Charbel **KINDJI**, PhD



AI Scientist · Tabular Foundation Models & Structured Data

(+33) 06 48 47 15 65 | kindji.charbel@gmail.com | charbelkindji.github.io | charbel-kindji | Scholar

Publications & Talks

Publications

- ***Robust Detection of Synthetic Tabular Data Under Schema Variability***, G. Charbel N. Kindji, Elisa Fromont, Lina Maria Rojas-Barahona, Tanguy Urvoy, AAAI 2026 (**A* conference, top 3 in AI**), Singapore, Jan. 2026, `doi:10.1609/aaai.v40i27.39422`
- ***Tabular Data Generation Models: An In-Depth Survey and Performance Benchmarks with Extensive Tuning***, G. Charbel N. Kindji, Lina Maria Rojas-Barahona, Elisa Fromont, Tanguy Urvoy, Neurocomputing (Q1 Journal, Elsevier), 2025, `doi:10.1016/j.neucom.2025.131655`
- ***Synthetic Tabular Data Detection In the Wild***, G. Charbel N. Kindji, Elisa Fromont, Lina Maria Rojas-Barahona, Tanguy Urvoy, International Symposium on Intelligent Data Analysis (IDA), Konstanz, Germany, May 2025, `doi:10.1007/978-3-031-91398-3_7`
- ***Cross-table Synthetic Tabular Data Detection***, G. Charbel N. Kindji, Lina Maria Rojas-Barahona, Elisa Fromont, Tanguy Urvoy, Workshop GenAIDetect, COLING 2025, Abu Dhabi, Jan. 2025, `aclanthology.org/2025.genaidetect-1.5`

Talks

- ***Invited talk: "Synthetic tabular data detection in the wild"***, CFE-CMStatistics 2025, Birkbeck, University of London, London, Dec. 2025 — International hybrid conference ∼1 800 participants (econometrics & computational statistics), session *Synthetic data: Generation and validation methods*, `cmstatistics.org`

Experience

CIFRE PhD Candidate — Tabular Foundation Models & Generative AI

Lannion / Rennes, France

Orange Innovation / IRISA - INRIA — *December 2022 – December 2025*

- Designed and trained **from scratch** a **foundation model for structured/tabular data** (<100M parameters) — full pipeline: architecture design, pre-training corpus curation, training at scale, evaluation on unseen schemas ; achieves competitive performance against large pre-trained models
- Designed an original Transformer architecture (*Datum-wise*) **from scratch** for heterogeneous tabular data — **+7 AUC points and +7 accuracy points** over the only published baseline, with only **3M parameters** (vs. 110M+) ; extended via domain adaptation (*gradient reversal*) for cross-schema generalisation: additional **+7 accuracy points**
- Fine-tuning and adaptation of **LLMs on structured data** — tabular serialisation, heterogeneous feature encoding, transfer to unseen schemas
- Rigorous benchmark of 5 generative model families (VAE, GAN, Diffusion, LLMs, classical) on **16 datasets** with **300 HPO trials per fold** — end-to-end pipeline: data curation, model design, training, evaluation and ablations
- **Large-scale distributed training** on HPC clusters (Jean-Zay, IGRIDA, Marcel/Orange) — Slurm job management, Tesla V100 32GB and RTX 4090 GPUs
- **First author** on 4 international publications: **AAAI 2026** (A* conference, top 3 worldwide in AI), *Neurocomputing* Q1, IDA 2025, COLING 2025 ; invited talk at **CFE-CMStatistics 2025** (Birkbeck University of London, ∼1 800 attendees)
- Co-supervised an M2 intern (6 months) on foundation model extension — topic definition, weekly technical mentoring, and result evaluation

Apprentice — Software Developer & Data Scientist

Rennes, France

Kurmi Software — *September 2020 – September 2022*

- Developed Kurmi connectors for integration with Microsoft 365 (Java, JavaScript, PowerShell)
- Behavioral analysis of Kurmi logs across **21 dimensions** (67 operation types, 19 entity types) for user segmentation — Python/KNIME pipeline, Power BI visualization
- Designed and maintained software tests to ensure the quality of the Kurmi suite

Research Intern — DRUID Team

Rennes, France

Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA - INRIA) — *May 2020 – August 2020*

- Analyzed the evolution of scientific fields through NLP co-occurrence graphs on **22 240 arXiv papers** (cs.AI, 1993–2019) — graphs up to 9 500 nodes and 2.8M edges
- Distributed computing on the IGRIDA grid (35 nodes, 696 cores) for graph generation and spectral analysis

Education

PhD in Computer Science: ***Synthetic Tabular Data: Generation and Detection*** — *Rennes, France*
University of Rennes — *2022 – 2025*

Bachelor's & Master's Degree in Computer Science (MIAGE) — *Rennes, France*
University of Rennes — *2019 – 2022*

Technical Skills

ML & Deep Learning	PyTorch, Hugging Face Transformers, Scikit-learn, Optuna, RayTune
Foundation Models	Pre-training from scratch, architecture design, tabular foundation models, LLM fine-tuning
Generative Models	VAE, GAN, Diffusion models, synthetic data generation & detection
Distributed Training & HPC	Slurm, Jean-Zay (IDRIS), IGRIDA (IRISA), Marcel (Orange), distributed GPU training
Structured Data	Tabular data modelling, heterogeneous schemas, synthetic data generation & detection
Languages/Tools	Python, Java, JavaScript, PowerShell, Git, Linux, Jupyter, VSCode

Languages

French	Native
English	Advanced (C1)
Fon	Native
Japanese	Beginner

Additional Information

Driving License	Full driver's license (Category B)
Interests	Video games, reading, piano, movies, manga, sports