Tomas Ruiz | Resume

Research Assistant

Status: Research Assistant at LMU Munich
Field: Machine Learning, Inference Optimization, Multi-modal AI
Techs: Python, R, C#, Java, AWS, Django, FastAPI, PostgreSQL, vLLM, CUDA

Munich, Germany
+49 176 817 933 17
tomas.ruiz.te@gmail.de

Summary

Research Assistant at LMU Munich at the intersection of Machine Learning and Multi-modal AI. My current work focuses on efficient and scalable inference for large language and vision-language models. I am the author of a paper on test-time scaling accepted at EMNLP 2025 and an active contributor to vLLM, where I implemented speculative decoding with significant throughput improvements. Previously, I worked as a software engineer building scalable systems in industry.

Research and Open Source Contributions

Paper: Test-Time Scaling for LLMs (EMNLP 2025) - EMNLP Shared Task — 2025

* First-author paper on test-time scaling, a technique for inference optimization that trades compute for task performance. Shows strong performance on verifiable-reward problems (e.g., math), accepted at EMNLP 2025. [arXiv:2510.12516](https

Contributor - vLLM Open Source Project — 2025 - Present

* Implemented speculative decoding ([PR 24322](https://github.com/vllm-project/vllm/pull/24322)), achieving 2x throughput improvements (mainly TPOT) and restoring a core feature lost in V1. Collaborate on large-scale inference benchmarking and performance tuning.

Experience

Senior Software Engineer - Allianz Global Investors - Munich, Germany — 03/2023 - 05/2024

* [Python, C#]: Wrote core libraries to simulate trading strategies for options and derivatives, as well as Python orchestration packages. Collaborated closely with quantitative researchers and financial engineers. Learned about financial mathematics, portfolio management and risk modeling.

Software Engineer - Preisenergie GmbH - Munich, Germany — 03/2021 - 02/2023

* [Python, R, REST, PostgreSQL, Django, FastAPI, DDD, TDD]: Designed and built RESTful applications with Django and FastAPI. Refactored a large untested legacy codebase towards DDD. Designed, implemented and tested a quadratic programmingbased price optimization algorithm with 100k variables, improving expected CLV by 20%. Extended a web application to manage and analyze data on energy contracts. Led code reviews and developer onboarding.

Software Engineering Intern - Core Machine Learning - Amazon - Berlin, Germany — 09/2017 - 02/2018

* [Java, AWS]: Wrote a tool to benchmark cutting-edge Multiarmed-Bandits models for product recommendation. Released it on AWS for internal use across Amazon and AWS development teams.

Bachelor's Thesis - BMW Group - Munich, Germany — 03/2017 - 08/2017

* [Python, Spark, Hadoop, Tableau]: Evaluated the query latency, throughput, and DB consistency of an application that monitored the engine manufacturing process.

Education

Research Assistant - Ludwig Maximilian University of Munich — 06/2024 - Present

* Researching on the intersection of machine learning and computational social science. Current work focuses on efficient inference and multimodal content understanding using VLMs (videos, images, audio, text).

Master of Science - Computer Science - Technical University of Munich — 04/2018 - 02/2021

* Thesis on Reinforcement Learning applied to robotics, learning skills with weak supervision. Relevant courses: Machine Learning, Deep Learning, Convex Optimization.

Bachelor of Science - Engineering Science - Technical University of Munich 09/2013 - 08/2017

* Thesis with BMW Group on evaluating Hadoop clusters for engine manufacturing monitoring. Courses in computer science (algorithms, databases) and engineering (PDEs, thermodynamics).