Glossario della Terminologia AI

• **Agente AI (AI Agent)**: Un'estensione delle funzionalità di un Large Language Model (LLM) che combina capacità di elaborazione del linguaggio naturale con componenti aggiuntivi. Permette all'LLM di eseguire compiti complessi, interagire con l'ambiente esterno e operare in modo autonomo. Gli agenti AI possono includere motori di ricerca per accesso a informazioni in tempo reale, API per prenotazioni o transazioni, strumenti computazionali per calcoli e analisi dati, e database per recuperare o memorizzare informazioni personalizzate. Esistono anche **Multi-Agenti LLM**, sistemi composti da più agenti autonomi che interagiscono tra loro per raggiungere obiettivi individuali o collettivi, sfruttando diversi modelli linguistici per creatività, etica, dati strutturati e assegnazione di ruoli specifici.

• **Algoritmi Generativi (Generative Algorithms)**: Algoritmi capaci di generare nuovi contenuti, a differenza di quelli che eseguono operazioni come classificazione o previsione. GPT è un esempio di algoritmo generativo.

• **AI Act**: Una normativa presentata dalla Commissione Europea nell'aprile 2021 che stabilisce un quadro normativo sull'IA, classificando i rischi e introducendo nuovi obblighi per le imprese che operano nei 27 Stati membri dell'Unione Europea. Mira a proteggere i diritti e le libertà dei cittadini promuovendo innovazione e imprenditorialità. Classifica la tecnologia in categorie di rischio (inaccettabile, alto, medio, basso) e include divieti su determinati casi d'uso, come i sistemi di riconoscimento delle emozioni sul luogo di lavoro. In Italia, il regolamento è stato pubblicato nella Gazzetta Ufficiale dell'Unione Europea il 12 luglio 2024 e la strategia italiana per l'Intelligenza Artificiale 2024-2026 è stata elaborata dal Dipartimento per la trasformazione digitale.

• **Allucinazioni (Hallucinations)**: Fenomeno in cui i modelli AI producono contenuti che sembrano plausibili ma sono in realtà inesatti o privi di fondamento.

• **API (Application Programming Interface)**: L'acronimo di "Application Programming Interface". Le API sono insiemi di definizioni e protocolli che consentono di costruire e integrare software applicativo. Una **API Key** è una stringa alfanumerica utilizzata per controllare l'accesso a un'API. Gli LLM possono essere utilizzati tramite API.

• **Attention Block (Blocco di Attenzione)**: Componente fondamentale del modello Transformer che permette al modello di pesare l'importanza di diverse parole in una sequenza rispetto ad altre, contribuendo a comprendere il contesto semantico e le relazioni a lungo raggio tra le parole.

• **Autoencoder**: Una architettura di rete neurale che riproduce l'input con meno informazioni, utile per generare nuovi contenuti o per la generazione di testo basata su domande.

• **Bias**: Preconcetti linguistici e culturali presenti nei dati di addestramento iniziali di un modello AI che possono portare a discriminazioni e contenuti inappropriati. Un esempio è l'indagine sui bias contro donne e ragazze nei Large Language Models.

• **Black Box (Scatola Nera)**: Si riferisce al fatto che, una volta allenata, una rete neurale è descritta dai suoi parametri, ma non è possibile risalire a cosa sia stato usato per l'allenamento, rendendola una "scatola nera". L'interpretazione dei processi interni di un LLM è complessa, il che rende difficile capire il ragionamento dietro le risposte.

• **CAG (Cache-Augmented Generation)**: Un approccio che prevede il pre caricamento e la memorizzazione in cache di una base di conoscenza fissa all'interno della finestra di contesto dell'LLM, eliminando il recupero in tempo reale delle informazioni. Il vantaggio principale è la **grande rapidità**, ma ha il contro di un corpus documentale limitato e poco modificabile.

• **Chain of Thoughts (CoT)**: Una tecnica per risolvere problemi complessi con gli LLM che consiste nell'indurre il modello a generare una "catena di pensieri", ovvero una sequenza di passaggi intermedi prima di fornire la risposta finale. Si può chiedere al modello di "lavorare passo-passo" o di "adottare un approccio passo-passo".

• **ChatGPT**: Un chatbot basato su GPT-3, lanciato da OpenAI il 30 novembre 2022, che ha generato un'ampia discussione sull'IA. Ha raggiunto 100 milioni di utenti registrati nei primi due mesi.

• **Distillation (Distillazione)**: Una tecnica che permette di trasferire conoscenza da un modello "grande" (teacher) a un modello "piccolo" (student) affinando il modello piccolo con informazioni generate dal modello grande. I vantaggi includono **efficienza** (il modello student è più veloce e richiede meno risorse), mantenimento delle prestazioni (spesso conserva gran parte dell'accuratezza del teacher) e **adattabilità** (può essere applicata a diversi LLM come BERT, GPT).

• **Embedding**: La rappresentazione numerica delle parole in uno spazio semantico multidimensionale. Le parole sono rappresentate come vettori, dove ogni dimensione del vettore può acquisire un significato (caratteristica). La quantità di dimensioni necessarie per rappresentare tutte le sfumature di significato della lingua può essere molto elevata (es. GPT-3-175B ha 12.288 dimensioni).

• **Fine-tuning**: Un processo in cui un modello di machine learning pre-addestrato (come GPT, Llama, Deepseek) viene ulteriormente addestrato su un dataset specifico per adattarsi a un compito o dominio particolare. Non aggiunge nuove conoscenze significative, ma modella il comportamento delle risposte per allinearlo a un contesto specifico.

◦ **Supervised Fine-Tuning (SFT)**: Addestramento supervisionato su un dataset etichettato per uno scopo specifico, come migliorare la rilevanza delle risposte per i clienti.

◦ **Parameter-Efficient Fine-Tuning (PEFT)**: Tecniche che aggiornano solo un sottoinsieme dei parametri del modello, preservando la maggior parte dei pesi pre-addestrati. Riduce i costi computazionali e previene l'"oblio catastrofico".

• **Generative Pre-trained Transformer (GPT)**: Una classe di algoritmi generativi basati sull'architettura Transformer.

◦ **Generative**: Un modello che può generare nuovi contenuti.

◦ **Pre-trained**: Un modello pre-allenato per scopi generici che può essere affinato (fine-tuned) per scopi specifici.

◦ **Transformer**: Un modello creato utilizzando una rete neurale Transformer.

• **Intelligenza Artificiale (AI)**: Un insieme di algoritmi e procedure che permette a un computer di eseguire compiti "intelligenti", tipicamente umani, come prendere decisioni, riconoscere modelli, creare e imparare. È una disciplina vasta e "antica".

• **Large Language Models (LLM)**: Modelli di linguaggio che, con l'introduzione dei Transformer, sono in grado di gestire testi molto lunghi e molteplici sfumature di significato. La loro capacità è spesso correlata al numero di parametri. Sono in grado di creare nuovo testo, classificare, riassumere, rispondere a domande e tradurre. Possono generare testo, una parola alla volta, basandosi sulla probabilità della parola successiva. Esistono numerosi LLM oltre a GPT, come Llama, Gemini, Deepseek, Mistral, Claude, Qwen.

• **Machine Learning (ML)**: Un sottogruppo dell'Intelligenza Artificiale. Un algoritmo di ML esegue un addestramento tramite dati iniziali, crea un modello tramite l'addestramento e applica il modello a nuovi dati. Richiede un "training set" etichettato.

• **Modello di Linguaggio (Language Model - LM)**: Una rappresentazione del linguaggio che permette di creare testo coerente e "con significato". I modelli di linguaggio possono essere probabilistici, il che significa che possono dare risposte diverse alla stessa domanda a causa di un parametro chiamato **Temperatura**. La temperatura (T) è un parametro applicato alla funzione softmax che il modello usa per ottenere le probabilità delle parole, influenzando la casualità dell'output.

• **Multilayer Perceptron (MLP)**: Un tipo di rete neurale in cui ogni parola viene influenzata dalla conoscenza pregressa del modello.

• **Neural Networks (Reti Neurali)**: Un tipo di IA ispirato alla struttura del cervello umano, composto da "neuroni" semplici e connessi tra loro che possono dare origine a comportamenti complessi. Possono essere addestrate per vari compiti, come il riconoscimento della scrittura.

• **Open Source LLM**: Modelli di linguaggio con codice sorgente aperto. Offrono vantaggi come una maggiore privacy e sicurezza dei dati sensibili, la possibilità di essere installati su server locali e maggiore flessibilità. Esempi includono Llama, Gemma, Deepseek, Qwen, Mistral. Esistono comunità come Hugging Face che offrono modelli, dataset e documentazione. È possibile utilizzare LLM open source in locale o come servizio API.

• **Prompt Engineering**: L'arte di formulare input efficaci (prompt) per gli LLM per ottenere le risposte desiderate. Un buon prompt è la chiave per l'utilizzo di un LLM. Elementi di un buon prompt includono specificare un ruolo, indicare lunghezza/tono/stile, specificare bene il contenuto da produrre e fornire esempi. Esistono "prompt hack" per migliorare le risposte, come chiedere all'AI di spiegare cosa farà o di fare un'analisi critica del suo output.

• **RAG (Retrieval-Augmented Generation)**: tecnologia ibrida che combina un sistema di recupero di informazioni (da database esterni, documenti, web) con un modello generativo per produrre risposte coerenti, precise e contestualizzate. A differenza dei normali sistemi di intelligenza artificiale che rispondono basandosi solo sulla loro "memoria" preesistente, il RAG prima cerca informazioni aggiornate in database, documenti o sul web, poi usa un modello generativo per elaborare queste informazioni e produrre risposte precise e sempre attuali. Il grande vantaggio è che può fornire risposte basate sui dati più recenti, superando il limite delle conoscenze "congelate nel tempo" tipiche degli altri sistemi. Un esempio concreto è NotebookLM, che usa proprio questa tecnologia per analizzare e rispondere basandosi su documenti sempre aggiornati.

• **Reinforcement Learning (Apprendimento per Rinforzo)**: Una metodologia di apprendimento automatico in cui il modello viene "premiato" quando trova la soluzione a un problema e si auto-regola per massimizzare i premi.

• **Token**: Nei modelli linguistici (LLM), i token rappresentano le unità base di testo (parole, caratteri o frammenti) e svolgono un duplice ruolo: da un lato misurano la quantità di testo elaborabile (limite del "contesto", dato da input + output), dall'altro determinano il costo d'uso dei servizi, solitamente calcolato in euro per token.

• **Training (Addestramento)**: Il processo di determinare i valori dei parametri di una rete neurale. I "buoni" valori sono quelli che permettono di ottenere output giusti su input noti.