Independent report on the PhD thesis of Nabeel Seedat

Date: 25-09-2025

Summary

This thesis focuses on Data-Centric AI and introduces an impressive set of innovations that contribute to the reliability and trustworthiness of machine learning (ML) systems applied in challenging real-world domains, especially the medical domain. It provides an ambitious, end-to-end framework that spans the entire ML pipeline, from data collection to model deployment.

The work is structured around five key data-related challenges identified within the author's "DC-Check" framework:

1. **Data Characterization:** During data collection, we need to better understand the quality and nature of training data. It introduces **Data-IQ**, a framework that uses a model's training dynamics — specifically prediction confidence and aleatoric (data) uncertainty — to categorize training as either "easy," "ambiguous," or "hard" to learn instances. This is shown to be useful for tasks like data sculpting, principled feature acquisition, and dataset auditing across tabular, image, and text data.
2. **Test-Time Incongruity Detection:** Moving to the deployment phase, the thesis aims to identify heterogeneous regions *within* the data that can lead to model failures. It proposes **Data-SUITE**, a model-agnostic framework that constructs feature-wise confidence intervals to flag test instances that are "inconsistent" or "uncertain" with respect to the training data.
3. **Synthetic Data Augmentation and Curation:** To address data scarcity, particularly in "ultra-low data regimes" (n<100), the thesis introduces **Curated LLM (CLLM)**. This framework leverages LLMs to produce high-quality, realistic synthetic data that improves the performance of downstream models, especially on underrepresented subgroups.
4. **Context-Aware Model Evaluation:** The thesis challenges "data-only" model testing. It formalizes **Context-Aware Testing (CAT)** and presents **SMART Testing**, an LLM-based system that generates and falsifies hypotheses about potential model failures. This approach aims to surface more meaningful and impactful model deficiencies than traditional data-slicing methods.
5. **Data-Enhanced Uncertainty Quantification:** Finally, the work aims to improve uncertainty quantification. It proposes **Self-Supervised Conformal Prediction (SSCP)**, a framework that enhances the efficiency and adaptivity of conformal prediction intervals. It achieves this by leveraging auxiliary signals from self-supervised learning tasks performed on both labeled and unlabeled data.

Together, these five contributions form a cohesive, principled, and end-to-end framework for operationalizing data-centric AI to advance the reliability and trustworthiness of ML systems, and rigorously evaluates in on tasks with real-world relevance, especially in the medical domain.

Main Strengths

1. The core contributions of the thesis have been **published in top-tier peer-reviewed conferences** (NeurIPS, ICML, ICLR, AISTATS), across 10 papers, which is very impressive within a single PhD. This provides strong external validation of the quality, novelty, and significance of the work.
2. The thesis is **very well structured** around the DC-Check framework. It presents an ambitious, clear, compelling, and holistic vision for data-centric AI, connecting disparate research problems (e.g., data generation, model testing) into a single logical pipeline. Each chapter tackles a distinct yet interconnected challenge, strengthening the overall data-centric argument.
3. The work is **creative** and **innovative** and truly at the forefront of the data-centric AI movement. It addresses highly relevant problems in ML and tackles them by leveraging the very latest AI developments (including LLMs). It is also holistic and the proposed frameworks (Data-IQ, Data-SUITE, CLLM, SMART, SSCP) together offer a novel and principled solutions to these challenges.
4. Each proposed method is **rigorously evaluated** by experiments across multiple datasets with **real-world relevance**, particularly in the medical domain (e.g., COVID mortality, prostate cancer registries). It uses relevant baselines, ablation studies, and sensitivity analyses. The 120-page appendix with in-depth evaluations and additional experiments is especially impressive and hints an impressive work ethic and scientific drive.
5. This work is not only academic, it offers **practical tools** for ML practitioners, especially to make ML more trustworthy in low-data regimes (CLLM), providing actionable insights for model deployment (Data-SUITE, SMART), and guiding data collection (Data-IQ).
6. The work integrates uncertainly quantification in a fundamental, principled way, contributing to trustworthy AI. It leverages the formal guarantees of conformal prediction for identifying data incongruity (Data-SUITE) and subsequently enhances conformal prediction itself with data-centric signals (SSCP).

Room for improvement and future work

The work is of excellent quality and includes an impressive quantity of ideas and experiments. It is hard to find significant faults or weaknesses. These points should therefore be seen as items for future improvement in subsequent work.

1. CLLM and SMART Testing are dependent on black-box LLMs like GPT-4, which can be biased or incorrect (and rapidly changing). The work touches upon biases but could further explore the risks of inheriting biases, factual incorrectness, or reasoning failures from the LLM.

2. Data-IQ and CLLM will filter and remove data samples deemed "hard," "ambiguous," or low-quality. This could be a risk, e.g. if they inadvertently discard rare but valid data from minority or underrepresented groups, especially in medical data. Of course, the work also notes that CLLM shows the largest gains in underrepresented subgroups. I'm still a bit worried that there may be implicit subgroups affected by this.
3. The methods are evaluated mainly on tabular data, although qualitative examples are also provided for images and text in Chapter 2. Hence, it is not always clear whether these methods would translate easily to unstructured data.
4. The different methods are separate, but could they actually be connected? For instance, could the "ambiguous" subgroups identified by Data-IQ be used as a source of context to generate hypotheses for SMART Testing? Or could Data-SUITE's incongruity score be integrated into the SSCP framework?

Overall, I believe that this work meets all the requirements of the PhD program.

dr. ir. Joaquin Vanschoren
Associate Professor
Eindhoven University of Technology