Mind the Gap: Large Language Models Get Smarter With Enterprise Data

NVIDIA NeMo service to help enterprises combine LLMs with their proprietary data to improve chatbots, customer service and more.

Author: Erik Pounds

Large language models available today are incredibly knowledgeable, but act like time capsules — the information they capture is limited to the data available when they were first trained. If trained a year ago, for example, an LLM powering an enterprise's AI chatbot won't know about the latest products and services at the business.

With the NVIDIA NeMo service, part of the newly announced NVIDIA AI Foundations family of cloud services, enterprises can close the gap by augmenting their LLMs with proprietary data, enabling them to frequently update a model's knowledge base without having to further train it — or start from scratch.

This new functionality in the NeMo service enables large language models to retrieve accurate information from proprietary data sources and generate conversational, human-like answers to user queries. With this capability, enterprises can use NeMo to customize large language models with regularly updated, domain-specific knowledge for their applications.

This can help enterprises keep up with a constantly changing landscape across inventory, services and more, unlocking capabilities such as highly accurate AI chatbots, enterprise search engines and market intelligence tools.

NeMo includes the ability to cite sources for the language model's responses, increasing user trust in the output. Developers using NeMo can also set up guardrails to define the AI's area of expertise, providing better control over the generated responses.

Quantiphi — an AI-first digital engineering solutions and platforms company and one of NVIDIA's service delivery partners — is working with NeMo to build a modular generative AI solution called baioniq that will help enterprises build customized LLMs to boost worker productivity. Its developer teams are creating tools that let users search up-to-date information across unstructured text, images and tables in seconds.

Analysts estimate that around two-thirds of enterprise data is untapped. This so-called dark data is unused partly because it's difficult to glean meaningful insights from vast troves of information. Now, with NeMo, businesses can retrieve insights from this data using natural language queries.

NeMo can help enterprises build models that can learn from and react to an evolving knowledge base — independent of the dataset that the model was originally trained on. Rather than needing to retrain an LLM to account for new information, NeMo can tap enterprise data sources for up-to-date details. Additional information can be added to expand the model's knowledge base without modifying its core capabilities of language processing and text generation.

Enterprises can also use NeMo to build guardrails so that generative AI applications don't provide opinions on topics outside their defined area of expertise.

By customizing an LLM with business data, enterprises can make their AI applications agile and responsive to new developments.

Chatbots: Many enterprises already use AI chatbots to power basic customer interactions on their websites. With NeMo, companies could build virtual subject-matter experts specific to their domains.

Customer service: Companies could update NeMo models with details about their latest products, helping live service representatives more easily answer customer questions with precise, up-to-date information.

Enterprise search: Businesses have a wealth of knowledge across the organization, including technical documentation, company policies and IT support articles. Employees could query a NeMo-powered internal search engine to retrieve information faster and more easily.

Market intelligence: The financial industry collects insights about global markets, public companies and economic trends. By connecting an LLM to a regularly updated database, investors and other experts could quickly identify useful details from a large set of information, such as regulatory documents, recordings of earnings calls or financial statements.

Enterprises interested in adding generative AI capabilities to their applications can apply for early access to the NeMo service.

Watch NVIDIA founder and CEO Jensen Huang discuss NVIDIA AI Foundations in the keynote address at NVIDIA GTC , running online through Thursday, March 23:

Original URL: https://blogs.nvidia.com/blog/2023/03/21/nemo-large-language-models-enterprise-data/