Feb 2, 2024

Dear National Institute of Standards and Technology (NIST) Representative:

Thank you for the opportunity to provide comments on behalf of NVIDIA Corporation pursuant to *Request for Information #NIST–2023–0009 Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence*. We appreciate the Department of Commerce's work in promoting innovation and equitable standards that foster a competitive industrial landscape. NVIDIA is committed to safe and trustworthy AI, in line with the White House Voluntary Commitments and other global AI Safety initiatives and are committed to helping to drive standards around the development and deployment of safe and trustworthy AI. NVIDIA believes AI should respect privacy and data protection regulations, operate in a secure and safe way, function in a transparent and accountable manner, and avoid unwanted biases and discrimination. More information about NVIDIA's Guiding Principles for Trustworthy AI can be found here.

Below are NVIDIA's responses to the Request for Information.

1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

Content authentication, provenance tracking, and synthetic content labeling and detection:

Several industry-driven initiatives focus on content authenticity. NVIDIA partners with Adobe in the Content Authenticity Initiative (CAI) to drive metadata tagging of content to identify creation and modification for audit purposes and collaborates with MLCommons to support community development of AI safety tests and benchmarks. NVIDIA also partners with creative platforms, including Getty Images and Shutterstock that provide strong provenance for training data to support generative AI development. NVIDIA hopes that industry-driven efforts such as these will lead to common content authenticity standards and measures.

AI Transparency and Documentation:

NVIDIA endorses model cards as a best practice to communicate common quality expectations to developers and nontechnical stakeholders alike. Model cards are short documents describing who made the model, the model's training and evaluation data, the model's intended use and intended outcome. Being transparent about dataset acquisition through model cards as a best practice also supports common provenance standards. NVIDIA has been using model cards since 2021, and in 2023 began updating our Model cards with additional information regarding bias, explainability, privacy, and safety and security. More information about NVIDIA's enhanced, next-generation Model Card++, can be found here and here.

Red-teaming best practices for AI safety:

Red-teaming is most effective when performed by an interdisciplinary team in addition to AI specialists. Red teams should include domain and interdisciplinary experts able to assess the risk of unintended outputs. NVIDIA also encourages and supports opportunities for red-teaming collaboration and information sharing among industry, academia, and civil society to build trust in AI.

How to design AI red-teaming exercises for different types of model risks, including specific security risks (*e.g.,* CBRN risks, etc.) and risks to individuals and society (*e.g.,* discriminatory output, hallucinations, etc.:

Red-teaming exercises must have well-defined objectives. Red-teaming should be clearly differentiated from more general model evaluation. Success criteria should be explicitly stated upfront and agreed upon by domain experts to validate and interpret red-teaming observations. Special attention should be paid to the training data, outputs, the role that the model will play in a potential deployed system, and consequences for out-of-distribution performance.

2. Reducing the Risk of Synthetic Content

Techniques for labeling synthetic content, such as using watermarking:

NVIDIA encourages synthetic content labeling, whether explicit (visible labels in the form of a text or an image) or implicit (introduced by altering audio, image, or video content that are not directly sensed by humans but can be extracted by technical

means). Synthetic content labels or markers should be detectable after common modifications like an image crop, rotation, or edit; designed to minimize impacts to model accuracy; and easily legible without significantly obscuring content. Although existing watermarking technologies (e.g. re-encoding) have limitations, they are helpful and readily available interim solutions.

3. Advance Responsible Global Technical Standards for AI Development

AI nomenclature and terminology:

Shared nomenclature and terminology are critical to advancing understanding and partnerships on AI safety and security. NVIDIA encourages NIST to continue to drive efforts towards the development of common nomenclature and terminology for AI.

Examples and typologies of AI systems for which standards would be particularly impactful (*e.g.,* because they are especially likely to be deployed or distributed across jurisdictional lines, or to need special governance practices):

NVIDIA endorses the following best practices for AI model training:

Model Parallelism (tensor, sequence & pipeline): A distributed training method that enables scaling across multiple devices, allowing developers to partition a model across more than one device. NVIDIA Megatron-LM supports model parallelism.

Efficiently Scale Large Language Model Training Across a Large GPU Cluster with Open-Source Frameworks: Open-source frameworks like Alpa.ai and Ray.io can help train large parameter models. Alpa is a unified compiler that automatically discovers and executes the best interoperator and intraoperator parallelism for large deep learning models. Ray.io offers a distributed computing framework that enables simplified scaling and management of resources across multiple machines. More information about both can be found here: Efficiently Scale LLM Training Across a Large GPU Cluster with Alpa and Ray | NVIDIA Technical Blog.

PEFT (Parameter-Efficient Fine-Tuning): PEFT methods fine-tune only a small number of extra model parameters, significantly decreasing computational and storage costs. Recent state-of-the-art PEFT techniques achieve performance

comparable to that of full fine-tuning. (Example: huggingface/peft is a library for efficiently adapting pre-trained language models (PLMs) to downstream applications without fine-tuning all the model's parameters.)

Instruction Tuning for Large Language Models (LLMs): Instruction tuning, or tuning with pairs of input-output instructions, enables LLMs to increase their content generation capabilities. More information can be found here: Instruction Tuning for Large Language Models: A Survey.

Federated Learning for LLMs on Distributed Datasets: Federated learning shares only models with updated parameters, not the data they may have been trained on, preserving privacy of training data. More about Federated Learning can be found here: Adapting LLMs to Downstream Tasks Using Federated Learning on Distributed Datasets | NVIDIA Technical Blog.

Model Pruning and Quantization using Transfer Learning: Pruning removes nodes from the neural network that contribute less to the overall accuracy of the model, reducing the overall size of the model and its memory footprint while also increasing inference throughput, which are all important for edge deployments. Quantization transforms deep learning models to use parameters and perform computations quicker and more efficiently by using lower precision. Both pruning and quantization can be accomplished using NVIDIA Transfer Learning Toolkit TAO Toolkit (5.2.0) - NVIDIA Docs.

Ways to improve the inclusivity of stakeholder representation in the standards development process:

Standards should be developed with multi-stakeholder input including industry, government experts, academia and civil society across diverse experiences, disciplines, backgrounds, and demographics.

Strategies for driving adoption and implementation of AI-related international standards:

NVIDIA endorses and promotes collaboration through well-recognized standards development organizations such as the International Organization for Standardization and International Electrotechnical Commission. NVIDIA encourages NIST and other government agencies to facilitate robust discussion

between the United States and partners and allies to align on respective domestic standards (for example, mapping exercises between national standards and guidelines). Harmonizing international standards and guidelines will help companies like NVIDIA and other companies innovate by applying a consistent and predictable set of standards and best practices.

NVIDIA is grateful for the work of NIST and the Department of Commerce to support a vibrant and innovative ecosystem and to develop and deliver safe, secure, and trustworthy artificial intelligence.

- NVIDIA Corporation