Questo glossario, come tutto il progetto “E poi arrivò DeepSeek”, è la prosecuzione ideale del glossario di “In principio era ChatGPT”. Come allora, lo abbiamo scritto seguendo i principi di semplificazione del linguaggio – ci sono molte metafore che riguardano la vita reale per spiegare i termini tecnici – e di correttezza tecnica.
Il glossario di “In principio era ChatGPT” e quello di “E poi arrivò DeepSeek” sono un esempio di collaborazione fra persone e macchine: deciso lo schema e il modo in cui volevamo farlo, abbiamo usato le intelligenze artificiali generative per fare le bozze di ogni definizione. Questa volta è stato ancora più facile farlo, perché noi e le nostre macchine avevamo imparato dall’esperienza precedente e siamo cresciute insieme.

Benchmarking

Il benchmarking è il processo di misurazione e confronto delle prestazioni di un sistema, un prodotto o un'organizzazione rispetto a standard di riferimento o concorrenti. Nell’ambito dell’Intelligenza Artificiale, il benchmarking viene utilizzato per valutare l’efficacia di un modello rispetto ad altri, attraverso test su dataset standardizzati. Un’IA viene sottoposta a prove su set di dati specifici e confrontata con i risultati di altri modelli per misurare precisione, velocità, efficienza computazionale e altri parametri chiave.

Esistono diverse tipologie di benchmarking:

- competitivo: confronta un prodotto o un modello con i migliori competitor sul mercato.
- funzionale: analizza processi specifici rispetto a best practice del settore, indipendentemente dai concorrenti diretti.
- interno: mette a confronto diverse unità o moduli di una stessa organizzazione o sistema.
- generativo: valuta le capacità creative di un'IA rispetto a modelli preesistenti o alla produzione umana.

Nel caso dei modelli di IA, il benchmarking può essere immaginato come una gara a partire da un test: tutti i partecipanti partono dallo stesso problema (dataset di test) e vengono valutati in base a tempo, precisione, costi, qualità. Un buon benchmark non solo permette di individuare il modello più performante, ma aiuta anche a identificare le aree di miglioramento e a ottimizzare lo sviluppo delle nuove tecnologie.

Chain of thought

La chain of thought può essere vista proprio come uno studente che spiega al professore tutti i passaggi di un compito. Invece di limitarsi a dare la risposta finale, lo studente descrive passo dopo passo come è arrivato alla soluzione, spiegando il ragionamento, le operazioni fatte e le decisioni prese lungo il percorso. Questa approccio esplicativo aiuta non solo a chiarire la logica dietro la risposta finale, ma anche a verificare il processo che ha portato a quella conclusione. Allo stesso modo, l'AI che usa la Chain of Thought "racconta" i suoi pensieri intermedi, rendendo il ragionamento più trasparente e facilmente comprensibile.

Distillazione
La distillazione è un processo in cui un modello di intelligenza artificiale più grande e complesso (modello maestro) trasferisce la sua conoscenza a un modello più piccolo e leggero (modello studente). È un processo che, in qualche modo, ricorda i passaggi di conoscenza tra chi insegna e chi studia. Il processo di distillazione permette di ridurre la dimensione del modello senza compromettere troppo le prestazioni, migliorando l'efficienza computazionale e adattandosi a dispositivi con risorse limitate. Il processo di distillazione avviene in due fasi principali:

1) addestramento del modello maestro su un grande dataset
2) trasferimento delle conoscenze al modello studente, che impara dai risultati del maestro anziché dai dati grezzi.

La distillazione è utile per rendere i modelli più veloci, meno costosi da eseguire e facilmente implementabili in ambienti con risorse limitate, come dispositivi mobili o edge computing.

GPU (Graphic Processing Unit)
La **GPU** è un processore specializzato progettato per eseguire calcoli in parallelo su grandi quantità di dati. Inizialmente sviluppata per l'elaborazione grafica nei videogiochi e nelle applicazioni di rendering, la GPU è diventata un elemento chiave nell'intelligenza artificiale, nel machine learning e nel calcolo scientifico ad alte prestazioni. A differenza della CPU (Central Processing Unit), che gestisce una vasta gamma di operazioni sequenziali, una GPU è ottimizzata per elaborare molte operazioni contemporaneamente, grazie alla sua architettura parallela. Questo la rende particolarmente efficace nell'addestramento e nell'inferenza di modelli di deep learning, dove milioni di parametri devono essere aggiornati simultaneamente.

I vantaggi principali dell'uso di una GPU per l'IA sono la velocità di calcolo, l'efficienza nell'elaborazione di matrici (fondamentali per il deep learning), l'ottimizzazione dell'hardware. Immagina la GPU come una squadra di operai che lavorano contemporaneamente su un enorme mosaico, assemblando centinaia di pezzi nello stesso momento, mentre una CPU è più simile a un artigiano che lavora pezzo per pezzo. Le principali aziende produttrici di GPU per l'IA sono NVIDIA, AMD e Intel, con NVIDIA che domina il settore grazie alla sua architettura CUDA, specificamente ottimizzata per il calcolo AI. Per non perderci mai più i pezzi dal resto del mondo, però, facciamo una mappa veloce degli altri produttori.
In Cina ci sono Huawei, Biren Technology e Cambricon. In Europa Graphcore (Regno Unito) e SiPearl (Francia). A Taiwan ci sono la TSMC (che produce per conto di NVIDIA, AMD e Biren) e la MediaTek. In Corea del Sud la Samsung e la SK Hynix. In Russia Yadro e Sberbank.

Guardrail
Il termine Guardrail (letteralmente "barriera di protezione") nell'intelligenza artificiale si riferisce a misure di sicurezza e restrizioni progettate per limitare e guidare il comportamento di un sistema AI, prevenendo azioni indesiderate, dannose, illegali, non etiche. I guardrail sono un insieme di regole, filtri o controlli che impediscono all'AI di agire al di fuori degli obiettivi di sicurezza, etici e legali stabiliti dai suoi sviluppatori. Per esempio, quando si usa un assistente virtuale, un guardrail potrebbe impedire che l'IA risponda con contenuti inappropriati o offensivi, mantenendo l'interazione rispettosa e sicura.

In pratica, i guardrail fungono da limiti di sicurezza che proteggono sia gli utenti sia il sistema, garantendo che l'AI agisca in modo controllato, anche in scenari complessi. Possono essere implementati in vari modi, come filtri automatici, supervisione umana o meccanismi di feedback che correggono o limitano azioni potenzialmente dannose. I guardrail, come tutte le regole, sono socio-costruiti e storicamente connotati e contestualizzati.

Mixture of Experts (MoE)
Immagina un team di specialisti in un ospedale: invece di sottoporre ogni paziente a tutti i medici disponibili, il sistema indirizza ciascun caso agli specialisti più competenti per il problema specifico. Il Mixture of Experts (MoE) si basa sul medesimo principio di divisione dei compiti: è un'architettura di rete neurale che suddivide il lavoro tra diversi esperti, ovvero sottoreti specializzate, per migliorare l'efficienza e le prestazioni di un modello di intelligenza artificiale. A differenza delle reti neurali tradizionali, in cui tutti i parametri sono attivati per ogni input, un modello MoE attiva solo alcuni esperti in base al tipo di dato in ingresso. Questo viene gestito da un gate, un meccanismo che decide quali esperti coinvolgere per ogni specifico compito.

Il vantaggio di questa architettura sta nel fatto che:

- aumenta la scalabilità: modelli con miliardi di parametri possono operare attivando solo una frazione della rete, riducendo il costo computazionale;
- migliora la specializzazione: ogni esperto si allena su un sottoinsieme specifico di dati, rendendo il modello più efficace nel gestire compiti differenti.
- ottimizza il consumo energetico: elaborando solo le informazioni necessarie, MoE riduce l'uso di risorse rispetto a un modello completamente denso.

Open source (nelle intelligenze artificiali)
L'intelligenza artificiale open source è un'IA il cui codice sorgente, i pesi dei modelli e i datii di addestramento sono resi accessibili a chiunque, permettendo di studiarne il funzionamento, modificarla e condividerla liberamente. Puoi immaginare un'IA open source come una ricetta di cucina completamente accessibile e completa: non solo puoi assaggiare il piatto, ma hai anche l'elenco completo degli ingredienti e delle loro quantità e le esatte istruzioni per cucinarla in autonomia. Chiaramente, puoi anche modificarla a tuo piacimento. Se, invece, hai solo il piatto pronto ma non conosci la ricetta, puoi gustarlo, ma non saprai mai come rifarlo esattamente.

Secondo la Open Source Initiative (OSI)[1], per essere davvero open source, un sistema di IA deve garantire alcune libertà fondamentali:

- libertà di utilizzo: chiunque deve poter usare l'IA per qualsiasi scopo, senza restrizioni arbitrarie.
- libertà di studio: il codice sorgente e i dati di addestramento devono essere accessibili, per permettere di comprendere come funziona il modello.

[1] https://opensource.org/ai/open-source-ai-definition

- libertà di modifica: è possibile adattare il modello alle proprie esigenze, migliorarlo o correggerne eventuali problemi.
- libertà di condivisione: le versioni modificate possono essere ridistribuite, mantenendo aperta l'evoluzione del modello

Molti modelli vengono definiti "aperti", ma in realtà non lo sono del tutto: alcuni permettono solo di scaricare i pesi del modello senza offrire accesso ai dati di addestramento o alle metodologie di training, rendendo difficile verificarne il funzionamento o riprodurli completamente.
L'open source nell'IA è una scelta politica oltre che tecnica: apre la porta alla trasparenza, alla collaborazione e alla decentralizzazione dello sviluppo, contrastando il monopolio delle grandi aziende e permettendo una maggiore democratizzazione delle tecnologie intelligenti.

Reinforcement learning
Il Reinforcement Learning (RL) è un metodo di addestramento in cui un agente impara interagendo con un ambiente e ricevendo ricompense o penalità in base alle sue azioni. L'obiettivo è ottimizzare il comportamento per massimizzare la ricompensa nel lungo termine. Per esempio, un'IA che gioca a scacchi impara dalle sue partite, premiando mosse vincenti e penalizzando quelle che portano alla sconfitta.

Supervised Fine-Tuning
Il Supervised Fine-Tuning è una tecnica di addestramento in cui un modello di intelligenza artificiale pre-addestrato viene raffinato su un dataset etichettato. Durante questo processo, l'IA apprende associazioni specifiche tra input e output corretti, migliorando le sue prestazioni per un compito mirato. Per esempio, un modello linguistico generale viene addestrato su dati specifici di medicina, etichettati da esseri umani, per diventare un assistente AI per dottori.