Vinícius Conte Turani

Date of Birth: May 20, 2003
Age: 21 years old
Address: Porto Alegre, RS, Brazil
Contact: +55 54 99617-3830

E-mail: viniciuscturani@hotmail.com
LinkedIn: linkedin.com/vinícius-c-turani
GitHub: github.com/ViniTurani

PROFILE

I am seeking challenges in artificial intelligence and deep learning, where I can expand my knowledge. I aim to absorb as much as possible while contributing to innovative and advanced projects. In the future, I see myself in leadership positions, coordinating a team through innovation projects

---

SUMMARY OF QUALIFICATIONS

- Fluent in English.
- Experienced with programming languages: Python, Java, C, C++.
- Experienced with NLP, LLMs, and Optimization/Quantization.
- Proficient with libraries and frameworks: PyTorch, TensorRT, NumPy, FastAPI, Pytest DeepEval, Prompt Engineering, and others.
- Skilled in SQL, MongoDB, and Milvus – Zilliz vector database.
- Responsible, creative, determined, persistent, empathetic, and a good team player.
- Experienced with the Agile Scrum methodology.

EDUCATION

Bachelor's in Computer Science - Pontifical Catholic University of Rio Grande do Sul (PUCRS) in Brazil
(Currently in 8th semester) *Expected Completion: 2025/2*

NATIONALITY

- **Brazilian –** My current location.
- **European/Italian.**

PROFESSIONAL EXPERIENCE

Teia Labs, Porto Alegre, RS, Brazil:

- **AI Software Engineer** - *March 2025 – Present*

  - Working full-time on AI infrastructure and machine learning applications.
  - Designed and maintained microservices for scalable AI systems (e.g., completions, OpenAI API integrations, data persistence, and Retrieval-Augmented Generation pipelines).
  - Built LLM solutions using AI agents, hybrid search, and reranking techniques.
  - Contributed to prompt engineering and LLM customization for enterprise solutions (e.g., Salesforce-integrated systems like Athena).
  - Applied software architecture best practices in backend development.
  - Improved RAG pipeline retrieval through prompt tuning and evaluation with DeepEval.
  - Used Agents to improve the initial enterprise quality results.
  - **Tech stack:** Python, MongoDB, Milvus, AWS, FastAPI, DeepEval, Hugging Face.

- **Data Science Intern** - *September 2024 – March 2025*

  - Supported backend and ML model integration into microservices.
  - Participated in research and development of LLM-driven features.
  - Contributed to containerized infrastructure with Docker and AWS.
  - Enhanced Linux/Unix scripting and operational skills.
  - **Tech stack:** Python, Docker, Linux, AWS, LLMs, Data Science.

PROJECTS AT TEIALABS:

- **Conversational AI Platform for OSF – (AllAI platform)**

  - Contributed to "**Athena"**, a conversational AI assistant platform supporting chat threads and memory, enabling seamless user interaction via custom prompts and tools.
  - Developed and improved "**MeltingFace"**, a model-agnostic prompt-completion backend that supports various LLM integrations and enables tool use in completions via APIs.
  - Built and maintained modules for "**Datasources"**, facilitating vector-based information retrieval using Atlas (MongoDB) and Milvus for semantic search. For this one, I've implemented the entire hybrid search, with Milvus, the reranking strategy, and based on the evals from the "**Specializations"**, improved the RAG pipeline.
  - The "**Specializations"** is an internal application on DeepEval that evaluates the agents' accuracy and other metrics. It is used as a baseline. I've helped to develop

this application from the beginning, including how to evaluate a RAG pipeline, the retrieval of the vector database, and the model tool calling parameters.
  - Wrote and maintained unit tests using pytest, followed best practices (e.g., python.analysis.typeCheckingMode, security checks, linters), and ensured high code quality through GitHub PR reviews.
  - Deployed microservices to **AWS ECS** via containerization pipelines, integrating with **FastAPI**, **Pydantic**, and **OpenAI** APIs.

- **A POC of an RTC application to help OSF employees learn faster, Salesforce Development**

  A local Python-based Proof of Concept application that leverages OpenAI's Realtime API to enhance the onboarding experience for new Salesforce developers at OSF. The application facilitates real-time, interactive learning sessions, enabling trainees to engage in dynamic conversations with an AI assistant tailored to Salesforce development topics.

  - OpenAI Realtime API Integration: Utilized OpenAI's Realtime API to establish a WebRTC connection, allowing for seamless, low-latency communication between users and the AI assistant.
  - Interactive Learning Sessions: Implemented real-time transcription and response generation, enabling users to receive immediate feedback and guidance on Salesforce development queries for faster communication with the model.

Software Innovation Laboratory (LIS) – HP Inc./PUCRS, Porto Alegre, RS, Brazil:

- **Software Development Intern** - *March 2023 – September 2024*
  - Engaged in multiple projects focusing on machine learning model training/fine-tuning and infrastructure development, including chatbot creation using LLMs.

PROJECTS AT LIS:

1. **Machine Learning Project – Image Style Transfer**

   Developed a deep learning-based solution to transform enterprise logos into stylized, hand-drawn images, mimicking the sketch aesthetic of the Excalidraw platform. The goal was to allow users to import their logos and seamlessly edit them within Excalidraw, preserving their visual identity and flexibility.

   - Created a custom dataset by scraping and curating over 3,000 images from Excalidraw's open libraries and the web, using crawlers. The final dataset included ~500 unique logos manually filtered and categorized by fill style (hashed vs. solid).
   - **Developed, trained, and fine-tuned** U-Net-based models for style transfer, with various encoder-decoder architectures (e.g., DenseNet, ResNeXt) to learn the hand-drawn style from limited data.

- Preprocessed and augmented data using custom pipelines (color normalization, batch-wise transformations, image augmentations) to address dataset imbalance and overfitting.
- Analyzed multiple architectural strategies, including experimenting with different encoder/decoder pairs to optimize for blurry line issues and improve output sharpness.
- Evaluated limitations of the deep learning approach, including style inconsistency and failure to properly learn hashed filling with small datasets.
- Explored classical computer vision alternatives, building an algorithmic pipeline (color quantization + edge detection + texture overlay) that outperformed ML in style fidelity, though lacked user customization flexibility.
- Assessed and compared results qualitatively across all strategies (ML, CV, hybrid) to determine trade-offs between automation and user control.
- **Technologies:** Python, PyTorch, OpenCV, Weights & Biases. U-Net, DenseNet, ResNeXt

Although the ML-based pipeline could stylize logos with some success, its results were ultimately limited by data scarcity and inconsistent line rendering. These findings informed a pivot toward a user-centered solution implemented as a web app (hosted via GitHub Pages), offering customizable vectorization and Excalidraw-ready exports, maximizing user flexibility while leveraging insights from the ML experiments.

2. **Infrastructure Project – GPU Recommendation System**
    - Developed the backend for an application recommending GPUs based on user requirements.
    - **Technologies:** Python, FastAPI, SQLAlchemy, Pytest, Juicify.
3. **Machine Learning Project – Local LLM Chatbot**
    - Developed a modular and privacy-focused chatbot platform using local LLMs to support HP's AI Studio users.
    - Designed and implemented custom components (e.g., tokenizer, vector DB, memory, embeddings, prompt template) inspired by LangChain but fully developed in-house to optimize control and customization.
    - Integrated Retrieval-Augmented Generation (RAG) for accurate document-grounded responses and applied prompt engineering to enhance interaction quality.
    - Contributed to performance optimization and model loading using TensorRT and LlamaCpp with quantized LLaMA2 models (7B).
    - **Technologies**: Python, PyTorch, TensorRT, LangChain (concepts), HuggingFace, ChromaDB, LlamaCpp, PyMuPDF.

SENAI – Moveis Foscarini, RS Brazil

- **Apprentice with Emphasis on Carpentry and Administration** - *February 2018 – June 2019*

- Engaged in a comprehensive apprenticeship program focusing on carpentry and administration.
- Learned technical drawing, production organization, cost formation, and manufacturing processes.

LANGUAGES

English: Fluent

- Studied at Wizard Language School (2010–2018)
- Continued studies at Yazigi International (2019–2021)
- Total of 11 years of language education.

Portuguese: Native

COMPUTER SKILLS AND PROGRAMMING LANGUAGES

- Professional experience with:
  - Python, TensorRT, PyTorch, FastAPI, Flask, Pytest, HuggingFace, SpaCy, LangChain.
  - LLMs and Deep Learning.
  - LLM Quantization and Optimization using TensorRT.
  - U-Net Fine Tuning using PyTorch.
  - NLP, Prompt Engineering, and NLP frameworks like SpaCy.
  - RESTful APIs.
  - Langchain, ChromaDB vector database
  - MongoDB database tool.
  - Milvus - Zilliz vector database
  - SQL and SQLAlchemy.
  - DeepEval and Prompt Engineering.
  - Git and GitHub.
  - Linux Terminal – Bashrc, etc.
  - AWS (Lambda, general deployment, infrastructure).
  - Docker and containerized applications.
- Proficient in Java programming language.
- Familiarity with C and C++ programming languages.

COURSES

- Computer and Motherboard Maintenance - Flexxo Center, IT Training, 60 hours (2019).

- Programming for Beginners and Advanced – Udemy, 18 hours (2020).
- Algorithms and Programming Logic from Basic to Advanced – Udemy, 33 hours (2022).
- The Complete Database and SQL Course – Udemy, 38 hours (2023).
- CS229: Machine learning (Andrew Ng - Autumn 2018) – Stanford Online, 27 hours (2024).
- PyTorch for Deep Learning Bootcamp – Udemy, 52 hours (2025).