

University of Salerno

Ph.D. Program in Computer Science

Curriculum Internet of Things and Smart Technologies

Software Sustainability in the ML Era: Empirical Insights, Trade-offs, and LLM Methodologies

Author
Vincenzo De Martino

Advisor
Prof. Fabio Palomba

XXXVIII Cycle

Abstract

Machine Learning (ML) has become a vital part of modern software systems, driving data-driven intelligence and automation across various fields such as healthcare, finance, mobility, and education. As ML technologies are increasingly integrated into critical infrastructures, they challenge traditional software engineering principles, introducing new complexities and quality concerns that extend beyond mere functional correctness. The development of ML-enabled systems is influenced by uncertainty, a continuous dependency on data, and evolving models that require ongoing retraining and maintenance.

In this context, it is essential to ensure compliance with non-functional requirements (NFRs), and among these, the concept of sustainability remains largely unexplored. Though sustainability has recently gained attention in both research and practice, its definition is fragmented, its evaluation inconsistent, and its integration into engineering processes limited. These problems lead to a persistent gap between high-level ideals and practical applications, where sustainability is often cited as a principle but seldom translated into measurable and actionable properties.

This dissertation aims to fill a significant gap by integrating empirical, analytical, and methodological investigations into the sustainability of ML-enabled systems. The research is guided by four main questions that collectively explore the understanding, evaluation, scaling, and methodological support of sustainability. Through a multiphase empirical design, this thesis examines how sustainability is defined and perceived in both academia and industry, how it interacts with other NFRs, how its study can be automated through repository mining, and how the use of Large Language Models (LLMs) can be designed to ensure rigor and energy awareness in empirical software engineering.

The first part of the dissertation explores the NFRs in ML-enabled systems, the related challenges, and how sustainability is perceived, as well as the practices and challenges faced by ML practitioners. A systematic literature review reveals 31 NFRS and 26 related challenges, and that most research tends to frame sustainability primarily as an environmental issue, focusing on aspects like energy efficiency and carbon reduction. While this perspective is important, it often overlooks other relevant dimensions, such as social and cost.

Complementing this academic view, an empirical study conducted with ML practitioners indicates that the industry takes a much more pragmatic interpretation of sustainability. For

these professionals, sustainability is a multifaceted goal that encompasses not only training and inference efficiency, but also the maintainability of ML pipelines and the ethical implications of algorithmic design.

The contrast between these two perspectives reveals a serious structural misalignment, where academic literature tends to frame sustainability in abstract and theoretical terms, often detaching it from practical application. In contrast, practitioners integrate sustainability into their decision-making, viewing it as a concrete engineering challenge that requires immediate and actionable solutions.

This disconnect clearly demonstrates that sustainability in ML-based systems cannot be confined to a single dimension. It must be recognized as a complex and multifaceted quality that arises from trade-offs between technical capabilities, environmental considerations, and ethical responsibilities while reducing the overall production costs. A comprehensive understanding of sustainability requires balancing these different factors in the development and implementation of ML-enabled systems, ensuring that approaches are both responsible and effective.

Building on this conceptual basis, the second part of the thesis empirically examines how sustainability interacts with other NFRs. An experimental benchmark evaluates the impact of bias mitigation algorithms, which are commonly introduced to enhance fairness across multiple dimensions of sustainability. The results show that improvements in fairness are often accompanied by increased energy consumption and decreased accuracy, indicating that socially desirable interventions can have significant environmental and technical costs. Conversely, highly energy-efficient configurations frequently lead to reductions in accuracy or an increase in the ML model bias. These results empirically confirm that sustainability optimization practices for an ML-enabled system have different quality characteristics that must be controlled during the development process, rather than a single objective to be optimized. They also highlight the need for a multidimensional evaluation approach that can quantify and balance trade-offs among fairness, energy efficiency, and performance. In this regard, the study contributes to reframing sustainability as a property that must be managed holistically throughout the ML system lifecycle.

To enable the large-scale and reproducible study of sustainability practices, this thesis introduces MARK (Machine Learning Automated Rule-based Classification Kit), a rule-based classifier designed to automatically categorize ML repositories according to their functional

roles within the ML lifecycle. The motivation behind the development of MARK stems from a recurring limitation observed in the literature that, although thousands of ML repositories are available to the public, their internal organization and role within the ML ecosystem remain largely undocumented. Manual inspection of these projects is not only time-consuming but also error-prone and non-reproducible, especially when dealing with large-scale datasets. Sustainability analyses, in particular, require clear differentiation between repositories that develop models (ML-Model Producer), those that apply pre-trained models (ML-Model Consumer), and those that integrate both functionalities (ML-Model Producers & Consumers). Without this distinction, comparing practices and measuring sustainability-related properties becomes infeasible. MARK was therefore conceived as a scalable, transparent, and reproducible mechanism for repository classification, serving as a basis for subsequent sustainability mining and benchmarking tasks.

Building on the datasets classified by MARK, this dissertation advances the investigation of sustainability practices by exploring how LLMs can be employed to extract and characterize sustainable practices from real-world ML repositories. The rationale behind this study stems from the awareness that while research is advancing in identifying sustainable practices, it has not been determined whether developers actually adopt these practices. Traditional mining techniques, which rely on static or keyword-based analysis, often fail to capture these latent and context-dependent signals. LLMs, by contrast, possess the ability to reason over unstructured text and heterogeneous artifacts, providing a novel means to bridge the semantic gap between sustainability concepts and their technical realization in software projects.

In this work, LLMs are integrated into a prompt-based pipeline designed to detect, describe, and elicit the green architectural tactics, design strategies, and coding practices that promote energy efficiency, maintainability, and responsible resource usage within ML-enabled systems. Each prompt was carefully constructed and iteratively refined to balance task clarity, contextual grounding, and reproducibility. The models were asked to analyze Python code to identify green tactics correctly. The use of LLMs has also enabled the identification of new tactics not found in existing literature, highlighting the need to continue identifying sustainable practices in the development process. LLMs facilitate semi-automated annotation of repositories, significantly reducing the manual effort needed to identify sustainability tactics. These findings also aid in the creation of larger and more accurate datasets, both in terms of scale and granularity. This phase illustrates that LLMs can be effective knowledge extractors, capable of identifying

important features from code during software development. Additionally, the development of this LLM-based mining approach has highlighted the need for pipelines that enhance reusability, energy efficiency, and maintainability by establishing guidelines for conducting mining software repository studies properly.

Nevertheless, the empirical evaluation also exposes the limitations and risks inherent in deploying LLMs for sustainability mining. The outputs of the models exhibit variability across runs, even when identical prompts and repositories are used. This stochastic behavior complicates reproducibility and raises concerns about the stability of results. In some cases, the models generate partial or hallucinated descriptions—plausible but unfounded interpretations that do not correspond to actual repository content. Such inaccuracies can propagate into the analysis if not carefully validated through human inspection or automated consistency checks. Beyond these risks, the computational cost of executing LLMs at scale poses a significant challenge. The inference process is energy-intensive and costly, particularly when using private LLMs, which require substantial hardware resources and large-scale mining. This directly contradicts the study's sustainability goals. Although LLMs can accelerate sustainability research and reveal insights that traditional methods cannot access, their energy consumption and unpredictability raise concerns about the very principles they are meant to support.

To address these challenges and to guide the systematic use of LLMs in empirical studies, the dissertation proposes PRIMES 2.0 (Prompt Refinement and Insights for Mining Empirical Software repositories), an extended methodological framework for LLM-based mining of software repositories. PRIMES 2.0 structures the entire process of LLM-based research into distinct and interconnected stages, encompassing prompt design, pilot testing, model comparison, validation, and synthesis. Each stage is associated with potential methodological threats and corresponding mitigation strategies. PRIMES 2.0 promotes a principled approach grounded in reproducibility, transparency, and sustainability. It encourages detailed documentation of prompt versions, model configurations, and evaluation procedures, as well as the integration of energy-awareness metrics into experimental planning. Moreover, the framework advocates practices such as dual annotation, statistical validation, and iterative refinement to improve the reliability of LLM-generated results. By incorporating sustainability principles into its methodological core, PRIMES 2.0 improves the rigor of LLM-based studies and also aligns their execution with ethical and environmental objectives.

Through these interconnected contributions, the thesis offers a comprehensive overview of sustainability in ML-based systems, ranging from conceptual analysis to empirical measurement, from tool development to methodological design. It demonstrates that sustainability cannot be considered a secondary requirement; instead, it must be seen as a fundamental property integrated early in the software design phase and analyzed throughout the entire lifecycle. The results show that sustainability emerges from balancing conflicting goals such as environmental efficiency, fairness, maintainability, and accuracy, which must be considered throughout the lifecycle of ML systems. Through MARK, the work provides a scalable means to operationalize sustainability analysis in large repositories. Through the LLM-based mining pipeline, it expands our empirical understanding of sustainable architectural practices. Through PRIMES 2.0, it establishes the methodological foundation for conducting these studies responsibly.

In conclusion, this thesis lays the foundation for further studies that include short- and long-term sustainability to design responsible systems. It contributes to the promotion of sustainable ML engineering with quantitative and qualitative studies, automated tools, and methodological guidelines. It offers both researchers and professionals concrete means to design ML-based systems that are not only high-performing and fair, but also energy-efficient, maintainable, and socially responsible.

Abstract

Il Machine Learning (ML) è diventato una componente fondamentale dei moderni sistemi software, guidando l'intelligenza data-driven e l'automazione in diversi settori quali sanità, finanza, mobilità ed educazione. Con l'integrazione crescente delle tecnologie ML nelle infrastrutture critiche, emergono nuove sfide per i principi tradizionali dell'ingegneria del software, introducendo complessità e problemi di qualità che vanno oltre la semplice correttezza funzionale. Lo sviluppo dei sistemi abilitati al ML è infatti caratterizzato da incertezza, continua dipendenza dai dati e modelli in evoluzione che richiedono aggiornamenti e manutenzione costanti.

In questo contesto, diventa essenziale garantire il rispetto dei requisiti non funzionali (NFR), tra i quali la sostenibilità rimane ampiamente inesplorata. Nonostante l'attenzione crescente nella ricerca e nell'industria, la sostenibilità presenta definizioni frammentate, valutazioni incoerenti e un'integrazione limitata nei processi ingegneristici. Questi problemi producono un divario persistente tra principi di alto livello e applicazioni concrete, in cui la sostenibilità è

spesso citata come valore, ma raramente tradotta in proprietà misurabili e azionabili.

Questa dissertazione mira a colmare tale divario integrando indagini empiriche, analitiche e metodologiche sulla sostenibilità dei sistemi basati su ML. La ricerca è guidata da quattro domande principali che esplorano la comprensione, la valutazione, la scalabilità e il supporto metodologico della sostenibilità. Attraverso un disegno empirico multifase, la tesi analizza come la sostenibilità viene definita e percepita nel mondo accademico e industriale, come interagisce con altri NFR, come il suo studio possa essere automatizzato tramite il mining di repository, e come l'uso dei Large Language Models (LLM) possa essere progettato per garantire rigore e consapevolezza energetica nell'ingegneria del software empirica.

La prima parte della dissertazione esplora i requisiti non funzionali nei sistemi ML, le sfide correlate e la percezione della sostenibilità, insieme alle pratiche e difficoltà affrontate dai professionisti del settore. Una revisione sistematica della letteratura rivela 31 NFR e 26 sfide collegate, mostrando che gran parte della ricerca interpreta la sostenibilità principalmente come un tema ambientale, focalizzandosi sull'efficienza energetica e sulla riduzione delle emissioni. Sebbene questa prospettiva sia importante, spesso trascura altre dimensioni rilevanti, come quelle sociali ed economiche.

A completamento della visione accademica, uno studio empirico condotto con professionisti del ML mostra come l'industria adotti un'interpretazione molto più pragmatica della sostenibilità. Per questi esperti, la sostenibilità rappresenta un obiettivo multifattoriale che comprende non solo l'efficienza del training e dell'inferenza, ma anche la manutenibilità delle pipeline di ML e le implicazioni etiche della progettazione algoritmica.

Il contrasto tra queste due prospettive rivela un serio disallineamento strutturale: mentre la letteratura tende a trattare la sostenibilità in termini astratti e teorici, spesso slegati dalle applicazioni pratiche, i professionisti la integrano direttamente nei processi decisionali, considerandola una sfida ingegneristica concreta che richiede soluzioni immediate e operative.

Questo divario dimostra chiaramente che la sostenibilità nei sistemi di ML non può essere ridotta a una singola dimensione. Deve invece essere concepita come una qualità complessa e multifaccettata che emerge dai trade-off tra capacità tecniche, considerazioni ambientali e responsabilità etiche, contribuendo al tempo stesso alla riduzione dei costi complessivi di produzione. Una comprensione completa della sostenibilità richiede un bilanciamento tra questi fattori lungo l'intero ciclo di sviluppo e di implementazione dei sistemi ML.

Sulla base di queste fondamenta concettuali, la seconda parte della tesi esamina empiricamente come la sostenibilità interagisca con gli altri NFR. Un benchmark sperimentale valuta l'impatto degli algoritmi di mitigazione del bias — comunemente introdotti per migliorare la fairness — sulle diverse dimensioni della sostenibilità. I risultati mostrano che i miglioramenti in fairness sono spesso accompagnati da un aumento del consumo energetico e da una riduzione dell'accuratezza, indicando che gli interventi socialmente desiderabili possono comportare costi ambientali e tecnici rilevanti. Al contrario, configurazioni altamente efficienti dal punto di vista energetico possono ridurre l'accuratezza o aumentare il bias. Questi risultati confermano empiricamente che l'ottimizzazione della sostenibilità nei sistemi ML coinvolge molteplici caratteristiche di qualità da gestire durante lo sviluppo, non un singolo obiettivo da ottimizzare. Evidenziano inoltre la necessità di un approccio valutativo multidimensionale capace di quantificare e bilanciare i trade-off tra fairness, efficienza energetica e performance. In questa prospettiva, lo studio contribuisce a riconfigurare la sostenibilità come una proprietà da gestire in modo olistico lungo l'intero ciclo di vita del sistema.

Per consentire studi di sostenibilità su larga scala e in modo riproducibile, la tesi introduce MARK (Machine Learning Automated Rule-based Classification Kit), un classificatore basato su regole progettato per categorizzare automaticamente i repository ML in base al loro ruolo funzionale nel ciclo di vita del ML. La motivazione nasce da una limitazione ricorrente nella letteratura: sebbene migliaia di repository ML siano pubblicamente disponibili, la loro struttura interna e il loro ruolo nell'ecosistema ML sono per lo più non documentati. L'ispezione manuale è lenta, soggetta a errori e non riproducibile, soprattutto su larga scala. Le analisi di sostenibilità, in particolare, richiedono una chiara distinzione tra repository che sviluppano modelli (ML-Model Producer), che applicano modelli pre-addestrati (ML-Model Consumer), o che integrano entrambi i ruoli (ML-Model Producers & Consumers). Senza questa distinzione, risulta impossibile confrontare pratiche o misurare proprietà legate alla sostenibilità. MARK è pertanto concepito come un meccanismo scalabile, trasparente e riproducibile per la classificazione dei repository, costituendo la base per le successive fasi di mining e benchmarking.

Utilizzando i dataset classificati da MARK, la dissertazione approfondisce lo studio delle pratiche di sostenibilità esplorando come gli LLM possano essere impiegati per estrarre e caratterizzare le pratiche sostenibili presenti nei repository ML reali. La motivazione deriva dalla consapevolezza che, sebbene la ricerca stia avanzando nell'identificazione di pratiche sostenibili, non è ancora chiaro se e come gli sviluppatori le adottino effettivamente. Le tecniche tradizionali

di mining, basate su analisi statiche o keyword, spesso non colgono segnali latenti o dipendenti dal contesto. Gli LLM, al contrario, sono in grado di ragionare su testi non strutturati e artefatti eterogenei, colmando il divario semantico tra concetti di sostenibilità e loro realizzazione tecnica nei progetti software.

In questa ricerca, gli LLM vengono integrati in una pipeline basata su prompt per rilevare, descrivere e identificare tattiche architetturali "green", strategie progettuali e pratiche di codifica che favoriscono l'efficienza energetica, la manutenibilità e un uso responsabile delle risorse nei sistemi basati su ML. Ogni prompt è stato progettato e raffinato iterativamente per equilibrare la chiarezza del compito, la contestualizzazione e la riproducibilità. Ai modelli è stato chiesto di analizzare il codice Python per identificare correttamente le tattiche green. L'uso degli LLM ha inoltre consentito di individuare nuove tattiche non presenti nella letteratura esistente, evidenziando la necessità di proseguire l'esplorazione di pratiche sostenibili nei processi di sviluppo. Gli LLM supportano l'annotazione semi-automatica dei repository, riducendo significativamente lo sforzo manuale e contribuendo alla creazione di dataset più ampi e accurati, sia in termini di scala che di granularità. Questa fase dimostra che gli LLM possono agire come efficaci estrattori di conoscenza, in grado di identificare elementi rilevanti dal codice durante lo sviluppo software. Inoltre, lo sviluppo di questa pipeline ha messo in luce l'importanza di processi più riutilizzabili, efficienti e manutenibili, fornendo linee guida per condurre correttamente gli studi di mining dei repository software.

L'analisi empirica ha tuttavia evidenziato anche limiti e rischi nell'uso degli LLM per il mining della sostenibilità. Gli output mostrano variabilità tra esecuzioni, anche con gli stessi prompt e gli stessi repository. Questa natura stocastica ostacola la riproducibilità e solleva dubbi sulla stabilità dei risultati. In alcuni casi, i modelli generano descrizioni parziali o allucinate — plausibili ma prive di fondamento — che possono compromettere l'analisi se non validate tramite ispezione umana o controlli automatici. Oltre a ciò, il costo computazionale dell'esecuzione su larga scala rappresenta una sfida significativa: l'inferenza dei LLM è energivora e costosa, soprattutto se basata su modelli privati che richiedono infrastrutture hardware considerevoli. Ciò contraddice direttamente gli obiettivi di sostenibilità dello studio. Sebbene gli LLM possano accelerare la ricerca e rivelare insight non ottenibili con metodi tradizionali, il loro consumo energetico e la loro imprevedibilità sollevano questioni sui principi che dovrebbero sostenere.

Per affrontare tali sfide e guidare l'uso sistematico dei LLM negli studi empirici, la disser-

tazione propone PRIMES 2.0 (Prompt Refinement and Insights for Mining Empirical Software repositories), un framework metodologico esteso per il mining dei repository basato su LLM. PRIMES 2.0 struttura l'intero processo di ricerca in fasi distinte e interconnesse, includendo progettazione dei prompt, test pilota, confronto dei modelli, validazione e sintesi. Ogni fase è associata a potenziali minacce metodologiche e relative strategie di mitigazione. Il framework promuove un approccio rigoroso, fondato su riproducibilità, trasparenza e sostenibilità. Incoraggia la documentazione dettagliata delle versioni dei prompt, delle configurazioni dei modelli e delle procedure di valutazione, oltre all'integrazione di metriche energetiche nella pianificazione sperimentale. Inoltre, raccomanda pratiche come doppia annotazione, validazione statistica e raffinamento iterativo per migliorare l'affidabilità dei risultati generati dai LLM. Integrando la sostenibilità nel suo nucleo metodologico, PRIMES 2.0 non solo migliora il rigore degli studi basati su LLM, ma ne allinea l'esecuzione a obiettivi etici e ambientali.

Attraverso questi contributi interconnessi, la tesi offre una panoramica completa della sostenibilità nei sistemi basati su ML, dalla concettualizzazione alla misurazione empirica, dallo sviluppo di strumenti al design metodologico. Dimostra che la sostenibilità non può essere considerata un requisito secondario, ma deve essere integrata sin dalle prime fasi di progettazione e analizzata durante l'intero ciclo di vita del software. I risultati mostrano che la sostenibilità emerge dal bilanciamento tra obiettivi spesso in conflitto — efficienza ambientale, fairness, manutenibilità e accuratezza — che devono essere gestiti come proprietà co-esistenti nei sistemi ML. Attraverso MARK, la tesi fornisce un mezzo scalabile per supportare analisi su larga scala. Attraverso la pipeline basata su LLM, amplia la comprensione empirica delle pratiche architetturali sostenibili. Attraverso PRIMES 2.0, stabilisce le basi metodologiche per condurre questi studi in modo responsabile.

In conclusione, questa tesi pone le basi per studi futuri che integrino visioni di sostenibilità a breve e lungo termine nella progettazione di sistemi responsabili. Contribuisce alla promozione dell'ingegneria sostenibile per ML mediante studi quantitativi e qualitativi, strumenti automatizzati e linee guida metodologiche. Offre a ricercatori e professionisti strumenti concreti per progettare sistemi ML che non siano solo performanti ed equi, ma anche efficienti dal punto di vista energetico, manutenibili e socialmente responsabili.