R. P. Ruiz

Seeking Founding, Senior Architect, R&D or Research positions working with Agentic AI

SUMMARY: I'm a creative, highly productive, and perpetual autodidact with twenty years of significant and broad cutting edge software R&D experience. I enjoy prototyping new products and services utilizing AI, LLMs, Deep Learning, Chatbots, Agents, and NLP. I am especially interested in Agentic AI as Physical & Cognitive Prosthesis.

EXPERIENCE:

Senior Generative AI Field Solutions Architect @ Google: 2024

I help Google's customers get down off the fence and commit to AI powered solutions by creatively integrating Google's generative AI services while focusing on amplifying customer impact and business value.

- Presented "Easy PEFT" utility from my "*Collection of Small Agents*" (CoSA) project for automating the creation and fine-tuning of Low Rank Adapters (LoRA)
- Presented my work on CoSA outlining how Google could slash inference costs and improve performance by using semantic caching
- Created, demoed and blogged about multiple reusable assets for document extraction and evaluating extraction accuracy

***Results**: Using Generative AI, I helped Equifax to reduce a five hour process to just five minutes, saving them millions of dollars.*

Founder & Senior AI Architect @ Deepily.ai: 2023

I founded Deepily.ai and created "Genie-in-the-Box", a voice-driven UX that's an extensible platform utilizing highly performant AI agents that convert your voice queries and commands into actionable speech and text.

I re-architected Deepily.ai's entire infrastructure, migrating away from OpenAI's GPT models to one based entirely on highly optimized and fine-tuned open source LLMs and Deep Learning models hosted on Deepily.ai's development servers.

***Results**: Development is faster, cheaper, and performant • Fine tuned models are 99.9% accurate • I am now 2x-3x more productive than before I created and began using Genie in the Box.*

Senior Architect & Data Scientist @ HelioCampus: 2019 to 2023

According to CEO Darren Catalano, I helped transform the Data Science group into a "small, but mighty" example of tech enablement for the entire organization to follow.

- Architect: Developed AMPE (Abstracted Modeling and Prediction Engine), a framework that facilitated code reusability and repeatability, simplifying ML pipelines.

- Architect: Composed process and orchestrated utilization of Docker containers from development through production stages.

Senior Data Science Engineer @ Department of Labor (Appteon): 2018 – 2019

I made my colleagues more productive through tool chain analysis and tech enablement.

- Refactored a key DoL report generator, reducing the Spark application size by 60% (from 900K to 360K of source code) and improving runtime performance by over 300%.
- Created an interactive Spark shell utility in Scala, enhancing report query development productivity by 20-100x.

Senior Data Scientist @ TransVoyant: 2016 – 2017 | Washington D.C.

*I increased the accuracy of TransVoyant's shipping logistics predictions by **365%**.*

- Re-engineered the Data Science team's workflow by implementing a polyglot toolchain (Scala, Python, R) using Apache Spark for notebook-based development, boosting iterative build & run times by 20x-200x.

Senior Researcher @ Comcast Data Science: 2013 – 2015 | Washington D.C.

I made the Data Science Research group's great work visible, demonstrable, and interactive

- Prototyped, integrated and demo'd new products and services utilizing Machine Learning, Natural Language Processing (NLP), Voice Recognition, Video Content Search & Analysis plus Dynamic Sports Event Metadata.

Senior Technology Leader @ Marriott International: 2012 – 2013 | Washington D.C.

- *I helped move Marriott from a desktop-centric web experience to a mobile centric view*

Founder and President @ Valeso: 2006 – 2011 | Washington D.C.

I made industrial-strength standards-based cryptographic security easy and transparent.

- Conceived and created AutonomyCentral, a cryptographically secure platform for spam-free email in any language, secure password and file storage, implemented in Java across multiple operating systems.
- Collaborated with Front Line and Tactical Technology to translate AutonomyCentral into Spanish, Russian, French, and Arabic, resulting in the "VaultletSuite 2 Go" inclusion in the "Security in a box" compilation.
- Conducted numerous Internet Security workshops globally for Human Rights defenders, citizen journalists, and activists, catering to diverse international and multilingual contexts.

EDUCATION

- Masters (MA) in Applied Linguistics & Foreign Languages from West Virginia University
- Bachelors of Arts (BA) in Applied Linguistics & Foreign Languages from West Virginia University

PROGRAMMING LANGUAGES

Python · JavaScript · Scala · SQL

SPOKEN LANGUAGES

- Fully bilingual: I speak English and Spanish with native proficiency

CERTIFICATIONS

- Google Cloud: Professional Machine Learning Engineer · 2024
- Generative AI with large language models, Coursera · 2023 (1 course)
- Natural Language Processing, Coursera · 2023 (4 courses)
- Statistics with Python Specialization, Coursera · 2020 – 2021 (Courses: 1, 2 & 3)
- Deep Learning Specialization: Deeplearning.ai, Coursera · 2017 – 2018 (5 courses)
- Functional Programming in Scala: École Polytechnique Fédérale de Lausanne (ÉPFL) Coursera · 2016 - 2017 (4 courses: 1, 2, 3 & 4)
- Data Science and Engineering with Spark and Python: Berkeley, edX · 2016 (Courses: 1, 2 & 3)
  Machine Learning: University of Washington (UW), Coursera · 2016 (Courses: 1, 2, 3 & 4)
  Data Sciences in R: Johns Hopkins (JHU) Coursera · 2015 – 2016 (Courses: 1, 2, 3, 4, 5, 6, 7 & 8)
- Related Coursework (Python, R, & Statistics): Coursera and Edx · 2015 – 2016 (Courses: 1, 2, 3 & 4)

CONCEPTS, LIBRARIES, MODELS, TECHNIQUES & KEY WORDS

Artificial Intelligence (AI) · Artificial Generative Intelligence (AGI) · Deep Learning · Neural Networks · Natural Language Processing (NLP) · Data Science · Machine Learning (ML) · Large Language Models (LLMs) · Prompt Engineering · Fine Tuning · Quantization · PEFT · QLoRA · LoRA · AWQ · WandB · Whisper.ai · ChatGPT 3.5 & 4.0 · OpenAI's API · Phind-CodeLlama-34B-v2 · Mistral-7B · Mixtral-8x7B · Whisper · Distil-Whisper · Groq · Google Gemini · Coqui text-to-speech · Text Embeddings · Semantic Similarity · Computational Similarity · Open Interpreter · Super AGI · AutoGPT · LangChain Agents & Chat bots · LangChain Question & Answering (GQA) · OpenLLM (Inference Server) · GitHub Copilot · Cursor.ai · Codeium · Replit · Hugging Face · Transformers · Google Collab · Dall-E · Pycharm · JupyterLab · Google Speech to Text API · Docker · CUDA · PyTorch · Jax · TensorFlow · Keras · RNNs · LSTM · CNNs · GPUs · Sentiment Analysis · Document Classification · TF-IDF · Classification & Clustering · Logistic & Linear Regression · Pandas · SMOTE · NLTK · NLP · GloVe · Word2Vec · openai-whisper · scikit-learn · Pandas · SHAP · spaCy · Gensim · XGBoost · LightGBM · GGPlot2 · Linear Optimization · Python · R · Scala · Java · Apache Spark · GitHub · RStudio · Shiny · Matplotlib · Plotly · Databricks · Domino Data Lab · Tableau · Apache Spark · Parallel Computing · Distributed Computing · Performance Analysis and Optimization · Code Refactoring · SQL · MySQL · Applied Linguistics · Open Source · Linux · Agile Software Development