IBM Italy News Room

IBM presenta Granite 3.0: modelli di AI ad alte prestazioni creati per il business

• I nuovi modelli Granite 3.0 8B e 2B, rilasciati sotto licenza Apache 2.0, offrono prestazioni innovative nei benchmark accademici e aziendali, superando o eguagliando modelli di dimensioni simili.

• I nuovi modelli Granite Guardian 3.0 offrono le funzionalità di protezione più complete per un'AI sicura e affidabile.

• I nuovi modelli Granite 3.0 Mixture-of-Experts consentono un'inferenza estremamente efficiente e una bassa latenza, adatta a implementazioni basate su CPU e sull'edge computing.

• Il nuovo modello Granite Time Series garantisce prestazioni all'avanguardia, superiori a quelle di modelli 10 volte più grandi

• IBM presenta la nuova generazione di watsonx Code Assistant, basato su Granite, per il supporto alla generazione di codice; introduce nuovi strumenti per la creazione e la distribuzione di applicazioni e agenti AI.

• Annuncio di Granite come modello predefinito di Consulting Advantage, una piattaforma di delivery basata sull'intelligenza artificiale e utilizzata da 160.000 consulenti IBM per offrire ai clienti nuove soluzioni più velocemente.

ARMONK, NY – 21 ottobre 2024 –Oggi**,** in occasione dell'evento annuale TechXchange di IBM (NYSE:IBM) è stata annunciata Granite 3.0: la famiglia di modelli di intelligenza artificiale finora più avanzata. I modelli linguistici di terza generazione Granite di IBM superano o eguagliano sui principali benchmark le prestazioni dei modelli di dimensioni simili dei principali fornitori di modelli con prestazioni, trasparenza e sicurezza migliori della categoria.

In linea con l'impegno dell'azienda nei confronti dell'AI open-source, i modelli Granite sono rilasciati con una licenza Apache 2.0 senza restrizioni, che li rende unici per la combinazione di prestazioni, flessibilità e diritti che offre alle imprese e alla comunità.

La famiglia Granite 3.0 di IBM include:

- Utilizzo generale/Linguaggio: **Granite 3.0 8B-Instruct**, **Granite 3.0 2B-Instruct, Granite 3.0 8B Base, Granite 3.0 2B Base**
- Protezione e sicurezza: **Granite Guardian 3.0 8B**, **Granite Guardian 3.0 2B** •
- Mix di Esperti: **Granite 3.0 3B A800M Instruct**, **Granite 3.0 1B A400M Instruct, Granite 3.0 3B A800M Base, Granite 3.0 1B A400M Base**

I nuovi modelli Granite 8B e 2B sono stati progettati come modelli solidi e affidabili per l'AI in ambito aziendale, in grado di offrire prestazioni all'avanguardia e costi contenuti per attività quali RAG (Retrieval Augmented Generation), classificazione, riepilogo, estrazione di entità e utilizzo di strumenti. Questi modelli compatti e versatili sono progettati per essere addestrati con i dati aziendali e integrati in modo fluido in qualsiasi ambiente aziendale o flusso di lavoro.

Mentre la maggior parte dei modelli linguistici di grandi dimensioni (LLM) viene invece addestrata su dati pubblici, la gran parte dei dati aziendali non viene sfruttata. Combinando Granite con i dati aziendali e utilizzando metodi di riaddestramento come InstructLab, - introdotto da IBM e RedHat a maggio - IBM ritiene che le aziende possano ottenere prestazioni specifiche in grado di competere con modelli più grandi a una frazione del costo (sulla base di una gamma osservata di costi inferiori di 3-23 volte rispetto ai modelli di frontiera di grandi dimensioni in diversi iniziali proof-of-concept)[(1)].

La pubblicazione di Granite 3.0 riafferma l'impegno di IBM per la trasparenza, la sicurezza e l'attendibilità. La relazione tecnica di Granite e la guida all'uso responsabile forniscono un'ampia documentazione dei set di dati utilizzati per addestrare questi modelli, i dettagli delle fasi di filtraggio, pulizia e cura applicate e i dati completi sulle prestazioni dei modelli rispetto ai principali benchmark accademici e aziendali.

In particolare, IBM fornisce una garanzia di proprietà intellettuale per tutti i modelli Granite su watsonx.ai, in modo che i clienti i possano essere più sicuri di integrare i loro dati ai modelli.

Nuovi traguardi: i benchmark di Granite 8B e 2B

I modelli linguistici Granite stanno dimostrando anche risultati promettenti in termini di prestazioni.

Sui benchmark accademici standard definiti dalla OpenLLM Leaderboard di Hugging Face, le prestazioni complessive del modello Granite 3.0 8B Instruct sono in media superiori a quelle dei modelli open-source di dimensioni simili di Meta e Mistral. Nel benchmark di sicurezza AttaQ di IBM, il modello Granite 3.0 8B Instruct è in testa in tutte le dimensioni della sicurezza rispetto ai modelli Meta e Mistral[(2)].

Su attività di core business come il Retrieval Augmented Generation, e di Cybersecurity, il modello Granite 8B mostra in media prestazioni complessivamente migliori rispetto ai modelli open-source di dimensioni simili di Mistral e Meta[(3)].

I modelli Granite 3.0 sono stati addestrati su oltre 12 trilioni di token e su dati provenienti da 12 lingue e 116 linguaggi di programmazione diversi, utilizzando un nuovo metodo di addestramento in due fasi, avvalendosi dei risultati di diverse migliaia di esperimenti concepiti per ottimizzare la qualità e la selezione dei dati e i parametri di addestramento. Entro la fine dell'anno, i modelli 8B e 2B includeranno anche il supporto per la lunghezza del contesto estesa a 128K e le capacità di comprensione multimodale dei documenti.

Dimostrando un eccellente equilibrio tra prestazioni e costi di inferenza, IBM offre i suoi modelli con architettura Granite Mixture of Experts (MoE), Granite 1B A400M e Granite 3B A800M più piccoli e leggeri che possono essere utilizzati per applicazioni a bassa latenza e per implementazioni basate su CPU.

IBM annuncia anche una versione aggiornata dei suoi modelli Granite pre-addestrati, le cui prime versioni sono state rilasciate all'inizio di quest'anno. Questi nuovi modelli sono addestrati su un numero di dati 3 volte superiore e offrono prestazioni ineguagliabili su benchmark di serie temporali, superando modelli 10 volte più grandi come quelli di Google e Alibaba. I modelli aggiornati offrono inoltre una maggiore flessibilità di modellazione, grazie al supporto di variabili esterne e previsioni continue[(4)].

Granite Guardian 3.0: una nuova frontiera per l'AI etica

Nell'ambito di questa release, IBM introduce anche una nuova famiglia di modelli Granite Guardian che consentono agli sviluppatori di applicazioni di implementare barriere di sicurezza controllando i prompt degli utenti e le risposte dell'LLM rispetto

ad una serie di rischi. I modelli Granite Guardian 8B e 2B offrono la serie più completa di funzionalità di rilevamento dei rischi e dei danni oggi disponibile sul mercato.

Oltre alle misurazioni di aspetti dannosi legati ad esempio al pregiudizio sociale, odio, tossicità, blasfemia, violenza, jailbreaking e altro ancora, questi modelli forniscono anche una serie di controlli unici e specifici per i RAG, come la fondatezza, la rilevanza del contesto e la rilevanza delle risposte. Nei test approfonditi condotti su oltre 19 benchmark di sicurezza e RAG, il modello Granite Guardian 3.0 8B ha ottenuto un'accuratezza complessiva nel rilevamento di questi aspetti superiore in media a tutte e tre le generazioni di modelli Llama Guard di Meta. Inoltre, ha mostrato prestazioni complessive pari a quelle dei modelli specializzati nel rilevamento delle allucinazioni WeCheck e MiniCheck[(5)].

Sebbene i modelli Granite Guardian siano derivati dai corrispondenti modelli linguistici Granite, possono essere utilizzati da chiunque per implementare protezioni da abbinare a qualsiasi modello di intelligenza artificiale open o proprietario.

Disponibilità dei modelli Granite 3.0

L'intera suite di modelli Granite 3.0 e i modelli time series aggiornati sono disponibili per il download su HuggingFace sotto la licenza senza restrizioni Apache 2.0. Le varianti di istruzione dei nuovi modelli linguistici Granite 3.0 8B e 2B e i modelli Granite Guardian 3.0 8B e 3B sono disponibili da oggi per uso commerciale sulla piattaforma IBM watsonx. Una selezione dei modelli Granite 3.0 sarà disponibile anche come microservizi NVIDIA NIM e attraverso le integrazioni Vertex AI Model Garden di Google Cloud con HuggingFace.

Per aiutare gli sviluppatori a scegliere, a semplificare l'uso e a supportare le implementazioni locali, una serie curata di modelli Granite 3.0 è disponibile anche su Ollama e Replicate.

L'ultima generazione di modelli Granite amplia il solido catalogo open-source di IBM con potenti LLM adatti allo scopo. IBM ha collaborato con partner dell'ecosistema come AWS, Docker, Domo, Qualcomm Technologies, Inc. attraverso il suo Qualcomm® AI Hub , Salesforce, SAP e altri per integrare i modelli Granite nelle offerte di questi partner o per rendere i modelli Granite disponibili sulle loro piattaforme, offrendo una maggiore scelta alle aziende di tutto il mondo.

Dagli Assistenti agli agenti: il futuro dell'AI per le imprese

IBM sta facendo progredire l'AI per le imprese attraverso uno spettro di tecnologie che varia dai modelli, agli assistenti e strumenti necessari per mettere a punto e distribuire l'AI basandosi sui dati e i casi d'uso specifici delle aziende.

IBM sta anche aprendo la strada a futuri agenti di intelligenza artificiale in grado di auto-dirigersi, riflettere ed eseguire compiti complessi in ambienti aziendali dinamici.

IBM continua a far evolvere il suo portafoglio di tecnologie per assistenti AI: da watsonx Orchestrate, che aiuta le aziende a creare i propri assistenti tramite strumenti e automazione low-code, a un'ampia gamma di assistenti precostituiti per attività e settori specifici come il servizio clienti, le risorse umane, le vendite e il marketing.

Le organizzazioni di tutto il mondo hanno utilizzato watsonx Assistant per costruire assistenti AI per attività come rispondere alle domande di routine dei clienti o dei dipendenti, modernizzare i loro mainframe e le applicazioni IT legacy, aiutare gli studenti a esplorare potenziali percorsi di carriera o fornire assistenza digitale per i mutui agli acquirenti di case.

Oggi IBM ha presentato la nuova generazione di watsonx Code Assistant, alimentata dai modelli di codice Granite, per offrire un'assistenza generica alla codifica in linguaggi come C, C++, Go, Java e Python, con funzionalità di modernizzazione delle applicazioni avanzate per le applicazioni Java aziendali[6].
Le funzionalità di codice di Granite sono ora accessibili anche attraverso un'estensione di Visual Studio Code, IBM IBM Granite.Code.

IBM sta inoltre introducendo nuovi strumenti per aiutare gli sviluppatori a costruire, personalizzare e distribuire l'AI in modo più efficiente tramite watsonx.ai, tra cui framework per agire in autonomia, integrazioni con ambienti esistenti e automazioni low-code per casi d'uso comuni come RAG e agenti[7].

IBM sta inoltre sviluppando tecnologie di agenti AI in grado di garantire una maggiore autonomia, un ragionamento sofisticato e la risoluzione di problemi in più fasi. La versione iniziale del modello Granite 3.0 8B è dotata di supporto per le principali funzioni per agire in autonomia, come il ragionamento avanzato e un modello di chat altamente strutturato e uno stile di prompting per l'implementazione di flussi di lavoro per l'utilizzo di strumenti. IBM sta inoltre introducendo una nuova funzionalità di chat con agenti AI in IBM watsonx Orchestrate, che utilizza le funzionalità per agire in autonomia per orchestrare assistenti AI, competenze e automazioni che aiutino gli utenti ad aumentare la produttività dei loro team[8]. Nel 2025 IBM continuerà a sviluppare funzionalità di agenti in tutto il suo portafoglio, compresi agenti precostituiti per domini e casi d'uso

specifici.

Ampliata la piattaforma di delivery basata sull'AI a disposizione dei consulenti IBM

IBM annuncia anche un'importante espansione della sua piattaforma di delivery basata sull'AI, IBM Consulting Advantage. La piattaforma multi-modello contiene agenti AI, applicazioni e metodi come framework ripetibili che consentono a 160.000 consulenti IBM di fornire ai clienti un valore migliore e più rapido a un costo inferiore.

Come parte di questo ampliamento i modelli linguistici di Granite 3.0 diventeranno il modello predefinito in Consulting Advantage. Grazie alle prestazioni e all'efficienza di Granite, IBM Consulting sarà in grado di massimizzare il ritorno sull'investimento per i progetti GenAI dei propri clienti.

Un'altra parte fondamentale è l'introduzione di IBM Consulting Advantage for Cloud Transformation and Management e IBM Consulting Advantage for Business Operations. Entrambi comprendono agenti, applicazioni e metodi di AI specifici per ogni settore, integrati con la proprietà intellettuale e le best practice di IBM, in modo che i consulenti possano accelerare le trasformazioni cloud e AI dei clienti su attività quali la modernizzazione del codice, la progettazione di qualità o la trasformazione e l'esecuzione di operazioni in settori come la finanza, le risorse umane e il procurement.

Per saperne di più su Granite e sulla strategia IBM AI for Business, visitate il sito https://www.ibm.com/granite

LinkedIn: IBM

Contatti:
Morgana Stell - *External Relations Leader, IBM Italia*
email: morgana.stell@it.ibm.com
mobile: 335 7693528

(1) - I calcoli dei costi si basano sui prezzi del costo API per milione di token di IBM watsonx per i modelli aperti e openAI dei modelli GPT4 (ipotizzando una mix di 80% inout, 20% output) per i proof-of-concept dei clienti

(2) - IBM Research technical paper: Granite 3.0 Language Models

(3) - IBM Research technical paper: Granite 3.0 Language Models

(4) - The Tiny Time Mixer: Fast Pre-Trained Models for Enhanced Zero/Few Shot Forecasting on Multivariate Time Series

(5) - Risultati della valutazione pubblicati nel report Granite Guardian GitHub

(6) - Disponibilità prevista 1Q 2025

(7) - Disponibilità prevista 1Q 2025

(8) - Disponibilità prevista 1Q 2025

---

https://it.newsroom.ibm.com/ibm-granite3