IBM Hong Kong newsroom

IBM 推出Granite 3.2 全新企業級多模態及推理的AI模型

- Granite 3.2是小型的 AI 模型 通過對開發人員友好的授權條款 提供推理、視覺和護欄功能
- 更新後的 Granite 時間序列模型可提供長期預測 參數少於10M 適用於金融與經濟趨勢分析、供應鏈需求預測 以及零售業的季節性庫存規劃等用例

香港2025年3月4日 /美通社/ -- IBM於2月底推出其 Granite 大型語言模型家族的最新一代 品 Granite 3.2 持續推動小型、高效、企業專用的 AI 為實際應用創造效益。

所有 Granite 3.2 模型均採用 鬆的 Apache 2.0 開源授權條款 可在 Hugging Face 下載。部分模型現已在 IBM watsonx.ai、Ollama、Replicate 和 LM Studio 上提供 預計不久將支援 RHEL AI 1.5 為企業與開源社區注入更強大的 AI 能力。

主要亮點

- **全新視覺語言模型** 專為理解文件任務而設計 在關鍵企業基準測試 DocVQA、ChartQA、AI2D 和 OCRBench[1] 中 表現可 美甚至超越更大規模的模型 如 Llama 3.2 11B 和 Pixtral 12B 。除了強大的訓練數據外 IBM 也利用其開源 Docling 工具包處理8千5百萬 PDF 文件 並生成2千6百萬個合成問答配對 提升視覺語言模型在處理大量文件工作流時的能力。
- **強推理功能** Granite 3.2的2B與8B模型加入了「思維鏈」 Chain of Thought CoT 推理機制 且使用者可以開 或關閉推理功能 以優化效率。通過這項能力 8B 模型在 ArenaHard 和 Alpaca Eval 等指令遵循基準測試中的表現[2] 比前一代優異比例達到兩位數 且不影響其他領域的安全性或性能。此外 通過創新的推理擴展方法 Granite 3.2 8B 模型可以調整至接近 Claude 3.5 Sonnet 或 GPT-4o 在數學推理基準 如 AIME2024 和 MATH500[3] 上的表現。
- **Granite Guardian 安全模型更輕巧** 在保持 Granite 3.1 Guardian 模型性能的同時 模型尺寸減少三成。此外 Granite 3.2 系列還引入了語言化信心評估 Verbalized Confidence 新功能 可提供更精細的風險評估 助安全監測系統識別不確定性。

IBM 持續推動企業專用的小型 AI 模型策略 並已在測試中展現高效能。例如 Granite 3.1 8B 模型在 Salesforce 大型語言模型CRM 基準測試中獲得高分 顯示其在實際應用中的準確度和可靠性。

IBM Granite 模型家族擁有廣大的合作夥伴生態體系 許多領先的軟件公司已將Granite模型嵌入其技術。Granite 3.2 是 IBM 在推動企業專用小型 AI 方面的重要進展 體現了 IBM 致力於提供小型、高效、實用 AI 的 品策略。

CrushBank 首席技術官 David Tan 表示 「在 CrushBank 我們親眼目睹了 IBM 開放、高效的人工智能模型如何為企業人工智能帶來真正的價 --在性能、成本效益和可擴展性之間實現適當的平衡。Granite 3.2 通過新的推理功能更進一 我們很高興能在構建新的代理 智能體 解決方案時探索這些功能。」

Granite 3.2 是 IBM 品組合和戰略發展的重要一 旨在為企業提供小型實用的 AI。雖然思維鏈在推理任務中表現強大 但它需要大量計算資源 並非所有任務都必須 用。因此 IBM 在 Granite 3.2 模型中加入了程式化開關功能 使用者可以根據需求開 或關閉推理模式 模型可在不 用推理的情況下運行較簡單的任務 以降低不必要的計算成本。

此外 其他推理技術 例如推理擴展 Inference Scaling 已顯示 Granite 3.2 8B 模型能 在標準數學推理基準測試中 美甚至超越更大模型的性能。持續發展這項推理技術也是 IBM 研究團隊的重點方向[4] 以進一 提升 AI 的效能與應用範圍。

除了 Granite 3.2 的指令、視覺和防護模型之外 IBM 也推出了新一代 TinyTimeMixers TTM 時間序列模型 這些模型的參數少於1千

萬 具備長期預測能力 可進行長達兩年的長期預測。這些模型為長期趨勢分析提供強大工具 適用於金融與經濟趨勢分析、供應鏈需求預測 以及零售業的季節性庫存規劃。

IBM AI 研究副總裁 Sriram Raghavan 表示 「AI 的下一個時代將聚焦效率、整合與實際應用的影響力 — 企業應該能 在不過度消耗計算資源的情境下 取得強大的 AI 效益。IBM 最新的 Granite 模型發展專注於開放式解決方案 逐 推動 AI 的普及 使其更具成本效益 為現代企業創造更大價 。」

欲了解Granite 3.2 的技術細節 請參 相關技術文章。

關於 IBM

IBM 是全球領先的混合雲與人工智能、以及企業服務提供商 為全球175個國家和地區的客 服務 助企業把握其數據洞察、簡化業務流程、降本 效 獲得行業競爭優勢。 IBM 混合雲平台和紅帽OpenShift 為全球超過4,000家政府和企業機構的關鍵性基礎設施提供有力支 例如來自金融服務、電訊和醫療健康等行業的客 助他們快速、高效、安全地實現數碼轉型。 IBM 在人工智能、量子運算、特定行業的雲解決方案以及企業服務等方面的突破性創新 使其可以為客 提供開放和靈活的選擇。 IBM 對信任、透明、責任、包容和服務的 久彌新的承諾 是我們業務發展的基石。 詢更多資料 請瀏覽 www.ibm.com/

傳媒 詢
郭韜 gguotao@cn.ibm.com

[1] 視覺模型 Vision Model 的基準測試結果可在 IBM技術文章《IBM Granite 3.2 推理、視覺、預測與更多應用》 2025 年 2 月 26 日發布 中 。

[2] 指令模型 Instruct Model 的基準測試結果可在 IBM 技術文章 《IBM Granite 3.2 推理、視覺、預測與更多應用》 2025 年 2 月 26 日發布 中 。

[3] 推理擴展 Inference Scaling 的基準測試結果可在 IBM技術研究部落格 《Granite 3.2 中的推理 利用推理擴展技術》 2025 年 2 月 26 日發布 中 。

[4] 推理擴展技術在 Granite 3.2 中的應用 IBM 技術研究部落格 2025 年 2 月 26 日發布

SOURCE IBM Hong Kong

Additional assets available online: Photos

https://hongkong.newsroom.ibm.com/2025-03-04-IBM-Granite-3-2-AI