NEWS RELEASE

ServiceNow, Hugging Face, and NVIDIA launch New Open-Access LLMs Help Developers Tap Generative AI to Build Enterprise Applications

2024-02-28

StarCoder2 - Created With the BigCode community and Trained on 600+ Programming Languages, Advances Code Generation, Transparency, Governance, and Innovation

SANTA CLARA, Calif., Feb. 28, 2024 — **ServiceNow** (NYSE: NOW), **Hugging Face**, and **NVIDIA**, today announced the release of **StarCoder2**, a family of open-access large language models (LLMs) for code generation that sets new standards for performance, transparency, and cost-effectiveness.

StarCoder2 was developed by the **BigCode** community, stewarded by **ServiceNow**, the leading digital workflow company making the world work better for everyone, and **Hugging Face**, the most-used open-source platform where the machine learning community collaborates on models, datasets and applications.

Trained on 619 programming languages, StarCoder2 can be further trained and embedded in enterprise applications to perform specialized tasks such as application source code generation, workflow generation, text summarization, and more. Developers can use its code completion, advanced code summarization, code snippets retrieval, and other capabilities to accelerate innovation and improve productivity.

StarCoder2 offers three model sizes: a 3 billion-parameter model trained by ServiceNow, a 7 billion-parameter model trained by Hugging Face, and a 15 billion-parameter model built by NVIDIA with NVIDIA NeMo and trained on NVIDIA accelerated infrastructure. The smaller variants provide powerful performance while saving on compute costs, as fewer parameters require less computing during inference. In fact, the new StarCoder2 3 billion-parameter

model also matches the performance of the original StarCoder 15 billion-parameter model.

“StarCoder2 stands as a testament to the combined power of open scientific collaboration and responsible AI practices with an ethical data supply chain,” emphasized Harm de Vries, lead of ServiceNow's StarCoder2 development team, and co-lead of BigCode. "The state-of-the-art open-access model improves on prior generative AI performance to increase developer productivity and provides developers equal access to the benefits of code generation AI, which in turn enables organizations of any size to more easily meet their full business potential.”

"The joint efforts led by Hugging Face, ServiceNow and NVIDIA enable the release of powerful base models that empower the community to build a wide range of applications more efficiently with full data and training transparency," said Leandro von Werra, machine learning engineer at Hugging Face and co-lead of BigCode. “StarCoder2 is a testament to the potential of open-source and open science as we work toward democratizing responsible AI."

"Since every software ecosystem has a proprietary programming language, code LLMs can drive breakthroughs in efficiency and innovation in every industry,” said Jonathan Cohen, vice president of applied research at NVIDIA. “NVIDIA’s collaboration with ServiceNow and Hugging Face introduces secure, responsibly developed models, and supports broader access to accountable generative AI that we hope will benefit the global community.”

Fine-Tuning Advances Capabilities with Business-Specific Data

StarCoder2 models share a state-of-the-art architecture and carefully curated data sources from BigCode that prioritize transparency and **open governance** to enable responsible innovation at scale.

The foundation of StarCoder2 is a new code dataset called **The Stack v2** which is more than 7x larger than **The Stack v1**. In addition to the advanced data set, new training techniques help the model understand low-resource programming languages (such as COBOL), mathematics, and program source code discussions.

StarCoder2 advances the potential of future AI-driven coding applications, including text-to-code and text-to-workflow capabilities. With broader, deeper programming training, it provides repository context, enabling accurate, context-aware predictions. These advancements serve seasoned software engineers and citizen developers alike, accelerating business value and digital transformation.

Users can fine-tune the open-access models with industry or organization-specific data using open-source tools such as NVIDIA NeMo or Hugging Face TRL.

Organizations have already fine-tuned the foundational StarCoder model to create specialized task-specific

capabilities for their businesses.

ServiceNow's text-to-code Now LLM was purpose-built on a specialized version of the 15 billion-parameter StarCoder LLM, fine-tuned and trained for ServiceNow workflow patterns, use-cases, and processes. Hugging Face also used the model to create its StarChat assistant.

BigCode Fosters Open Scientific Collaboration in AI

BigCode represents an open scientific collaboration jointly led by Hugging Face and ServiceNow. Its mission centers on the responsible development of LLMs for code.

The BigCode community actively participated in the technical aspects of the StarCoder2 project through working groups and task forces, leveraging ServiceNow's Fast LLM framework to train the 3 billion-parameter model, Hugging Face's nanotron framework for the 7 billion-parameter model, and the end-to-end NVIDIA NeMo cloud-native framework and **NVIDIA TensorRT-LLM** software to train and optimize the 15 billion-parameter model.

Fostering responsible innovation is at the core of BigCode's purpose, demonstrated through its open governance, transparent supply chain, use of open-source software, and the ability for developers to opt data out for training. StarCoder2 was built using responsibly sourced data under license from the digital commons of **Software Heritage**, hosted by **Inria**.

"StarCoder2 is the first code generation AI model developed using the Software Heritage source code archive and built to align with our policies for responsible development of models for code," stated Roberto Di Cosmo, Director at Software Heritage. "The collaboration of ServiceNow, Hugging Face, and NVIDIA exemplifies a shared commitment to ethical AI development, advancing technology for the greater good."

StarCoder2, as with its predecessor, will be made available under the BigCode Open RAIL-M license, allowing royalty-free access and use. Furthermore, the supporting code for the models resides on the BigCode project's GitHub page.

All StarCoder2 models will also be **available for download** from Hugging Face and the StarCoder2 15B model is available on **NVIDIA AI Foundation models** for developers to experiment with directly from their browser, or through an API endpoint.

For more information on StarCoder2, visit **https://huggingface.co/bigcode**.

About ServiceNow

ServiceNow (NYSE: NOW) makes the world work better for everyone. Our cloud based platform and solutions help digitize and unify organizations so that they can find smarter, faster, better ways to make work flow. So employees and customers can be more connected, more innovative, and more agile. And we can all create the future we imagine. The world works with ServiceNow™. For more information, visit: **www.servicenow.com**.

About Hugging Face

Hugging Face is the leading open source and community driven AI platform, providing tools that enable users to build, explore, deploy and train machine learning models and datasets. For more information, visit: **www.huggingface.co**

About NVIDIA

Since its founding in 1993, **NVIDIA** (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing company with data center scale offerings that are reshaping the industry. More information at **https://nvidianews.nvidia.com/**.

NVIDIA Forward-Looking Statements

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, and performance of NVIDIA's products, services, and technologies, including NVIDIA NeMo and NVIDIA TensorRT-LLM; the benefits and impact of NVIDIA's collaboration with ServiceNow and Hugging Face, and the features and availability of its services and offerings are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

ServiceNow Forward-Looking Statements

This press release contains “forward-looking statements” about the expectations, beliefs, plans, and intentions relating to releasing a new open-access LLM for code generation. Such statements include statements regarding future product capabilities and offerings and expected benefits to ServiceNow. Forward-looking statements are subject to known and unknown risks and uncertainties and are based on potentially inaccurate assumptions that could cause actual results to differ materially from those expected or implied by the forward-looking statements. If any such risks or uncertainties materialize or if any of the assumptions prove incorrect, ServiceNow’s results could differ materially from the results expressed or implied by the forward-looking statements made. ServiceNow undertakes no obligation, and does not intend, to update the forward-looking statements. Factors that may cause actual results to differ materially from those in any forward-looking statements include: (i) delays and unexpected difficulties and expenses in executing the partnership or delivering the product capabilities and offerings, (ii) changes in the regulatory landscape related to AI and (iii) uncertainty as to whether sales will justify the investments in the product capabilities and offerings. Further information on factors that could affect ServiceNow’s financial and other results is included in the filings ServiceNow makes with the Securities and Exchange Commission from time to time.

Media Contacts:

ServiceNow
Jacqueline Velasco
press@servicenow.com

NVIDIA
Anna Pari Kiachian
akiachian@nvidia.com

Hugging Face
Brigitte Tousignant
brigitte@huggingface.co