

Johanna Männistö

+358 40 855 7212 | johanna_mannisto@hotmail.com | linkedin.com/in/jmannisto

Work Experience

Nokia

Espoo, Finland

Data Scientist

May 2024 - Current

- Designed and executed controlled experiments across RAG system components (data processing, indexing, retrieval, generation) to identify performance bottlenecks and optimization opportunities for each customer use case. Delivered 10-30% improvements in retrieval metrics across 5 different RAG use cases
- Coordinated team efforts with external stakeholders using Jira, translated user stories into technical requirements and project roadmaps while documenting decisions to ensure team alignment
- Developed custom evaluation metrics aligned with user-identified downstream task output style preferences, enabling more targeted optimization beyond standard automated metrics
- Documented methodologies, relevant literature, results, and conclusions alongside code to ensure reproducibility and maximize knowledge sharing
- Built feedback/annotation tool with a Streamlit and MongoDB, enabling error analysis by visualizing application outputs, traces, and facilitated user feedback collection
- Conducted exploratory data analysis on large-scale unstructured datasets (PDFs, videos, PowerPoint, logs), to develop domain-specific processing strategies and RAG setting recommendations, such as identifying and removing noise patterns (e.g., boilerplate text) that reduced noise in knowledge base
- Built Airflow pipelines for automated data processing and synthetic data generation to produce training sets and evaluation benchmarks
- Fine-tuned and evaluated decoder-only models (SFT, RAFT), embedding models, and cross-encoder models for RAG systems, conducting experiments to optimize retrieval quality and end-to-end performance

Nokia

Espoo, Finland

Machine Learning Trainee

May 2023 - May 2024

- Trained and evaluated over 200 models (ranging from 89M to 15.5B parameters) across multiple architectures with pipelines build HuggingFace, PyTorch, MLflow, and Flyte
- Iterated on training pipelines to accommodate multiple fine-tuning approaches (Full, (Q)LoRA, Prefix-and Prompt-Tuning) and quantization techniques (NF4 and INT8)

Education

University of Helsinki

Helsinki, Finland

Masters in Language Technology | Cum.GPA: 4.38/5.0

August 2022 - June 2024

- Courses:** Engineering of Machine Learning Systems, Approaches to Natural Language Understanding, Advanced Machine Learning, Machine Learning, Neural Machine Translation, Deep Learning, Data Science, Speech Synthesis & Recognition, Statistics for Data Science, Models and Algorithms for NLP, Big Data Platforms, Computational Morphology, Computational Semantics, Computational Syntax

Projects

Lit Review Agent

Helsinki, Finland

GitHub Repository

January 2026

- Build literature review agent with DSPy and custom MCP tools with FastMCP for searching academic article databases (arXiv and Semantic Scholar), assessing relevance, and summarizing contents into markdown files
- Implemented MLflow tracing to identify performance bottlenecks for future iteration

A Comparative Study of PEFT Approaches for Language Models of Code

Helsinki, Finland

Masters Thesis | Nokia

November 2023 - May 2024

- Explored the impact of using different PEFT techniques for fine-tuning on ability of code language models to generate code across model sizes and architectures
- Orchestrated fine-tuning and evaluation pipelines on SOTA models (up to 8B parameters), with MLflow tracking to ensure reproducibility through systematic logging of hyperparameters and metrics
- Published in NoDaLiDa 2025

Multilingual Knowledge Distillation with Transformers

Helsinki, Finland

Research Project | Green NLP

February 2023 - May 2023

- Conducted research on compressing (multilingual) neural machine translation encoder-decoder models while retaining task performance via knowledge distillation methods
- Implemented various data cleaning techniques and quantized distillation methods to analyzed their impact on distilled models' performance
- Leveraged transfer learning to distill multiple large Transformer models into one smaller model with the aim of generating high performing models with smaller environmental footprints

Skills

Tools & Frameworks

Bash/zsh, Git, Jira, Confluence, HuggingFace, MLflow, Airflow, Docker, MongoDB, PostgreSQL, OpenSearch, FastAPI, Jupyter

Languages

Python, English (Native), Finnish (B2), French (B1)