IBM DACH Newsroom

IBM stellt Granite 3.0 vor: Hochleistungsfähige, vertrauenswürdige KI-Modelle für Unternehmen

• Die neuen Granite 3.0 8B- und 2B-Modelle, die unter einer uneingeschränkten Apache 2.0-Lizenz veröffentlicht wurden, zeigen in vielen akademischen und Unternehmens-Benchmarks eine starke Leistung und können Modelle ähnlicher Größe übertreffen oder mit ihnen gleichziehen

• Die neuen Granite Guardian 3.0-Modelle bieten die umfassendsten Leitplankenfunktionen für sichere und vertrauenswürdige KI

• Die neuen Granite 3.0 Mixture-of-Experts-Modelle ermöglichen extrem effiziente Inferenzen und niedrige Latenzzeiten und eignen sich für CPU-basierte Implementierungen und Edge-Computing

• Das neue Granite Time Series-Modell erreicht Spitzenleistung bei Zero/Few-Shot-Prognosen und übertrifft 10-mal größere Modelle

• IBM präsentiert die nächste Generation des Granite-gestützten watsonx Code Assistant für allgemeine Codierung; stellt neue Tools in watsonx.ai für die Entwicklung und den Einsatz von KI-Anwendungen und Agenten vor

• Ankündigung von Granite als Standardmodell von Consulting Advantage, einer KI-gestützten Bereitstellungsplattform, die von den 160.000 IBM Berater_innen genutzt wird, um Kund_innen schneller neue Lösungen zu bieten

ARMONK, N.Y., Oct. 21, 2024 /PRNewswire/ -- Auf der jährlichen TechXchange-Veranstaltung von IBM (NYSE:IBM) gab das Unternehmen heute die Veröffentlichung seiner bisher fortschrittlichsten Familie von KI-Modellen bekannt, Granite 3.0. IBMs Granite-Sprachmodelle der dritten Generation übertreffen bei vielen akademischen und industriellen Benchmarks ähnlich große Modelle von führenden Modellanbietern und zeichnen sich durch hohe Leistung, Transparenz und Sicherheit aus.

Im Einklang mit dem Engagement des Unternehmens für Open-Source-KI werden die Granite-Modelle unter einer uneingeschränkten Apache 2.0-Lizenz veröffentlicht, was sie in ihrer Kombination aus Leistung, Flexibilität und Autonomie, die sie Unternehmenskunden und der Community insgesamt bieten, einzigartig macht.

Die Granite 3.0-Familie von IBM umfasst:

- Allgemeine Anwendungen/Sprachen: **Granite 3.0 8B-Instruct**, **Granite 3.0 2B-Instruct, Granite 3.0 8B Base, Granite 3.0 2B Base**
- Leitplanken und Sicherheit: **Granite Guardian 3.0 8B**, **Granite Guardian 3.0 2B**
- Mixture-of-Experts: **Granite 3.0 3B A800M Instruct**, **Granite 3.0 1B A400M Instruct, Granite 3.0 3B A800M Base**, **Granite 3.0 1B A400M Base**

Die neuen Granite 8B- und 2B-Modelle sind als „Arbeitspferde“ für die KI in Unternehmen konzipiert und bieten modernste Leistung und Kosteneffizienz für Aufgaben wie Retrieval Augmented Geneneration (RAG), Klassifizierung, Zusammenfassung, Entitätsextraktion und Toolnutzung. Diese kompakten, vielseitigen Modelle sind für die Feinabstimmung mit Unternehmensdaten konzipiert und lassen sich nahtlos in jede Geschäftsumgebung und jeden Arbeitsablauf integrieren.

Während viele große Sprachmodelle (LLMs) auf öffentlich verfügbaren Daten trainiert werden, bleibt ein Großteil der Unternehmensdaten ungenutzt. Durch die Kombination eines kleinen Granite-Modells mit Unternehmensdaten, insbesondere unter Verwendung der revolutionären Alignment-Technik InstructLab - die von IBM und RedHat im Mai vorgestellt wurde - können Unternehmen nach Ansicht von IBM eine aufgabenspezifische Leistung erzielen, die mit größeren Modellen konkurriert, und das zu einem Bruchteil der Kosten (basierend auf einer beobachteten Spanne von 3- bis 23-fach geringeren Kosten als bei großen Frontier-Modellen in mehreren frühen Proof-of-Concept[1]).

Die Veröffentlichung von Granite 3.0 unterstreicht das Engagement von IBM, Transparenz, Sicherheit und Vertrauen in KI-Produkte zu schaffen. Der technische Bericht zu Granite 3.0 und der Leitfaden zur verantwortungsvollen Nutzung enthalten eine Beschreibung der Datensätze, die zum Trainieren dieser Modelle verwendet wurden, Einzelheiten zu den angewandten Filter-, Bereinigungs- und Kuratierungsschritten sowie umfassende Ergebnisse der Modellleistung in wichtigen akademischen und Unternehmens-Benchmarks.

Entscheidend ist, dass IBM eine IP-Entschädigung für alle Granite-Modelle auf watsonx.ai anbietet, damit Unternehmenskunden ihre Daten vertrauensvoll mit den Modellen zusammenführen können.

Die Messlatte höher legen: Granite 8B- und 2B-Benchmarks

Die Granite 3.0-Sprachmodelle zeigen auch bei der reinen Leistung vielversprechende Ergebnisse.

Bei akademischen Standard-Benchmarks, die vom Hugging Face OpenLLM Leaderboard definiert wurden, liegt die Gesamtleistung des Granite 3.0 8B Instruct-Modells im Durchschnitt über der Leistung von Open-Source-Modellen ähnlicher Größe von Meta und Mistral. Bei IBMs modernstem Sicherheitsbenchmark AttaQ ist das Granite 3.0 8B Instruct Modell in allen gemessenen Sicherheitsdimensionen führend im Vergleich zu Modellen von Meta und Mistral.[2]

Bei den zentralen Unternehmensaufgaben der RAG, der Werkzeugnutzung und den Aufgaben im Bereich Cybersecurity zeigt das Granite 3.0 8B Instruct Modell im Vergleich zu ähnlich großen Open-Source-Modellen von Mistral und Meta eine durchschnittlich führende Leistung.[3]

IBM bietet seine Granite Mixture of Experts (MoE) Architecture-Modelle, Granite 3.0 1B-A400M und Granite 3.0 3B-A800M, als

kleinere, leichtgewichtige Modelle an, die sowohl für Anwendungen mit geringer Latenz als auch für CPU-basierte Implementierungen eingesetzt werden können und ein hervorragendes Gleichgewicht zwischen Leistung und Inferenzkosten bieten.

IBM kündigt außerdem eine aktualisierte Version seiner vortrainierten Granite-Zeitreihenmodelle an, deren erste Versionen Anfang des Jahres veröffentlicht wurden. Diese neuen Modelle wurden mit dreimal so vielen Daten trainiert und liefern eine starke Leistung bei allen drei wichtigen Zeitreihen-Benchmarks und übertreffen die 10-mal größeren Modelle von Google, Alibaba und anderen. Die aktualisierten Modelle bieten auch eine größere Flexibilität bei der Modellierung mit Unterstützung für externe Variablen und rollierende Prognosen.[4]

Granite Guardian 3.0: läutet die nächste Ära der verantwortungsvollen KI ein

Im Rahmen dieser Version führt IBM auch eine neue Familie von Granite Guardian-Modellen ein, die es Anwendungsentwicklern ermöglichen, Sicherheitsleitplanken zu implementieren, indem sie Benutzeraufforderungen und LLM-Antworten auf eine Vielzahl von Risiken überprüfen. Die Granite Guardian 3.0 8B- und 2B-Modelle bieten die umfassendste Palette an Risiko- und Schadenserkennungsfunktionen, die derzeit auf dem Markt erhältlich ist.

Zusätzlich zu den Schadensdimensionen wie soziale Voreingenommenheit, Hass, Toxizität, Obszönität, Gewalt, Gefängnisausbruch und mehr bieten diese Modelle auch eine Reihe einzigartiger RAG-spezifischer Prüfungen, wie z. B. Fundiertheit, Kontextrelevanz und Antwortrelevanz. In umfangreichen Tests mit 19 Sicherheits- und RAG-Benchmarks hat das Granite Guardian 3.0 8B-Modell eine höhere Gesamtgenauigkeit bei der Schadenserkennung als alle drei Generationen der Llama Guard-Modelle von Meta. Auch bei der Erkennung von Halluzinationen liegt es im Durchschnitt gleichauf mit den spezialisierten Halluzinationserkennungsmodellen WeCheck und MiniCheck.[5]

Obwohl die Granite Guardian-Modelle von den entsprechenden Granite-Sprachmodellen abgeleitet sind, können sie zur Implementierung von Leitplanken neben allen offenen oder proprietären KI-Modellen verwendet werden.

Verfügbarkeit der Granite 3.0-Modelle

Die gesamte Suite der Granite 3.0 Modelle und die neuen Zeitreihenmodelle stehen auf HuggingFace unter der uneingeschränkten Apache 2.0 Lizenz zum Download bereit. Die Instruct-Varianten der neuen Granite 3.0 8B- und 2B-Sprachmodelle und die Granite Guardian 3.0 8B- und 2B-Modelle sind ab heute für die kommerzielle Nutzung auf der watsonx-Plattform von IBM verfügbar. Eine Auswahl der Granite 3.0-Modelle wird auch als NVIDIA NIM Microservices und über die Vertex AI Model Garden-Integrationen von Google Cloud mit HuggingFace verfügbar sein.

Um Entwickler_innen eine größere Auswahl und eine einfachere Nutzung zu ermöglichen und lokale Edge-Implementierungen zu unterstützen, ist eine Auswahl der Granite 3.0-Modelle auch auf Ollama und Replicate verfügbar.

Die neueste Generation der Granite-Modelle erweitert den robusten Open-Source-Katalog der leistungsstarken LLMs von IBM. IBM hat mit Ökosystempartnern wie AWS, Docker, Domo, Qualcomm Technologies, Inc. über seinenQualcomm® AI Hub, Salesforce, SAP und anderen zusammengearbeitet, um eine Vielzahl von Granite-Modellen in die Angebote dieser Partner zu integrieren oder Granite-Modelle auf ihren Plattformen verfügbar zu machen, was Unternehmen auf der ganzen Welt eine größere Auswahl bietet.

Von Assistenten zu Agenten: die Zukunft der KI in Unternehmen realisieren

IBM treibt die KI in Unternehmen durch ein breites Spektrum an Technologien voran - von Modellen und Assistenten bis hin zu den Tools, die für die Abstimmung und den Einsatz von KI speziell für die Daten und Anwendungsfälle von Unternehmen erforderlich sind. IBM ebnet auch den Weg für zukünftige KI-Agenten, die sich selbst steuern, reflektieren und komplexe Aufgaben in dynamischen Geschäftsumgebungen ausführen können.

IBM entwickelt sein Portfolio an KI-Assistententechnologien kontinuierlich weiter - von watsonx Orchestrate, mit dem Unternehmen ihre eigenen Assistenten über Low-Code-Tools und Automatisierung erstellen können, bis hin zu einer breiten Palette an vorgefertigten Assistenten für spezifische Aufgaben und Bereiche wie Kundenservice, Personalwesen, Vertrieb und Marketing. Unternehmen auf der ganzen Welt haben watsonx Assistant genutzt, um KI-Assistenten für Aufgaben wie die Beantwortung von Routinefragen von Kunden oder Mitarbeitern, die Modernisierung ihrer Mainframes und älteren IT-Anwendungen, die Unterstützung von Studenten bei der Erkundung potenzieller Karrierewege oder die Bereitstellung digitaler Hypothekenunterstützung für Hauskäufer zu entwickeln.

IBM hat heute auch die nächste Generation von watsonx Code Assistant vorgestellt, die auf Granite-Code-Modellen basiert und allgemeine Unterstützung bei der Codierung in Sprachen wie C, C++, Go, Java und Python sowie erweiterte Funktionen zur Anwendungsmodernisierung für Enterprise-Java-Anwendungen bietet.[6] Die Code-Funktionen von Granite sind jetzt auch über eine Visual Studio Code-Erweiterung, IBM Granite.Code, zugänglich.

IBM plant außerdem die Veröffentlichung neuer Tools, die Entwicklern helfen sollen, KI effizienter über watsonx.ai zu entwickeln, anzupassen und einzusetzen - einschließlich agentenbasierter Frameworks, Integrationen in bestehende Umgebungen und Low-Code-Automatisierungen für gängige Anwendungsfälle wie RAG und Agenten.[7]

IBM konzentriert sich auf die Entwicklung von KI-Agententechnologien, die zu größerer Autonomie, ausgefeilten Schlussfolgerungen und mehrstufigen Problemlösungen fähig sind. Die erste Version des Granite 3.0 8B Modells bietet Unterstützung für wichtige agentenbasierte Fähigkeiten, wie z.B. fortgeschrittenes logisches Denken und eine hochstrukturierte Chat-Vorlage sowie einen Prompting-Stil für die Implementierung von Workflows für die Toolnutzung. IBM plant außerdem die Einführung einer neuen KI-Agenten-Chatfunktion für IBM watsonx Orchestrate, die KI-Assistenten, Skills und Automatisierungen orchestriert, die den Nutzern helfen, die Produktivität ihrer Teams zu steigern.[8] IBM plant, bis 2025 weitere Agentenfunktionen für sein gesamtes Portfolio zu entwickeln, einschließlich vorgefertigter Agenten für bestimmte Domänen und Anwendungsfälle.

Erweiterte KI-gestützte Bereitstellungsplattform, um IBM Berater_innen mit KI zu unterstützen

IBM kündigt außerdem eine bedeutende Erweiterung seiner KI-gestützten LieferplattformIBM Consulting Advantage an. Die Multi-Modell-Plattform enthält KI-Agenten, -Anwendungen und -Methoden wie wiederholbare Frameworks, mit denen 160.000 IBM-Berater_innen einen besseren und schnelleren Kundennutzen zu geringeren Kosten liefern können.

Im Rahmen der Erweiterung werden die Sprachmodelle von Granite 3.0 zum Standardmodell in Consulting Advantage. Durch die Nutzung der Leistung und Effizienz von Granite wird IBM Consulting in der Lage sein, den Return-on-Investment für die generativen KI-Projekte der IBM Kunden zu maximieren.

Ein weiterer wichtiger Bestandteil der Erweiterung ist die Einführung von IBM Consulting Advantage for Cloud Transformation and Management und IBM Consulting Advantage for Business Operations. Beide umfassen domänenspezifische KI-Agenten, -

Anwendungen und -Methoden, die mit den Best Practices von IBM angereichert sind, so dass IBM-Berater_innen Kund_innen dabei helfen können, Cloud- und KI-Transformationen bei Aufgaben wie Code-Modernisierung und Quality Engineering zu beschleunigen oder Abläufe in Bereichen wie Finanzen, HR und Beschaffung zu transformieren und durchzuführen.

Wenn Sie mehr über Granite und die IBM-Strategie für KI für Unternehmen erfahren möchten, besuchen Sie https://www.ibm.com/de-de/granite.

[1] Die Kostenberechnungen basieren auf den API-Kosten pro Million Token von IBM watsonx für offene Modelle und openAI für GPT4-Modelle (unter der Annahme einer Mischung von 80 % Inout, 20 % Output) für Kunden-Proofs-of-Concept.

[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] Evaluation results published in Granite Guardian GitHub Repo

[6] Geplante Verfügbarkeit für Q4 2024

[7] Geplante Verfügbarkeit für Q4 2024

[8] Geplante Verfügbarkeit für Q1 2025

Medienkontakte

Annette Hodapp
Annette_Hodapp@de.ibm.com

Amy Angelini
alangeli@us.ibm.com

QUELLE IBM

IBM-Granite

https://de.newsroom.ibm.com/IBM-stellt-Granite-3-0-vor-Hochleistungsfahige,-vertrauenswurdige-KI-Modelle-fur-Unternehmen