LUCA MORONI

CURRICULUM VITAE

EDUCATION AND TRAINING

01/11/2023 – IN CORSO | Roma, Italia

Ph.D. in Engineering in Computer Science

Sapienza Università di Roma

Livello EQF: EQF level 8

01/09/2021 – 21/07/2023 | Pisa, Italia

Laurea Magistrale in Computer Science (Artificial Intelligence)

Università di Pisa

Livello EQF: EQF level 7

01/09/2018 – 21/07/2021 | Perugia, Italia

Laurea Triennale in Informatica

Università di Perugia

Livello EQF: EQF level 6

PROJECTS

01/04/2021 – 01/10/2021

Machine Learning Applied to Cardiorespiratory Disease Detection Through Mobile Application

- Studio su come gli strumenti di Machine Learning (ML) possano essere applicati per rilevare malattie cardiorespiratorie, utilizzando il giroscopio e l'accelerometro dei dispositivi mobili in contesti a basse risorse.
- Sviluppo di un'applicazione mobile in grado di individuare e classificare le malattie cardiorespiratorie con un'accuratezza del ~90% sulla popolazione target.

PUBLICATIONS

2025

Optimizing LLMs for Italian: Reducing Token Fertility and Enhancing Efficiency Through Vocabulary Adaptation

Abstract: Il numero di Large Language Models (LLM) pre-addestrati è in costante aumento, sebbene la maggior parte sia progettata prevalentemente per la lingua inglese. In questo lavoro confrontiamo approfonditamente una varietà di tecniche di adattamento del vocabolario per ottimizzare gli LLM inglesi per la lingua italiana e proponiamo la *Semantic Alignment Vocabulary Adaptation (SAVA)*, un nuovo metodo che sfrutta la mappatura neurale per la sostituzione del vocabolario. SAVA ottiene prestazioni competitive in molteplici task a valle, migliorando le strategie di allineamento radicate. Abbiamo adattato due LLM: Mistral-7B-v0.1, riducendo la fertilità dei token del 25%, e Llama-3.1-8B, ottimizzando il vocabolario e riducendo il numero di parametri di 1 miliardo. Mostriamo che, in seguito all'adattamento del vocabolario, questi modelli possono recuperare le loro prestazioni con una fase relativamente limitata di addestramento continuo sulla lingua target. Infine, testiamo le capacità dei modelli adattati su vari task a scelta multipla e generativi.

Autori: *Luca Moroni, Giovanni Puccetti, Pere-Lluís Huguet Cabot, Andrei Stefan Bejgu, Alessio Miaschi, Edoardo Barba, Felice Dell'Orletta, Andrea Esuli, Roberto Navigli*

Rivista: *Findings of the Association for Computational Linguistics: NAACL 2025*

Editore: *Association for Computational Linguistics*

Multi-LMentry: Can Multilingual LLMs Solve Elementary Tasks Across Languages?

Abstract: Con il continuo miglioramento dei modelli linguistici di grandi dimensioni (LLM), la loro valutazione si concentra sempre più su task complessi e di alto livello, spesso a scapito di una valutazione sistematica delle capacità fondamentali. Per colmare questo divario, un lavoro recente ha proposto *LMentry*, un benchmark compatto che comprende task banali per gli esseri umani ma sorprendentemente difficili per gli LLM. Tuttavia, LMentry è limitato all'inglese, lasciando le sue intuizioni linguisticamente ristrette. In questo articolo presentiamo *Multi-LMentry*, una ricreazione da zero di LMentry che consente la valutazione sistematica degli LLM su task di ragionamento e comprensione di base in nove lingue diverse. Multi-LMentry include l'inglese ed si espande a basco, portoghese brasiliano, catalano, galiziano, tedesco, italiano, coreano e spagnolo, sottolineando l'importanza dei contesti cross-lingua e a basse risorse. Attraverso ampi esperimenti, riveliamo che gli LLM multilingue open-weight allo stato dell'arte non raggiungono ancora le prestazioni umane su task elementari in molte lingue. I nostri risultati mostrano nuove modalità di errore che rimangono nascoste nella valutazione monolingua, sottolineando la necessità di "unit test" rigorosi e linguisticamente diversificati delle capacità fondamentali dei modelli.

Autori: *Luca Moroni, Javier Aula-Blasco, Simone Conia, Irene Baucells, Naiara Perez, Silvia Paniagua Suárez, Anna Sallés, Malte Ostendorff, Júlia Falcão, Guijin Son, Aitor Gonzalez-Agirre, Roberto Navigli, Marta Villegas*

Rivista: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*

Editore: *Association for Computational Linguistics*

In the eyes of a language model: A comprehensive examination through eye-tracking data

Abstract: I segnali cognitivi, in particolare i dati di eye-tracking, offrono una prospettiva unica per comprendere l'elaborazione delle frasi umane. Sfruttando i dati dello sguardo della sezione inglese e italiana del *Multilingual Eye-Movement Corpus (MECO)*, abbiamo progettato una serie di esperimenti volti a esplorare se i modelli linguistici neurali (NLM) pre-addestrati codifichino pattern rappresentativi del comportamento di lettura umano e se l'incorporazione diretta di queste informazioni attraverso un processo di fine-tuning influenzi la plausibilità cognitiva del modello. I nostri risultati rivelano che i transformer codificano le informazioni relative allo sguardo durante il pre-addestramento e che l'integrazione esplicita delle caratteristiche di eye-tracking aumenta l'allineamento del modello con l'attenzione umana. In conclusione, la nostra valutazione completa degli NLM informati dai pattern di attenzione umana offre un grande potenziale per far progredire il campo dell'Intelligenza Artificiale Spiegabile (XAI).

Autori: *Luca Dini, Luca Moroni, Dominique Brunato, Felice Dell'Orletta*

Rivista: *Neurocomputing*

Editore: *Neurocomputing*

2024

Minerva LLMs: The First Family of Large Language Models Trained from Scratch on Italian Data

Abstract: La crescente popolarità dei Large Language Models (LLM) ha portato a un'ondata di ricerche sull'adattamento dei modelli esistenti a diverse lingue. Tuttavia, il pre-addestramento di LLM non inglesi è ancora un'area poco esplorata e non esiste un tentativo open-source che esplori ciò che è realizzabile con dati italiani aperti. Per affrontare questo problema, presentiamo *Minerva*, la prima famiglia di LLM addestrati da zero su dati italiani. La creazione di Minerva è un'opportunità per esplorare e studiare il pre-addestramento di LLM per la lingua italiana, delineando le sfide che emergono quando si addestrano LLM con testi nativi italiani. Minerva dimostra che un LLM per una lingua specifica porta una serie di vantaggi pratici rispetto all'adattamento di uno esistente, tra cui un controllo profondo sulla composizione del vocabolario e sui dati di addestramento. Con questo articolo, miriamo a fornire una panoramica completa delle scelte di progettazione, dei risultati e della valutazione dei nostri modelli Minerva, mostrando risultati promettenti sui benchmark italiani e sui task a valle.

Autori: *Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, Roberto Navigli*

Rivista: *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*

Editore: *CEUR Workshop Proceedings*