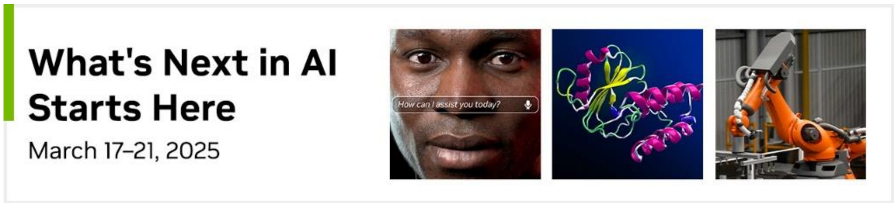

AI Engineers and Developers - Discover the Future of Generative AI at NVIDIA GTC 2025
Join us in San Jose from March 17–21, 2025, for NVIDIA GTC, the premier event where technology meets transformation. Dive into the world of Generative AI and explore its groundbreaking impact on the Federal government's most complex challenges.

Deep Dive into Generative AI: Explore the latest advancements in Generative AI technologies, including foundational models and creative AI applications. Learn how these innovations drive significant improvements across government operations. Experience over 500 sessions led by industry pioneers, more than 300 exhibits showcasing next-gen technology, and exclusive hands-on training opportunities. GTC is your gateway to networking with peers and experts in an immersive AI environment.

Secure your spot today and be at the forefront of AI innovation!
Use my employee referral for 25% off early bird discount: : www.nvidia.com/gtc/?ncid=GTC-NVD4DXLNZ

Engage with Thought Leaders and the Generative AI Ecosystem All in One Place:
Unlock the transformative potential of Generative AI at NVIDIA GTC, where developers like you converge to:

- **Experience Cutting-Edge Generative AI Applications**: Dive deep through technical sessions and hands-on training to learn how Generative AI is revolutionizing industries across the board—from the public sector to healthcare to finance. See how these innovations can be applied to enhance your projects and workflows.
- **Network with Industry Leaders**: Connect with thought leaders, innovators, and potential collaborators. Don't miss the opportunity to meet NVIDIA's CEO, Jensen Huang. He's often available on the expo floor and at special sessions, sharing insights that can redefine your understanding of Generative AI's possibilities.
- **Enhance Your Professional Skills**: Gain invaluable insights from real-world implementations of Generative AI by industry pioneers. These learnings can catapult your career to the next level, preparing you for the future demands of tech roles.

GTC Keynote: March 18th at 10:00AM Pacific Time

Don't miss this keynote from NVIDIA founder and CEO Jensen Huang. He'll share how NVIDIA's accelerated computing platform is driving the next wave in AI, digital twins, cloud technologies, and sustainable computing.

I've created a curated list of sessions below. There are plenty of other sessions in the catalog, but I wanted to give you a taste of what is at GTC 2025. See the full Generative AI track here.

Session and ID	Description
Transform an Enterprise Data Platform With Generative AI and RAG [S72205]	Trillions of PDF files are generated every year, each file likely consisting of multiple pages filled with various content types, including text, images, charts, and tables. Learn how generative AI and retrieval-augmented generation (RAG) is enabling enterprises to extract massive volumes of data to quickly empower employees with valuable expertise.
Pioneering the Future of Data Platforms for the Era of Generative AI [S71650]	In the era of generative AI, retrieval augmented generation (RAG) is emerging as a vital tool, enabling enterprises to capitalize on their data. As data evolves and volumes grow exponentially, traditional storage solutions can no longer meet the rigorous demands of enterprise AI systems. GenAI inferencing has become a rack-scale computing workload, necessitating robust, enterprise-grade infrastructure. Discover how NVIDIA is collaborating with storage trailblazers to create a new class of accelerated data platforms — optimized for GenAI workloads and real-time RAG pipelines. Powering these next-generation systems are NVIDIA GPUs, high-performance NVIDIA networking, NVIDIA NIMs, and an AI-native data ingestion stack, delivering efficient pre-processing, storage, and retrieval at unprecedented speeds. Explore the innovations that are redefining enterprise data platforms and shaping the future of GenAI.
How to Create Your Organization's AI Flywheel [S71784]	The new wave of generative AI is here, powered by the availability of advanced open-source foundation models, as well as advancements in agentic AI that are improving efficiency and autonomy of AI workflows. Enterprises can build and operationalize custom AI applications — creating data-driven AI flywheels — using NIM Agent Blueprints along with NVIDIA NIM microservices and NVIDIA NeMo framework, all part of the NVIDIA AI Enterprise Platform. Learn how NVIDIA NIM Agent Blueprints enable you to quickly build and deploy customizable Gen AI applications.
How to Build an Agentic AI System Using the Best Tools and Frameworks [S73739]	AI agents are the new digital workforce, working for and with us. They can reason about a mission, create a plan, and retrieve data or use tools to generate a quality response. Data is the fuel for AI agents, but the magnitude and scale of enterprise data often make it too expensive and time-consuming to leverage effectively. For enterprises to thrive in the AI era, they must find a way to make use of all of their data. Join this session to learn about tools and frameworks to more easily build agentic AI systems that connect AI agents to a library of reusable tools that unlock your data, and drive efficiency gains across the organization.
How to Build Multimodal Agentic AI Retrieval Systems [S72208]	Join NVIDIA technical product architects for an in-depth tutorial demonstrating how to build agentic AI pipelines that integrate diverse data types—including text, images, audio, and video—into enterprise AI applications. This session will cover the full journey, from designing production-ready retrieval systems to effectively using them in Enterprise use cases.

For those interested in full-day workshops, hands-on training labs, and certifications, I recommend the following Deep Learning Institute sessions on Generative AI that will be given on Sunday, March 16 or Monday, March 17:

DLI Session and ID	Description
Deploying RAG Pipelines for Production at Scale [DLIW73634] *Full-Day Workshop*	Retrieval-Augmented Generation (RAG) pipelines are revolutionizing enterprise operations. However, most existing tutorials stop at proof-of-concept implementations that falter when scaling. This workshop aims to bridge that gap, focusing on building scalable, production-ready RAG pipelines powered by NVIDIA NIM microservices and Kubernetes. Participants will gain hands-on experience deploying, monitoring, and scaling RAG pipelines with the NIM Operator and learn best practices for infrastructure optimization, performance monitoring, and handling high traffic volumes.

Building AI Agents With Multimodal Models [DLIW73633] *Full-Day Workshop*	Just like how humans have multiple senses to perceive the world around them, more and more computer sensors are being developed to capture a wide variety of data. In the health industry, computed tomography (CT) scans provide a 3D representation to detect potentially dangerous abnormalities. In the robotics industry, lidars help robots see depth and navigate their complex environments. In this course, learners will develop neural network agents that can reason using many different data types by exploring multiple fusion techniques.
Efficient Large Language Model Customization [DLIW73630] *Full-Day Workshop*	Enterprises need to execute language-related tasks daily, such as text classification, content generation, sentiment analysis, and customer chat support. And they need to do so in the most cost-effective way. Large language models can automate these tasks, and efficient LLM customization techniques can increase a model's capabilities and reduce the size of models required for use in enterprise applications. In this course, you'll go beyond prompt-engineering LLMs and learn techniques to efficiently customize pretrained LLMs for your specific use cases. We'll cover how to do this without engaging in the computationally intensive and expensive process of pre-training your own model or fine-tuning a model's internal weights. Using NVIDIA NIM microservices, NeMo Curator, and NeMo Framework, you'll learn various parameter-efficient fine-tuning methods to customize LLM behavior for your organization.
Blueprints for Success: Navigating NIM Agent Workflows for Real-World Multimodal Retrieval [DLIT71592] *Training Lab*	In this comprehensive course, you'll delve into the real-world applications of multimodal retrieval-augmented generation (RAG). Through a series of hands-on modules, learn how to leverage NIM agent blueprints as a one-stop solution for your AI projects. Much of the course is a deep dive into a real-life multimodal RAG application at scale, utilizing NVIDIA NIMs. You'll gain the knowledge and skills needed to implement and optimize multimodal data ingestion, embedding, and retrieval. You'll also learn how NVIDIA Blueprints can guide your AI solution from proof of concept (POC) to production using self-hosted microservices. This holistic approach ensures that you are equipped to handle every stage of the AI development lifecycle. In the end, you'll have a solid understanding of NIM agent workflows and be able to apply them effectively in real-world scenarios, taking your AI projects to the next level.
Build Visual AI Agents With RAG Using NVIDIA Morpheus, RIVA, and Metropolis [DLIT72055] *Training Lab*	Discover how to utilize NVIDIA Visual Insights Agent (VIA) microservices to create your own visual AI agents. You'll construct an agentic retrieval-augmented generation (RAG) pipeline, employing NVIDIA Morpheus SDK and NVIDIA Riva Speech Services NIMs in conjunction with VIA Microservice. We'll show you a practical example aimed at performing agentic RAG on egocentric (first-person) video feeds for dynamic, open-world QA.
Evaluating RAG and Semantic Search Systems [DLIT71593] *Training Lab*	As adoption of LLMs and retrieval-augmented generation (RAG) becomes more prevalent in enterprise, the demand for robust evaluation arises, ensuring these new tools meet the high regulatory standards of various industries. Gain an in-depth understanding of the unique requirements as well as practical tools to robustly evaluate such systems addressing the complex nature of the underlying data. We'll introduce evaluation techniques addressing domain-specific language and knowledge, specific evaluation metrics, and independent assessment of the retrieval and generation steps, all while taking temporal information into account.