Case Study

Software services
AI integration

Prediction Guard De-Risks LLM Applications at Scale

Intel Developer Cloud provides AI startup with resilient computing resources, ensuring peak performance and consistency in cloud operations for generative AI applications.

Large language models can drive AI-enabled innovation, but many businesses are finding they lack the resources and expertise to get the results they desire.

Large language models (LLMs) hold tremendous potential for helping companies realize operational efficiencies and build compelling AI-driven tools. Prediction Guard, a startup company operating at the forefront of AI integration, wants to help organizations unleash that potential faster.

While LLMs are proving their utility across a wide range of generative AI (GenAI) applications, they can bring risks when it comes to reliability, variability and security. That makes them challenging to deploy for many use cases, especially ones in which human safety and enterprise compliance are top considerations.

Prediction Guard is working to reduce those risks, while leveraging Intel® Developer Cloud and Intel® Gaudi® 2 processors to run more efficiently and effectively as it helps organizations integrate LLMs into their AI-enabled business applications.

At-a-glance

Prediction Guard's API platform enables enterprises to harness the full potential of large language models while mitigating security and trust issues such as hallucinations, harmful outputs and prompt injections.

Intel® Developer Cloud is helping the company increase throughput while providing scalability, availability, security and cost savings.

The Intel® Liftoff program enables the company to maximize operational capability, value and efficiency with Intel Developer Cloud.

Challenge

Founded at the beginning of 2023, Prediction Guard has rapidly built a client portfolio spanning a range of industries. Many of the company's clients recognize that LLMs can be game-changers. Those organizations are looking to leverage LLMs in GenAI applications such as information extraction, customer service and analytics, marketing campaign planning, financial reporting and "co-pilot" solutions.

One prime use case supports medical providers by automatically drafting medical forms based on information extracted from physician dictations (a scenario in which private data and LLM outputs need to be handled securely and reliably). Other use cases integrate private customer data to generate answers to supply chain questions or to respond to retail customers based on purchase behavior, inventory and order information.

"The use cases are almost limitless," said Daniel Whitenack, Prediction Guard's founder. "Businesses see opportunities everywhere they look. They see how they can use GenAI tools such as large language models to unlock the full power of their data — to create new insights and experiences that can take business outcomes to the next level. But they also recognize that there are still huge obstacles to overcome."

One key challenge, according to Whitenack, is that the output of LLMs can vary, making them unreliable. They are sometimes prone to "hallucinations," with results that are factually inaccurate or toxic. And they are also vulnerable to an emerging type of security threat known as prompt injections, in which an attacker uses a malicious input to elicit an unintended response or data breach from the model.

"All of these challenges can make it tricky for organizations to gain traction with LLMs. They want to avoid liabilities, and they don't want to risk the loss of sensitive data such as personally identifiable information. Many of them also lack the data talent and other resources needed to address these issues," Whitenack said. "Our solution helps remove these barriers to adoption, but it requires significant and reliable computing power — something we have been able to access through Intel Developer Cloud."

Solution

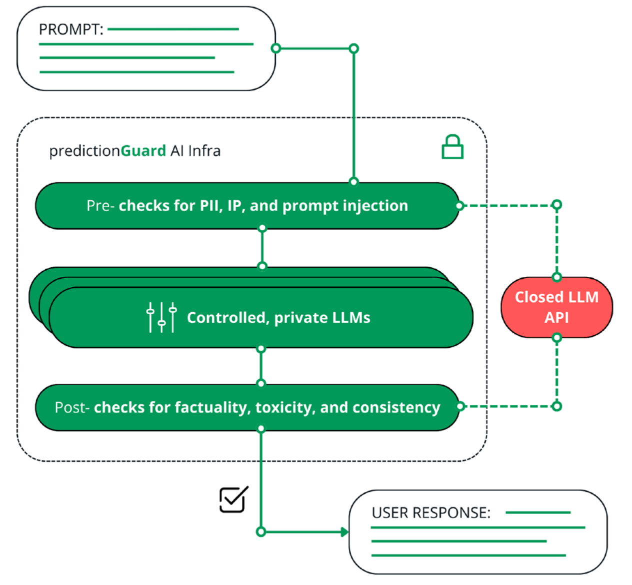
Prediction Guard's platform provides its customers with access to multiple LLMs while integrating additional capabilities and technologies that deal with harmful inputs (like prompt injections) and harmful outputs (like inaccuracies or toxicity) — all hosted in a secure, private environment on Intel Developer Cloud, running Intel Gaudi 2 deep-learning processors. The company is also using the Hugging Face Optimum Habana library (a collaboration between Intel and Hugging Face) in Intel Developer Cloud to optimize the models it runs on the processors.

Based on an open software foundation with oneAPI, Intel Developer Cloud allows users to learn, test and run applications on a cluster of the latest Intel hardware and software. It gives developers access to Intel technologies for pre-launch development and testing, as well as supplying a full stack solution for building and deploying AI applications at scale. Intel Developer Cloud also offers developers hardware choice and freedom from proprietary programming models, which supports accelerated computing, code reuse and portability.

The Intel Gaudi 2 processors provided in Intel Developer Cloud are helping Prediction Guard address some of the big needs that come with a growing business and with client expectations for rapid, reliable GenAI outcomes. Designed as an AI accelerator, the Intel Gaudi 2 processor features 7 nanometer process technology, heterogeneous compute, 24 tensor processor cores, and dual-matrix multiplication engines. It also includes 24 100 gigabit Ethernet integrated on chip, 96 GB HBM2E memory on board, 48 MB SRAM and integrated media control.

Before moving to Intel Gaudi 2 processors on Intel Developer Cloud, Prediction Guard tried processors from another provider. The new Intel-based environment has provided a significant boost for Prediction Guard and its clients in multiple areas, according to Whitenack.

"For certain models, following our move to Intel Gaudi 2, we have seen our costs decrease while throughput has increased by 2x — which means we can help customers generate outputs faster and more cost-effectively.[1] It's also now easier for us to get access to the processing power we need in the cloud — something that was often scarce with our previous provider," he said. "That greater availability also allows us to scale more readily to meet the growing needs of our clients and our own business."

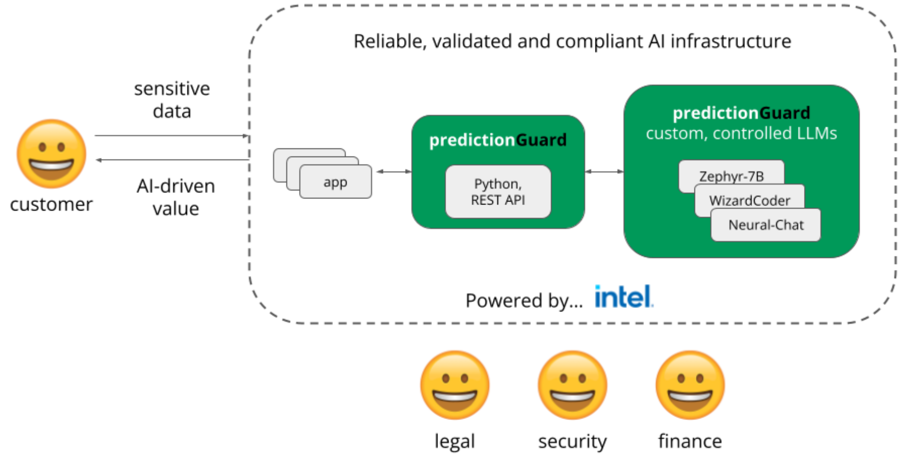

PROMPT:
predictionGuard AI Infra
Pre- checks for PII, IP, and prompt injection
Controlled, private LLMs
Post- checks for factuality, toxicity, and consistency
Closed LLM API
USER RESPONSE:

Playing a vital role between prompt and user response

Source: Prediction Guard

For Prediction Guard clients with especially sensitive data considerations, Whitenack said Intel Developer Cloud allows the company to create client-specific deployments that protect customer data while maintaining the easy-to-use Prediction Guard API.

One other advantage of Intel Developer Cloud stems from Prediction Guard's involvement in Intel® Liftoff for Startups, a program that provides startups with access to the computational power they need and fosters collaboration among startups to help them refine and enhance their offerings.



Leveraging Intel® technology to help provide AI-driven value for customers

Source: Prediction Guard

[1] Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

"Our participation in Intel Liftoff was critical for getting us where we are today. It allowed us to explore the art of the possible with Intel Gaudi 2 processors and demonstrate early on that our system could run effectively on Intel Developer Cloud at the speed and scale we needed," Whitenack said.

As Prediction Guard has moved fully into production with Intel Gaudi 2 processors on Intel Developer Cloud, the company continues to see benefits from Intel Liftoff. Whitenack said the program provides rapid access to Intel experts who help answer questions and address issues as Prediction Guard continues to innovate using Intel technology.

"That's one of the greatest things about this relationship. The compute and the community are here," Whitenack said. "And so are the results. Intel Developer Cloud provides a full-fledged AI cloud environment that can support our business as we continue to innovate and grow."

Result

Prediction Guard is realizing greater speed, security, scalability and reliability, as well as reduced costs, using Intel Gaudi 2 AI accelerator processors on Intel Developer Cloud. The technology is providing a 2x throughput increase for some large language models, for example, allowing the company to support clients using LLMs for a variety of innovative GenAI applications.[2]

And Intel Developer Cloud is having a measurable impact on Prediction Guard's bottom line. Because the company has ported critical model hosting onto Intel Gaudi 2 machines in Intel Developer Cloud, it has a better view of its fixed costs. That means the company can conduct more accurate cost planning and flexibly adjust the models it runs.

Solution ingredients

Intel® Developer Cloud

Intel® Gaudi® 2 processors

Optimum Habana

Intel® Extension for PyTorch*

Intel® Extension for Transformers*

Where to get more information

Prediction Guard

Intel® Liftoff

Intel® Developer Cloud

Intel® Gaudi® 2 AI accelerator

oneAPI

Notices and disclaimers
1, 2. As reported by Prediction Guard as of January 31, 2024.

Performance varies by use, configuration, and other factors. Learn more at intel.com/PerformanceIndex.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.
Not all features are available on all SKUs.
Not all features are supported in every operating system.
Intel may change availability of products and support at any time without notice. All product plans are subject to change without notice.
Your costs and results may vary.
Intel® technologies may require enabled hardware, software, or service activation.