

# Hajar Sakai, Ph.D.

(607) 240-1073 ◊ [hajarsakai@gmail.com](mailto:hajarsakai@gmail.com) ◊ [LinkedIn](#) ◊ [Research Portfolio](#) ◊ Philadelphia, PA

Ph.D. in Industrial & Systems Engineering with 12+ publications and 3+ years building production ML/NLP systems at scale. Specialized in LLMs, multi-label text classification, and RAG architectures for healthcare.

Research interests span LLM stochasticity modeling, data contamination, and lookahead bias in backtesting.

## Education

---

**Ph.D. Industrial and Systems Engineering**, Binghamton University **May 2025**

*Focus:* LLMs for Healthcare Text Classification | GPA: 4.0

Binghamton, NY, USA

**M.Sc. Research in Finance**, Paris Dauphine University - PSL **Jul 2021**

*Focus:* Time Series Forecasting

Paris, France

**M.Sc. Industrial Engineering**, Mohammadia School of Engineers **Jul 2020**

*Focus:* Volatility Modeling, Time Series Forecasting

Rabat, Morocco

## TECHNICAL SKILLS

---

- **Languages:** Python, R, Julia, SQL
- **ML/DL Frameworks:** PyTorch, TensorFlow, Keras, Scikit-learn, HuggingFace Transformers
- **LLM & NLP:** LangChain, LlamaIndex, ChromaDB, Azure OpenAI, RAG, Prompt Engineering, FFT, PEFT
- **Cloud & MLOps:** Azure AI Foundry, Azure Blob Storage, Microsoft Fabric
- **Statistical & Quant Methods:** GARCH, ARIMA, Bayesian Inference, Time Series Forecasting, Classification
- **Optimization:** Gurobi, Linear & Integer Programming

## WORK EXPERIENCE

---

**Analytics Developer**, Cooper University Health Care

**July 2022 - Present**

*Promoted from Research Student to Analytics Developer*

- **Policies Chatbot:** Designed and deployed a production RAG-based chatbot for hospital policy Q&A using ChromaDB and Azure OpenAI; implemented multi-intent routing, conversation memory, and delta indexing, achieving sub-5s latency at concurrent load
- **Access Data Insights Chatbot Development:** Designed and deployed an Agentic RAG system that unified structured and unstructured data, enabling executives to obtain real-time insights and decision-ready summaries with significantly reduced response time
- **Call Mining:** Developed speech-to-text and LLM solutions to analyze +20,000 call recordings monthly for quality assurance and escalation decision support; Leveraged threading for parallel processing and inference time minimization
- **Surgical Forecasting System:** Built and deployed a time series ML model predicting surgical case volumes across 83 cases with 7% weighted average MAPE, optimizing resource allocation and scheduling efficiency, and deployed to production serving the different surgical case types
- **Call Center Optimization:** Developed predictive models delivering weekly volume forecasts with 5% average error rate, enabling data-driven staffing decisions and achieving  $\leq 5\%$  call abandonment rate
- **PHI Detection Framework:** Designed and implemented quarterly-updated Protected Health Information detection system using text mining, ensuring HIPAA compliance for patient feedback sharing on hospital website
- **Patient Experience Classification:** Created multi-label comment classification system using GPT-3.5/4/4o models with interactive monthly dashboard, identifying improvement areas and monitoring patient satisfaction trends
- **Clinical Communication Analysis:** Developed hierarchical multi-label classification system using GPT-4o to identify communication patterns between call center navigators and nursing staff
- **Statistical Analysis:** Conducted analysis on patient characteristics for denied MRI authorizations, nurse turnover factors, and patient no-show root causes, delivering actionable insights to leadership
- **Cross-functional Collaboration:** Partnered with clinicians, administrators, and operational teams to translate business requirements into ML solutions, ensuring business relevance and operational impact

**Teaching Assistant**, Binghamton University

**Aug 2021 - Dec 2024**

- **Advanced ML Education:** Instructed undergraduate courses in Statistics, Probability, and Python; mentored students on data analysis projects and statistical methods
- **Curriculum Development:** Graded and provided feedback for courses including Neural Networks, Healthcare Financial Engineering, and Adaptive Optimization

**Research Assistant**, DRM Finance Laboratory

**Apr 2021 - Jun 2021**

- **Financial ML Research:** Conducted research on asset pricing and time-series momentum, designed tree-based classifiers and neural networks achieving 90.7% accuracy in predicting US 2-Year Treasury Note Futures signals

**Data Science Intern**, OCP Group, Downstream Department

**Jan 2020 - Jul 2020**

- **Supply Chain Analytics:** Developed GARCH model integration with ML forecasts for dry bulk shipping strategies, achieving 70.22% predictive accuracy for strategic route optimization
- **Process Automation:** Automated freight reporting systems, reducing manual effort by 45 minutes daily for the managerial team

## SELECTED RESEARCH CONTRIBUTIONS

---

- **QUAD-LLM-MLTC:** Ensemble LLM framework for multi-label text classification; addresses label imbalance, LLM stochasticity, and inter-label dependency
- **KDH-MLTC:** Knowledge distillation approach compressing a larger LLM for efficient classification without significant accuracy loss
- **HAMLET:** Adaptive multilingual topic modeling framework leveraging LLMs and GNN, enabling cross-lingual generalization

## ACHIEVEMENTS & AWARDS

---

- Academic Excellence Award in Industrial & Systems Engineering **2025**
- Outstanding Teaching Assistant Award for Systems Science & Industrial Engineering **2025**
- **2nd place**, Data 4 Good Competition, Large Language Models for Medical Information Extraction **2023**
- Krishnaswami Srihari Graduate Scholarship, *Binghamton University* **2022 - 2024**
- Excellence Scholarship, *Paris Dauphine University* **2020 - 2021**

## Additional Information

---

- **Secretary**, *Alpha Pi Mu Honor Society* **2023 - 2024**
- **President and Co-founder**, *EMI-Économie* **2017 - 2019**
- **Certifications:** Deep Learning Specialization, Natural Language Processing Specialization, Lean Six Sigma Green Belt, AZ-900: Microsoft Azure Fundamentals