GABRIELE ROMANATO

Tipologie di dataset per il training degli LLM

I modelli linguistici di grandi dimensioni (Large Language Models, LLM) vengono addestrati su enormi quantità di dati testuali. La qualità, la varietà e la provenienza dei dataset utilizzati giocano un ruolo cruciale nelle capacità e nei limiti del modello. Di seguito vengono illustrate le principali tipologie di dataset impiegate nel training degli LLM.

1. Dati da Web Crawl

Una delle fonti più comuni è il crawling del web, ovvero la raccolta automatica di testi da siti web pubblicamente accessibili. Dataset come Common Crawl forniscono una base ampia e diversificata di contenuti, inclusi articoli, blog, forum e altro. Tuttavia, questi dati possono contenere rumore, duplicazioni e contenuti di bassa qualità, richiedendo filtraggio e pulizia.

2. Libri e Letteratura

I libri digitali, soprattutto quelli di pubblico dominio, sono utilizzati per esporre i modelli a un linguaggio ricco e strutturato. Dataset come The Pile includono collezioni di libri che spaziano dalla narrativa alla saggistica, offrendo contenuti complessi e stilisticamente variegati.

3. Enciclopedie e Documentazione Tecnica

Testi provenienti da enciclopedie libere (come Wikipedia) o documentazione tecnica (ad esempio, documenti scientifici e manuali) forniscono conoscenze strutturate, fatti e linguaggio formale. Queste fonti aiutano a migliorare la precisione e la capacità di rispondere a domande basate su conoscenze oggettive.

4. Conversazioni e Forum

I dati tratti da forum di discussione, chat pubbliche e piattaforme Q&A permettono al modello di apprendere stili linguistici colloquiali, domande frequenti e dinamiche conversazionali. Stack Exchange e Reddit sono esempi di fonti spesso utilizzate con opportune licenze e filtraggi.

5. Codice Sorgente

Per i modelli multimodali o focalizzati sulla programmazione, i repository di codice come GitHub sono una fonte chiave. Questi dataset includono esempi di codice, documentazione inline e discussioni tecniche che permettono al modello di comprendere linguaggi di programmazione e contesto tecnico.

6. Dati Proprietari e Curati

Oltre ai dati pubblicamente accessibili, molti LLM vengono addestrati anche su dati curati e proprietari, spesso raccolti o annotati manualmente. Questi includono dataset costruiti ad hoc per compiti specifici come il completamento di frasi, la traduzione, la classificazione o il ragionamento logico.

Conclusioni

La combinazione di diverse tipologie di dataset consente agli LLM di acquisire una comprensione ampia e flessibile del linguaggio. Tuttavia, la qualità e la rappresentatività dei dati rimangono fattori critici per garantire prestazioni elevate e una riduzione dei bias nei modelli finali.