

Jose Sanchez Gonzalez

Machine Learning Engineer / RAG and LLM Systems

(503) 660-9763 | joseszgz2021@gmail.com | github.com/joses2017smjh
linkedin.com/in/jose-j-sanchez-gonzalez-84a800257/ | jose-sanchez-portfolio-com.vercel.app

Summary

Machine learning engineer with an M.S. in Artificial Intelligence. Develops retrieval pipelines, agent workflows, and transformer models, with experience in ranking evaluation, constrained decoding, and containerized model serving.

Technical Skills

Languages and Learning: Python, SQL, C++, PyTorch, scikit-learn, XGBoost, Transformers

Retrieval: Retrieval-Augmented Generation (RAG), BM25, reciprocal rank fusion, pgvector, graph retrieval

Agents and Evaluation: LangGraph, Model Context Protocol (MCP), constrained decoding, nDCG, Recall@k

Systems: FastAPI, PostgreSQL, Docker, ONNX Runtime, Linux, Git, Slurm

Projects

MetaNaviT: Research File Retrieval | Python, PostgreSQL, pgvector, MCP

- Implemented hybrid BM25/dense retrieval, reciprocal rank fusion, staleness filtering, and typed graph traversal; added query routing and approval-gated file operations through Model Context Protocol (MCP).
- Measured Recall@50 of 0.938 and nDCG@10 of 0.493 on 136 labeled queries over a 61-file fixture corpus, using hash embeddings and overlap reranking.
- Improved exact-path nDCG@10 from 0.596 to 0.938 with query routing; bypassed embedding on 8 of 136 queries and reranking on 41 of 136.

Agentic Match Forecasting | XGBoost, LangGraph, FastAPI

- Built a forecasting prototype with XGBoost, isotonic calibration, and conformal prediction sets; connected three MCP servers to a LangGraph orchestrator with explicit approval for staked recommendations.
- Reported 62.7% accuracy versus a 47.1% prior baseline and 0.9013 log loss on 102 World Cup 2026 matches in a historical evaluation with pre-tournament training.

DoM: Math Program Generation | Python, PyTorch, Transformers

- Co-developed a decoder-only transformer in PyTorch to translate MathQA word problems into computation graphs; implemented tokenization, data loading, training, and constrained decoding.
- Achieved 77.45% exact graph match and 84.78% 2-gram BLEU on the MathQA test set; constrained generated graph structure while analyzing remaining semantic errors.

SPUR: Metric Depth for Robotic Pruning | PyTorch, DINOv2, FastAPI

- Reduced synthetic validation depth RMSE from 0.0550 m to 0.0445 m (19%) using DINOv2 RGB-D refinement across three stereo pairs; reported seed variation of 0.0057 m.
- Packaged inference in a Docker/FastAPI service with split ONNX export; benchmarked the PyTorch refiner at 156 ms p50 on a V100 (FP16, six precomputed-depth views).

Experience

Graduate Teaching Assistant, Oregon State University | Sep 2025 – Jun 2026

- Led 2-hour studio sessions for 25+ students in Operating Systems I (C++) and Computational Methods (COMSOL); supported debugging, concurrency, memory management, and numerical simulation.

Math and Computer Science Tutor, Oregon State University | Nov 2022 – Jun 2025

- Tutored 50+ students in mathematics, programming, and computer science; adapted explanations and debugging guidance to individual learning needs.

Education

Oregon State University, Corvallis, OR

Master of Science in Artificial Intelligence | Jun 2026

Bachelor of Science in Computer Science (Applied: Data Science); Mathematics Minor | Jun 2025

Additional Experience

Berkeley Humanoid Lite hardware build (in progress): Soldered 22 magnetic encoders to motor controllers, 3D-printed structural parts, and began actuator calibration and limb assembly.