

University of Salerno

Ph.D. Program in Computer Science

Curriculum Internet of Things and Smart Technologies

Thesis Abstract

Author:
Gabriele De Vito

Advisor
Prof. Filomena Ferrucci

XXXVIII Cycle

Abstract (English)

Large Language Models (LLMs) are reshaping the development of artificial intelligence applications, clearly displaying unexpected abilities in natural language understanding and generation tasks. As these systems evolve, their deployment in specific domains like software engineering and healthcare presents exciting opportunities but also new challenges. The textual nature of software artifacts and unstructured healthcare data makes these domains prime candidates for processing by language models, opening up previously unexplored possibilities for automating tasks that have traditionally been performed by humans. On the other hand, LLM-based systems in these domains must meet stringent requirements for accuracy, consistency, and safety, and must integrate smoothly into existing professional workflows.

Even though the interest from the research community is growing and researchers have probed diverse LLM architectures in order to balance their capabilities with deployment considerations, and investigated different methods for enhancing task performance (i.e., adaptation approaches), there are still limitations and challenges that threaten the practicality of LLM-based applications in real-world software engineering and healthcare scenarios. In particular, there is a fragmented evidence regarding (i) different adaptation strategies across specialized domains and complex tasks and how well they perform, (ii) different model architectures and the kinds of tasks they are good at (or not), with (iii) insufficient account taken of the need for reliable, consistent performance in LLM applications destined for high-stakes environments.

In this thesis, we aim to address these issues by developing novel LLM adaptation and deployment approaches specifically designed for software engineering and healthcare domains. We present a set of contributions, ranging from confidence-based frameworks to modular architectures and domain-specific adaptation methods, which we evaluate across diverse applications, including use case quality improvements, automated testing generation, functional size estimates, clinical decision support, and drug interaction predictions. Through an extensive analysis, we investigate (i) the performance of various adaptation strategies in domain-specific tasks, (ii) the factors influencing the stability and dependability of LLM performance, (iii) the features of the model architecture that affect performance in specialized settings, and (iv) the prerequisites for a successful integration into current workflows.

This dissertation reports findings that challenge several common assumptions about the optimal deployment strategies for LLMs. First of all, we find that model size does not correlate directly with performance effectiveness. Small models achieve results comparable to systems orders of magnitude larger after appropriate domain adaptation. Local deployment using open-weight models can match the performance of cloud-based, proprietary alternatives while providing enhanced data sovereignty and operational independence. Second, our research uncovers method-model interaction complexities that do not lend themselves to universal solutions by demonstrating that using the identical adaptation strategies can have dramatically different outcomes across LLMs. Third, we observe that confidence-based frameworks offer safety-first design capabilities while creating better protection and operational efficacy, contrary to the expectation that conflicts exist between safety and effectiveness. We also show that user acceptance is contingent on the transparency of the decision-making process and users' autonomy, not just on raw system performance.

Lastly, we outline a range of open issues that we believe need to be taken up by the research community, including the development of predictive model-method compatibility frameworks, dynamic confidence calibration across domains, systematic bias detection and mitigation in ensemble failures, and longitudinal performance stability mechanisms to ensure consistent performance over time.

Abstract (Italiano)

I Large Language Models (LLM) stanno trasformando radicalmente lo sviluppo di applicazioni di intelligenza artificiale, dimostrando capacità sorprendenti e inaspettate nei compiti di comprensione e generazione del linguaggio naturale. Con l'evoluzione di questi sistemi, la loro adozione in domini specifici come l'ingegneria del software e la sanità offre opportunità entusiasmanti ma pone anche nuove sfide. La natura testuale degli artefatti software e dei dati sanitari non strutturati rende questi domini particolarmente adatti all'elaborazione tramite modelli linguistici, aprendo possibilità prima inesplorate per l'automazione di compiti tradizionalmente svolti dall'uomo. D'altro canto, i sistemi basati su LLM in questi ambiti devono soddisfare requisiti rigorosi in termini di accuratezza, coerenza e sicurezza, e devono integrarsi senza soluzione di continuità nei flussi di lavoro professionali esistenti.

Sebbene l'interesse della comunità scientifica sia in crescita e i ricercatori abbiano esplorato diverse architetture di LLM nel tentativo di bilanciare le loro capacità con le esigenze di implementazione, e abbiano studiato diversi metodi per migliorare le prestazioni nei vari compiti (ossia, approcci di adattamento), persistono ancora limitazioni e sfide che compromettono l'applicabilità pratica dei sistemi basati su LLM in scenari reali di ingegneria del software e sanità. In particolare, le evidenze risultano frammentate per quanto riguarda (i) le diverse strategie di adattamento applicabili a domini specializzati e compiti complessi e la loro efficacia, (ii) le diverse architetture di modelli e le tipologie di compiti in cui si dimostrano efficaci (o inefficaci), e (iii) la scarsa considerazione della necessità di prestazioni affidabili e coerenti per le applicazioni LLM destinate ad ambienti critici e ad alto rischio.

In questa tesi ci proponiamo di affrontare tali problematiche sviluppando nuovi approcci di adattamento e deployment di LLM specificamente progettati per i domini dell'ingegneria del software e della sanità. Presentiamo una serie di contributi che spaziano da framework basati sulla stima della confidenza ad architetture modulari e metodi di adattamento specifici per dominio, che valutiamo attraverso diverse applicazioni, tra cui il miglioramento della qualità dei casi d'uso, la generazione automatica di test, la stima delle dimensioni funzionali, il supporto alle decisioni cliniche e la previsione delle interazioni farmacologiche. Attraverso un'analisi approfondita, indaghiamo (i) le prestazioni di diverse strategie di adattamento in compiti specifici di dominio, (ii) i fattori che influenzano la stabilità e l'affidabilità delle prestazioni degli LLM, (iii) le caratteristiche architetturali del modello che incidono sulle prestazioni in contesti specializzati, e (iv) i prerequisiti necessari per un'integrazione efficace nei flussi di lavoro esistenti. Questa dissertazione presenta risultati che mettono in discussione diverse assunzioni comuni riguardo alle strategie ottimali di deployment degli LLM. In primo luogo, riscontriamo che la dimensione del modello non è direttamente correlata all'efficacia delle prestazioni. Modelli di piccole dimensioni raggiungono risultati paragonabili a sistemi di ordini di grandezza superiori dopo un adeguato adattamento al dominio. Il deployment locale mediante modelli open-weight può eguagliare le prestazioni di alternative proprietarie basate su cloud, garantendo al contempo una maggiore sovranità sui dati e indipendenza operativa. In secondo luogo, la nostra ricerca rivela complessità nell'interazione tra metodo e modello che non ammettono soluzioni universali, dimostrando che l'applicazione delle medesime strategie di adattamento può produrre risultati drasticamente diversi tra LLM differenti. In terzo luogo, osserviamo che i framework basati sulla confidenza offrono funzionalità di progettazione orientate alla sicurezza, garantendo al contempo maggiore protezione ed efficacia operativa, contrariamente all'aspettativa che esistano compromessi tra sicurezza ed efficacia. Dimostriamo inoltre che l'accettazione da parte degli utenti dipende dalla trasparenza del processo decisionale e dall'autonomia degli utenti, non solo dalle prestazioni complessive del sistema. Infine, delineiamo una serie di questioni aperte che riteniamo debbano essere affrontate dalla comunità scientifica, tra cui lo sviluppo di framework predittivi per la compatibilità tra modello e metodo, la calibrazione dinamica della confidenza attraverso diversi domini, il rilevamento sistematico e la mitigazione dei bias nei casi di fallimento degli ensemble, e meccanismi per garantire la stabilità delle prestazioni nel lungo periodo.