval techniques using ColBERT.
- Fine-tuning foundation models like Mistral, Llama, and BERT for multiple NLP problems like chat, classification, text generation, and NLU.
- Designed and implemented data ingestion pipelines using the user's personal data in order to build query-based chatbot systems.

04/2015 – 07/2023
San Francisco Bay Area

Staff Machine Learning Engineer
Mozilla

- **Firefox Translations:** Developed a web extension for secure, client-side translation that performs neural machine translation locally on the device using the Bergamot engine, eliminating the need for cloud-based processing.
- **Project Bergamot:** Contributed to the research and development of a set of neural machine translation tools in coordination with the Bergamot Project Consortium to enable offline translation capabilities within the browser.
- **Common Voice:** Worked on the open-source initiative to build a large-scale, public-domain voice database used for training and benchmarking speech recognition models.
- **Voice Fill:** Engineered a Firefox feature that integrated speech recognition technology directly into web forms and search engines, allowing for voice-based text input and navigation.
- **Firefox Reality:** Developed a standalone browser for VR/AR headsets (Oculus Go, Viveport,

Daydream) that supports voice commands, hand gestures, and 3D web content interaction.
- **Mozilla Web of Things (WoT):** Built an open-source framework and platform for smart device interoperability, focusing on using open standards to monitor and control connected devices.
- **Web Speech API:** Implemented the SpeechRecognition and SpeechSynthesis interfaces to provide standardized voice input and output capabilities for web applications.
- **Firefox OS and Vaani:** Architected Project Vaani, an open-source voice assistant for the Firefox OS platform designed to provide an extensible, web-based alternative to proprietary voice assistants.

01/2013 – 11/2014
Sao Paulo and Barcelona

Speech Software Engineer
Verbio

- **High-Volume IVR Infrastructure:** Architected and maintained the municipal IVR system for the city of São Paulo, managing high-availability infrastructure that processed nearly one million calls per month.
- **NLP-Enabled IVR:** Engineered speech-enabled IVR systems utilizing Natural Language Processing (NLP) to replace traditional touch-tone menus with conversational user interfaces.
- **Mobile Virtual Assistants:** Developed context-aware mobile assistants integrating speech recognition and NLP to automate complex user tasks on smartphone platforms.
- **Android Speech SDKs:** Designed and built specialized SDKs for the Android ecosystem, enabling third-party developers to integrate advanced speech features into their own applications.
- **Voice Biometric Authentication:** Developed iOS applications that utilized voice biometrics for secure identity verification, focusing on unique vocal characteristic extraction and matching.
- **Mobile VAD Algorithms:** Researched and implemented optimized Voice Activity Detection (VAD) algorithms specifically for mobile hardware to improve processing efficiency and battery life.
- **Pre-Sales Engineering:** Supported the sales lifecycle by providing technical architectural guidance and building functional prototypes to demonstrate the feasibility of speech-tech solutions.

12/2010 – 04/2015
São Paulo Area, Brazil

Founding Engineer
All Apps

- **Speech-Enabled IVR:** Developed a full-stack IVR solution integrating Asterisk, UniMRCP, and Kaldi to handle millions of monthly calls using automated speech recognition.
- **Asics Perfect Pace:** Lead developer for a specialized iOS application that received over 20 international awards, including the Cannes Lions, for its technical implementation.
- **Cross-Platform Development:** Architected and deployed applications for iOS, Android, Windows Phone, and FirefoxOS, utilizing real-time speech processing and geo-location.
- **Multi-Device Ecosystems:** Designed software for diverse form factors including Smart TVs, tablets, and wearables, focusing on speech interfaces and hardware integration.

01/2000 – 12/2012

Additional experience as an IC

- **UBS:** Engineered FIX-compliant high-frequency order routing, algorithmic derivative trading solutions, and real-time quote APIs with messaging-based risk system integration.
- **SystemGPS:** Architected and developed a technical suite including speech-enabled IVRs, SMS chatbots, and real-time web-based tracking systems integrated with GPS vehicle hardware.
- **Credit Suisse, Hedging-Griffo:** Developed low-latency, multithreaded C++ trading systems, migrating legacy DDE spreadsheets to high-frequency FIX-compliant exchange connectivity.
- **CMA**: Engineered the firm's first derivative algorithmic trading solution and a comprehensive real-time web-based trading platform for HSBC integrating live market data and analytics.
- **Bovespa (via 7Comm):** Developed a collateral loan exchange API enabling seamless integration between brokerage order management systems and the public exchange.
- **Globo.com:** Developed a certified Windows Media Server security addon for unauthorized access prevention and stream health monitoring tools using the WMS SDK.
- **Microsoft Consulting Services:** Architected and executed the end-to-end migration of

Work Experience

legacy judicial systems to a modernized, web-based .NET infrastructure.

- **Bechtel Corporation:** Developed Windows CE mobile applications for automated industrial data collection and real-time processing of routine unit tests within an oil & gas facility.
- **Webmotors:** Served as an early-stage full-stack engineer developing web and desktop features integrated with large-scale database architectures.
- **MTV Brazil:** Engineered award-winning interactive TV systems and early day streaming media under high-bandwidth constraints.

Education

09/2024 – present
Palo Alto, CA, USA

Artificial Intelligence Graduate Certificate
Stanford University

- CS224R - Deep Reinforcement Learning
- CS224S - Spoken Language Processing
- CS230 - Deep Learning
- in progress

01/2020 – 12/2022
Palo Alto, CA, USA

Artificial Intelligence Professional Program
Stanford University

- XCS224N – Natural Language Processing with Deep Learning.
- XCS224U – Natural Language Understanding with Deep Learning.
- XCS224W – Machine Learning with Graphs.

01/2013 – 12/2015
São Paulo, Brazil

Computer Science | Bachelor's degree
Universidade Anhembi Morumbi

Certificates

- **NVIDIA:** Building Agentic AI Applications with LLMs
- **NVIDIA:** Generative AI with Diffusion Models
- **NVIDIA:** NVIDIA Certified Associate: Generative AI LLMs
- **NVIDIA:** NVIDIA Certified Associate: Generative AI Multimodal
- **NVIDIA:** Accelerated Computing in CUDA C/C++
- **DeepLearning.AI:** Deep Learning Specialization (as part of Stanford's CS230)
- **Google:** Google Summer of Code Alumni
- **Microsoft:** Microsoft Certified Professional
- **Founder University:** Cohort 19

Projects & Awards

Full detailed list on:

https://andrenatal.com/projects/
https://andrenatal.com/media
https://www.linkedin.com/in/andrenatal