FRONTIER AI RISK ASSESSMENT

Barnaby Simkin, Nikki Pope, Leon Derczynski, Christopher Parisien

Applicable from August 2025

Abstract

As a leader in AI, NVIDIA recognizes the transformative potential of artificial intelligence to improve lives and solve some of the world's hardest challenges. This paper describes how NVIDIA's risk framework is applied to help identify, mitigate, and address potential harms arising from frontier AI. Even though frontier AI models are not currently under development at NVIDIA, our existing risk framework can be applied to identify emerging capabilities within our advanced AI models and put in place appropriate risk-mitigation measures. AI capabilities and their associated risks evolve rapidly. Therefore, our risk management framework will be regularly reviewed and updated to reflect new findings, emerging threats, and ongoing advancements in the industry. This iterative approach will ensure our risk assessment remains fit-for-purpose over time.

Executive Summary

For the purposes of this assessment, a 'frontier model' is defined as a highly capable general-purpose AI model that can perform a wide variety of undefined tasks and exceeds the capabilities present in the most advanced models currently in existence. These complex models may interact with the world in ways that may be unpredictable or even harmful if not managed responsibly. To address these challenges, our AI risk framework offers a structured and proportionate approach for analyzing risk that also takes into account the nature of the product under assessment. Our risk framework comprises two main components: a Preliminary Risk Assessment (PRA) and a Detailed Risk Assessment (DRA). The PRA functions as an initial filter during the design phase, assigning products to broad risk categories and identifying what sections of the DRA are applied on a recommended or mandatory basis. The DRA then examines the product's architecture and development processes in detail, identifies use case specific hazards, assigns more granular risk scores based on those hazards, and recommends methods for risk mitigation. Our risk evaluation process then estimates the residual risk after controls are applied and compares it against the potential initial risks posed by the AI-based product. Leveraging the results from the risk evaluation phase, it is possible to determine how residual risks correspond with NVIDIA's Trustworthy AI (TAI) principles and document any trade-offs made during the allocation of risk treatment measures. All relevant data from the risk evaluation process is then stored in our model cards.

Risk Assessment Process

Preliminary Risk Assessment (PRA)

The preliminary risk assessment uses three key criteria to separate AI models into different risk categories. Risk categories are allocated by looking at what the model is designed to do (its capabilities), where it will be deployed (its use case), and how autonomously it operates (level of autonomy). For example, an object detection model in a retail setting, used primarily to monitor stock levels or customer flows, may be classified as relatively low risk, especially if it operates under human supervision. However, the same type of object detection model used in a healthcare context to detect surgical instruments would be deemed higher risk, as mistakes or malfunctions have direct implications for patient health and safety. An object detection algorithm employed in retail to monitor store entries, but deployed as a fully autonomous agent would contain more risk than a semi-autonomous or supervised version. The increased autonomy heightens the potential impact of system errors through unauthorized interventions, thereby elevating its overall risk category.

Each risk criteria have discrete thresholds between 1 and 5 that are used to determine a model's risk category[1]. The PRA will assign a model risk (MR) score between 1 and 5 based on the highest MR score within this criteria. Below is a non-exhaustive list of attributes used to define the MR score.

	MR1	MR2	MR3	MR4	MR5
Intended use case	Retail Entertainment No intended industry	Manufacturing Financial services Education Agriculture	Healthcare Robotics Transportation Politics	Defense and security	
Capabilities	Object detection Clustering Recommendation engine Machine translation	Image / speech synthesis Visual reasoning Text generation	Molecule discovery Image manipulation	Biometric identification Wide-variety of distinct capabilities	Wide-variety of undefined capabilities
Level of autonomy	Inference API with or without user interface	Deterministic agent	Non-deterministic agent	Autonomous agent with human oversight	Autonomous agent without user approval

The MR score is correlated to the maximum permissible harm relative to our trustworthy AI principles[2]. High risk models require more intensive scrutiny, increased oversight and face stricter development and operational constraints[3]. The level of governance associated with each MR levels can be broadly grouped into the following categories:

- MR5 - A detailed risk assessment should be complete and approved by an independent committee e.g. NVIDIA's AI ethics committee.
- MR4 – A detailed risk assessment should be complete and business unit leader approval is required.

[1] Campos et al (2024). A Framework to Rate AI Developers' Risk Management Maturity - https://www.safer-ai.org/research-posts/a-framework-to-rate-ai-developers-risk-management-maturity

[2] Koessler et al (2024). Risk thresholds for frontier AI - https://arxiv.org/pdf/2406.14713

[3] Kolt et al (2024). Responsible Reporting for Frontier AI Development - https://arxiv.org/pdf/2404.02675

- MR3 - Risk mitigation measures and evaluation results are documented by engineering teams and periodically reviewed.
- MR2/MR1 – Evaluation results are documented by engineering teams.

It's important to note that initial MR score should not be reduced through typical risk mitigation measures e.g. a perception module for an automated vehicle will always have higher risk (MR3) even with perfect perception and redundancy measures in place. A frontier model would be classified as MR5 due to potential adversarial capabilities associated with these highly capable models operating within an undefined domain. Implementing a simplified method for risk categorization does present some known challenges around reliability. However, we believe these factors are a more effective proxy for risk than relying on compute thresholds[4].

Detailed Risk Assessment (DRA)

Summary:

Formal processes for risk assessments in the safety and security domain are very mature but the challenge comes when adapting existing processes to cover risks related to other trustworthy AI principles so that risks are directly comparable. The DRA covers an assessment of a product's architecture and development processes and consists of five key components: use case specification, hazard identification, risk analysis, risk mitigation and risk evaluation. These risk assessment components are integrated into the product development lifecycles. All product lifecycles follow a v-model used to structure the stages of design, development and testing. The left side of the 'v' represents the design stage and the decomposition of requirements and design of a product, and the right side represents verification and validation. Each development phase on the left has a corresponding testing phase on the right, ensuring that each requirement is verified and validated throughout the development lifecycle. V-models are established for the development of system, model and dataset lifecycles. The dependencies between the risk assessment and product lifecycles are shown in Figure 1. V-models are typically carried out in an agile manner and are run through multiple times as the product matures, or new potential hazards are identified. Risk assessments are periodically reviewed, and repeated if pre-defined thresholds are met e.g. technology matures, component is significantly modified, operating conditions change, or a hazard occurs with high severity or frequency. If a product's MR rating is increased during reassessment, then the new governance measures should be applied before the latest version of the product is released. For example, if a product's use case was changed from 'transportation' to 'defense' the model's risk category would change from MR3 to MR4 and it would require a business unit leaders' approval before release.

[4] Frontier Model Forum (2024). Components of Frontier AI Safety Frameworks - https://www.frontiermodelforum.org/updates/issue-brief-components-of-frontier-ai-safety-frameworks/

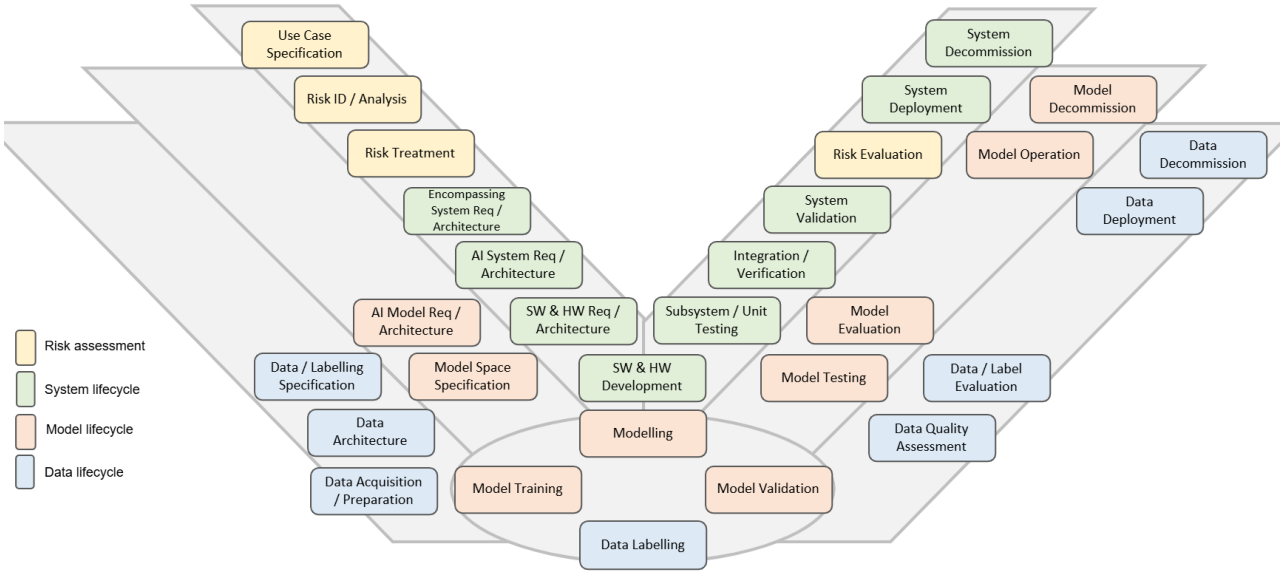

Figure 1: V-model for system, model and datasets for the providers of AI systems

If designing an AI-based system for a well-defined intended use case then the risk mitigation measures can be applied to the system, model and data lifecycles. However, if a model is being assessed without a specific use case then the controls are applied to the model and data lifecycles. When developing an AI model, it is important to record assumptions about the intended use case (if any) to provide context around model quality and any known limitations. The output from these assessments are documented in our model cards[5] and supports our customers when safely integrating our models into downstream applications or systems.

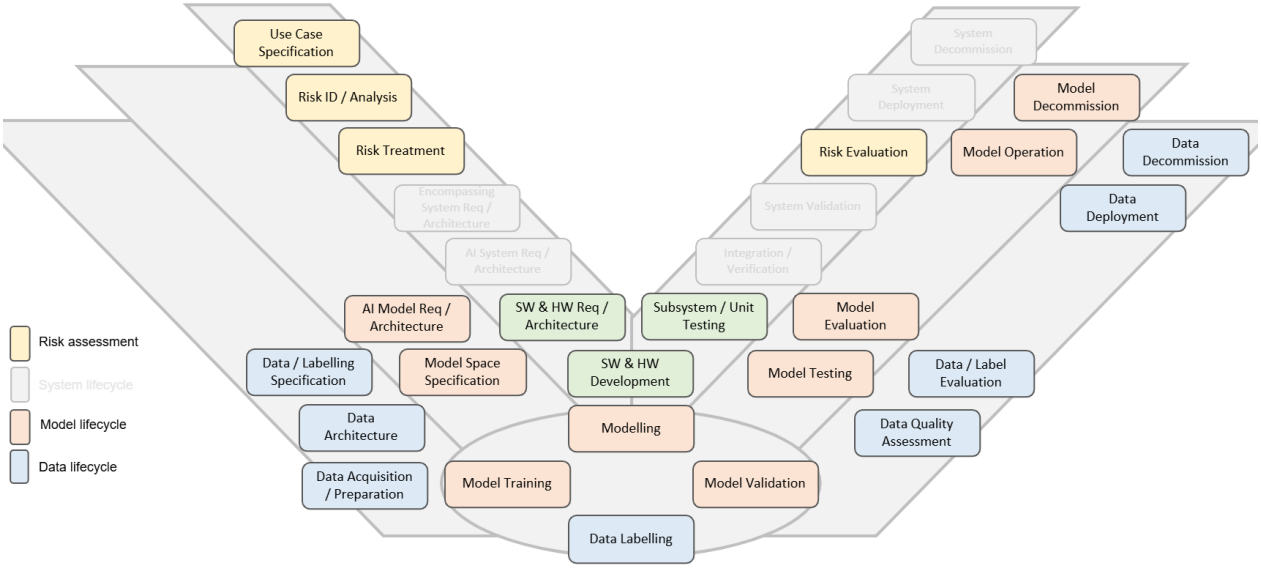

Figure 2: V-model for system, model and datasets for the providers of foundation models

[5]NVIDIA Model Cards - https://developer.nvidia.com/blog/enhancing-ai-transparency-and-ethical-considerations-with-model-card/

To determine which version of the DRA is carried out it is necessary to identify whether or not a product is classified as a system. We define a system as *'a set of integrated components that interact with external inputs or users to provide an end-to-end functionality or service, typically with a clearly defined purpose and measurable outcomes.'*

NVIDIA mainly provides the following types of AI-based products:

- 'AI model files' which are designed to have specific capabilities, such as image classification, language translation, or anomaly detection. On its own, an AI model is just a piece of software code plus learned parameters with no user interface, deployment infrastructure, or additional logic.
- 'Containers' which bundles an AI model and its dependencies (libraries, runtime environment) into a portable unit. Containers make it simpler to deploy and run AI models across various computing environments.
- 'Web-based inference API' which provides a cloud-based endpoint for sending input data and receiving outputs from an AI model.
- 'AI systems' that are embedded into an encompassing system e.g. robotics, IoT devices, automated vehicles.

Our AI models files and containers are not systems as they require additional layers to become a system. AI systems are embedded into encompassing systems and are classified as systems as they combine all necessary components to deliver a complete function or service without requiring further integration. Our web-based inference APIs on build.nvidia are designed for users to experiment with a model's functionality and for them to ultimately integrate those models into their downstream applications or systems. Web-based inference APIs with user interfaces could be classified as a system as they form a complete loop of user input, AI inference, and output. However, we do not classify them as normal systems because they do not have a clearly defined intended use case or cannot influence their environment. For this reason, we take a hybrid approach in the risk assessment. We document assumptions and limitations in the model card but also factor in controls that can be applied to the system architecture e.g. recording use, rate limiting, input/output restriction etc.

Understandably there are fewer opportunities to mitigate potential risks when an AI model is made available for download. That's not to say that the levels of residual risk in models are always higher than systems but the types of hazards, the capacity for risk propagation and methods for risk mitigation needs to be carefully considered based on the deployment strategy.

Use case specification:

A use case specification should include a description of: the product's objectives, architecture, development processes and the operational design domain (ODD). The ODD refers to the specific conditions under which a product is intended to operate effectively and includes a description of the impacted stakeholders. The ODD can therefore be used to help structure a suitable test dataset for model evaluation. The methods for describing the ODD depend on the data modality applicable to the model. For example, the ODD for image-based models should describe the visual characteristics of the operating environment. Additionally, it is essential to identify all ODD attributes that are significant for the model's prediction but are not measurable or annotated, as this signals a limitation in the evaluation process.

The difficulty in assessing the trustworthiness of frontier AI models is typically due to their open-ended ODDs and therefore their unpredictable nature in diverse scenarios[6]. The PRA sets a threshold for MR5 (frontier models) based on a 'wide variety of undefined capabilities' or high levels of autonomy. In contrast to what is noted above, MR levels can be reduced through use case restrictions. For example, a model classified as MR5 could be reduced to MR4 by evaluating the model quality within a well-defined ODD then technically restricting the model to being used inside the conditions in which the initial evaluation occurred. Use case restriction and preventing high levels of autonomy can dramatically reduce systemic risks associated with frontier AI models.

Risk Identification and Analysis:

NVIDIA's Trustworthy AI Principles are derived from human rights and legal principles. These principles are used as a foundation for defining a broad range of potential risks that a product may be exposed to. Based on the description of a product's architecture and development workflows it should be possible to identify possible hazards, estimate the level of risk for each hazard and categorize the cumulative risk relative to our trustworthy AI principles. As stakeholders may perceive risk differently, the exact quantity of each individual risk does not provide much insight. Understanding the relative perceived risk level across multiple models and risk types is far more valuable. We therefore developed a standardized approach for measuring risk across different trustworthy AI principles that enables engineers to make more informed decisions about trade-offs that need to be made when mitigating risk.

We defined risk as the potential for an event to lead to an undesired outcome, measured in terms of its likelihood (probability), its impact (severity) and its ability to be controlled or detected (controllability). The risk associated with each hazard is scored between 1 and 64, with the higher value indicating a higher risk.

Risk = likelihood x severity x observability
Risk = frequency x (duration + speed of onset) x (detectability + predictability)

A hazard that has a non-zero but very low probability of occurring, that is transient in nature, occurs gradually, easy to detect and localized has the lowest risk score. In contrast, a hazard that has a high probability of occurring, is permanent in nature, occurs instantaneously and randomly due to latent faults has the highest risk score.

[6] AI Safety Summit (2023). Capabilities and risks from frontier AI - https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf

Risk Level				
Risk measurements			**Individual Score**	**Total max score**
Likelihood	Frequency	High Probability	4	4
		Medium probability	3	
		Low probability	2	
		Very low probability	1	
Severity	Duration	Permanent	2	2
		Intermittent	1	
		Transient	0.5	
	Speed of onset	Instantaneous	2	2
		Rapid	1	
		Gradual / cumulative	0.5	
Observability	Detectability	Latent faults	2	2
		Hard to detect	1	
		Easily detectable	0.5	
	Predictability	Random	2	2
		Uniform	1	
		Localized	0.5	

Table 1: NVIDIA's methodology for estimating the level of risk associated with a hazard

NVIDIA has a comprehensive repository of potential hazards that has been carefully curated and mapped to assets to help guide teams to understand potential risks related with their products. This repository has been created using a variety of sources e.g. stakeholder consultation, market data, incident reports (AI Vulnerability database, AI Incident database, AAAIC database, OECD.ai AI Incidents Monitor). This approach is suitable when we have a well-defined set of capabilities and a known use case for a specific model. However, for frontier models we need to consider speculative risks that may or may not be present in the model.[7] To help detect specific adversarial capabilities, models will be stress-tested against extreme but plausible scenarios that may lead to systemic risks. This approach ensures that both known and emergent hazards are taken into account.[8] A list of potential systemic risks associated with frontier AI models were

[7] Anderljung et al (2023). Frontier AI Regulation, Managing Emerging Risks to Public Safety - https://arxiv.org/pdf/2307.03718

[8] Campos et al (2024). Campos A Framework to Rate AI Developers' Risk Management Maturity - https://www.safer-ai.org/research-posts/a-framework-to-rate-ai-developers-risk-management-maturity

identified using the risk analysis we designed and confirmed by reviewing existing literature[9] and academic research. In particular, frontier models may have the capacity to present the following hazards.[10]

- Cyber offence e.g. risks from using AI for discovering or exploiting system vulnerabilities.
- Chemical, biological, radiological, and nuclear risks[11] e.g. AI enabling the development and use of weapons of mass destruction.
- Persuasion and manipulation e.g. influence operations, disinformation, and erosion of democratic values through AI-driven content.
- At-scale discrimination e.g. bias and unlawful discrimination enabled by AI systems.

Risk Mitigation:

Due to the scale and diversity of AI models deployed at NVIDIA it is infeasible to map specific controls to risks and enforce their integration across all model types and use cases. Engineering teams are permitted to deploy controls that are relevant to their products, but increased levels of oversight are applied to models with higher risk. Recognizing that risk cannot be entirely eliminated, the effectiveness of each control is evaluated according to its impact on the attributes used to calculate the initial risk e.g. prompt-based guardrails that reduce the frequency of adversarial prompts being inputted into a model. Table 2 provides an example of how a risk analysis may be documented for models that have the capabilities to spread disinformation.

Risk identification and Analysis					Risk Mitigation			
Asset	**Hazard source**	**Hazard**	**TAI Principle**	**Risk Analysis**	**Control**	**Impacted asset**	**Risk Impact**	**Residual Risk**
Model	Adversarial prompt	Disinformation	Safety	49	Block toxic prompts	Input data	Reduce likelihood	5
					Rate limiting	System	Reduce speed of onset	
					Implicit watermark	Output data	Increase detectability	

Table 2: Template for risk analysis

As a provider of open-source models, NVIDIA is also committed to providing open-source tooling that enables that trustworthy development and integration of AI models into downstream applications. NVIDIA offers a range of tools that can be use to address risk: either through reducing the risk itself or by increasing the coverage of test cases for measuring risk.

[9] Government Office for Science (2023). Future Risks of Frontier AI - https://assets.publishing.service.gov.uk/media/653bc393d10f3500139a6ac5/future-risks-of-frontier-ai-annex-a.pdf

[10] OpenAI (2023). OpenAI's Approach to Frontier Risk - https://openai.com/global-affairs/our-approach-to-frontier-risk/

[11] Centre for Security and Emerging Technologies (2023). Anticipating and Managing Risks from Frontier AI Systems - https://cset.georgetown.edu/wp-content/uploads/Frontier-AI-Roundtable-Paper-Final-2023CA004-v2.pdf

Decreasing the frequency of a hazard

NeMo Guardrails[12] acts as a safety layer on top of language models to enforce predefined rules and policies during inference. NeMo Guardrails is an opensource framework that contains proprietary and 3rd party safeguards. As new threats emerge, Guardrails can be updated with new safeguards. This adaptability encourages deployed models to remain trustworthy over time, even as the landscape of acceptable AI behavior evolves. NeMo Guardrails library currently includes:

- Jailbreak detection techniques through Ardennes[13]
- Output checking through Presidio[14] or ActiveFence[15]
- Fact checking through AlignScore[16]
- Hallucination detection through Patronus Lynx[17] or CleanLab[18]
- Content safety through LlamaGuard[19] or Aegis[20] content safety

Deploying safeguards across various points in a model's architecture ensures that if one layer is compromised, others remain effective. This approach enhances resilience against potential risks by providing redundant protective measures[21]. Guardrails can be implemented at various locations in the model architecture

- Input rails are guardrails applied to the input from the user; an input rail can reject the input, stopping any additional processing, or alter the input (e.g., to mask potentially sensitive data, to rephrase). Cosmos pre-Guard leverages Aegis-AI-Content-Safety-LlamaGuard-LLM-Defensive-1.0, which is a fine-tuned version of Llama-Guard trained on NVIDIA's Aegis Content Safety Dataset and a blocklist filter that performs a lemmatized and whole-word keyword search to block harmful prompts. It then further sanitizes the user prompt by processing it through the Cosmos-Text2World Prompt Upsampler.
- Dialog rails influence how the LLM is prompted; dialog rails operate on canonical form messages and determine if an action should be executed, if the LLM should be invoked to generate the next step or a response, if a predefined response should be used instead, etc.
- Retrieval rails are guardrails applied to the retrieved chunks in the case of a RAG (Retrieval Augmented Generation) scenario; a retrieval rail can reject a chunk, preventing it from being used to prompt the LLM, or alter the relevant chunks (e.g., to mask potentially sensitive data).
- Execution rails are guardrails applied to input/output of the custom actions (a.k.a. tools), that need to be called by the LLM.

---

[12] NeMo Guardrails - https://docs.nvidia.com/nemo/guardrails/introduction.html
[13] Andennes - https://catalog.ngc.nvidia.com/orgs/nim/teams/nvidia/containers/nemoguard-jailbreak-detect
[14] Presidio - https://microsoft.github.io/presidio/analyzer/
[15] ActiveFence - https://www.activefence.com/solutions/automated-content-moderation/
[16] AlignScore - https://github.com/yuh-zha/AlignScore
[17] Patronus Lynx - https://docs.patronus.ai/docs/hallucination-detection
[18] CleanLab - https://cleanlab.ai/tlm/
[19] LlamaGuard - https://ai.meta.com/research/publications/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations/
[20] Aegis content safety - https://huggingface.co/nvidia/Aegis-AI-Content-Safety-LlamaGuard-Permissive-1.0
[21] Institute for AI Policy and Strategy (2024). Adapting cybersecurity frameworks to manage frontier AI risks: A defense-in-depth approach - https://arxiv.org/pdf/2408.07933

- Output rails are guardrails applied to the output generated by the LLM; an output rail can reject the output, preventing it from being returned to the user, or alter it (e.g., removing sensitive data). Cosmos post-Guard stage blocks harmful visual outputs using a video content safety classifier and a face blur filter.

Models are typically trained using both proprietary and publicly available datasets. To ensure effective distribution of the dataset we employ taxonomy-based classifiers to label data types and prune those that introduce unrealistic or noisy patterns. Certain categories relevant to the model's use case may be upsampled, while less critical ones are downsampled. A significant amount of initial training data can be semantically redundant, which may induce unwanted artifacts in the model's output if not appropriately handled. We therefore use a sequence of data processing steps to find the most valuable parts of the data for training. The risk of memorization is higher where data appears more than once in the training dataset.[22] , so a deduplication step is normally used to identify near-duplicate content and preserves the highest quality version for minimal data loss.

Data quality is a recurring focus across guardrails. In addition to conventional approaches for measuring data quality we have developed and shared novel methods for detecting AI generated content in training datasets through HIVE[23] as a NIM. We also filter training data to exclude examples that could result in capabilities that increase the likelihood of misuse, such known CSAM images. Ardennes, NVIDIA Guardrails Topical rails, and AEGIS rely heavily on data quality assessment and assurance through frequent model evaluation as part of their development processes. When directly working with downstream providers like Getty Images[24] we've also created datasets that have strict and comprehensive consent over the use of copyrighted data, ensuring that appropriate attribution is provided to the creators of the original data used to train an AI model.

Hazard detection

NeMo Evaluator[25] provides a microservice to assess generative AI models and pipelines across academic and custom benchmarks on any platform. It goes beyond simple accuracy metrics, offering comprehensive evaluations that highlight potential model vulnerabilities or unexpected failure modes. By integrating with CI/CD pipelines, Evaluator can continuously test new model versions or updates against controlled test sets. This systematic oversight ensures that changes do not degrade model performance or introduce new risks, increasing stakeholders' trust in ongoing development cycles. Typically, benchmarks produce aggregated metrics that allow for one model to be directly compared to another. These metrics provide a global analysis but doesn't reliably show where the weaknesses are. To support a deeper analysis of model quality we leverage technology from QuantPi**.** Their tools allow us to use open-source and NVIDIA models as feature embedders, enhancing test data context and supporting safety or bias evaluations.Other AI models can also be used as perturbers that can provide systematic noise or random noise to assess the robustness of models. Leveraging other AI models to augment test data can increase the breadth and depth of assessment beyond public benchmarks

[22]Carlini et al (2023). Extracting training data from diffusion models

[23] HIVE - https://hivemoderation.com/ai-generated-content-detection

[24] Getty Image's commercially safe AI - https://newsroom.gettyimages.com/en/getty-images/getty-images-introduces-updated-ai-model-with-increased-speed-quality-and-accuracy

[25] NeMo Evaluator - https://developer.nvidia.com/blog/streamline-evaluation-of-llms-for-accuracy-with-nvidia-nemo-evaluator/

Watermarks embedded at generation-time enable downstream detection and attribution of AI-produced outputs, providing a verifiable origin signal for both end-users and automated scanning tools. This ensures that when models produce text, images, or other media, it can be reliably traced back to a specific model version or developer entity. Implicit watermarks are subtle, often statistical or cryptographic markers embedded directly into the distributional properties of generated outputs. They are not perceivable by the human eye or ear and do not alter the visible or audible characteristics of the content. There are several opensource implicit watermarking tools that may be used to increase the detectability of content generated by NVIDIA models. Explicit watermarks are visible or otherwise directly perceivable indicators, such as logos, text overlays, or perceptible patterns. They clearly denote AI-generated content to end-users. The use of explicit watermarks are specific to downstream developers. NVIDIA encourages the use of explicit watermarks to help mitigate the spread of misinformation but is not integrated into NVIDIA's foundation models.

Increasing predictability of hazards

One effective approach to increase the predictability of a hazard is to restrict the scope and use of a model. This is achieved by imposing capability or feature restrictions, such as limiting the types of inputs a model can process. Additionally, models may be explicitly barred from prohibited applications through legal mechanisms such as NVIDIA's End User License Agreements for foundation models.[26] Another important strategy involves restricting advanced autonomy functions like self-assigning new sub-goals or executing long-horizon tasks, as well as tool-use functionalities like function calls and web browsing.

Lowering hazard duration

Lowering the duration of a hazard can be achieved by implementing establish robust protocols for managing AI-related incidents[27], including clear information-sharing mechanisms between developers and relevant authorities[28]. This encourages proactive identification of potential risks before they escalate. Additionally, reducing access to a model reactively when misuse is detected can help limit further harm. This can involve rolling back a model to a previous version or discontinuing its availability if significant misuse risks emerge during production. Lastly, conducting regular safety drills ensures that emergency response plans are stress-tested. By practicing responses to foreseeable, fast-moving emergency scenarios[29], NVIDIA is able to improve their readiness and reduce the duration of hazardous incidents.

Decreasing hazard onset speed

Decreasing the speed of onset for a hazard is essential in managing risks associated with frontier AI models. Key strategies include maintaining human oversight by avoiding full autonomy in critical systems and ensuring a human-in-the-loop for all decisions in high-stakes contexts. This slows down potentially harmful automated actions, allowing for

[26] NVIDIA EULA for Foundation models - https://www.nvidia.com/en-us/agreements/enterprise-software/nvidia-ai-foundation-models-community-license-agreement/

[27] NVIDIA security vulnerability reporting - https://www.nvidia.com/en-us/product-security/report-vulnerability/

[28] Centre for Security and Emerging Technologies (2023). Anticipating and Managing Risks from Frontier AI Systems - https://cset.georgetown.edu/wp-content/uploads/Frontier-AI-Roundtable-Paper-Final-2023CA004-v2.pdf

[29] Uuk et al (2024). Effective Mitigations for Systemic Risks from General-Purpose AI - https://arxiv.org/pdf/2412.02145

intervention. Access control measures further mitigate risks. These include ensuring only authorized users access the model through secure API keys and authentication protocols, performing Know-Your-Customer (KYC) screenings for users with high output needs, and limiting access frequency by capping requests[30] or instituting time-based quotas. Proactive monitoring is equally critical. This includes detecting and blocking misuse attempts using algorithmic classifiers, which can limit unsafe queries, modify responses, or block users attempting to bypass safeguards. Initially, model access can be restricted to a limited audience, expanding gradually as risks are better understood and mitigated. In cases of severe risk, notifying other developers of identified hazards through the proven channel of NVIDIA's security bulletin[31] allows for coordinated response efforts, mitigating widespread issues[32]. As a last resort, full market removal or deletion of the model and its components can be considered to prevent further misuse and contain hazards effectively[33].

Risk evaluation

We're committed to conducting comprehensive testing to identify our model susceptibilities related to systemic risks. This proactive approach aims to uncover and mitigate potential risks before public deployment. When a model shows capabilities of frontier AI models pre deployment we will initially restrict access to model weights to essential personnel and ensure rigorous security protocols are in place.[34] Measures will also be in place to restrict at-will fine tuning of frontier AI models without safeguards in NeMo customizer, reducing the options to retrain a model on data related to dangerous tasks or to reduce how often the model refuses potentially dangerous requests.

Identifying early warning signs for these potential hazardous capabilities are crucial to mitigating systemic risk in frontier AI models.[35] Common public benchmarks are excellent tools for providing broad coverage over curated data samples and easing comparison between published models. Public benchmarks are currently available to test for capabilities associated with manipulation or large-scale discrimination, with the current generation including e.g.

- TruthfulQA,[36] FEVER,[37] and GLUE[38] test a model's tendency to generate false or misleading content.
- BBQ[39] and BOLD[40] test open-ended generation for biased language.
- WMDP[41] benchmark serves as both a proxy evaluation for hazardous knowledge in large language models (LLMs) and a benchmark for unlearning methods to remove such knowledge.

Whilst many public benchmarks exist, not many are directly targeted to measure frontier risks. In such cases, existing benchmarks may need to be repurposed or combined to create robust testing environments.

---

[30] NVIDIA API credits - https://nvidia.github.io/GenerativeAIExamples/0.5.0/api-catalog.html
[31] NVIDIA's security bulletin - https://www.nvidia.com/en-us/security/
[32] Alaga et al (2023). Coordinated pausing, An evaluation-based coordination scheme for frontier AI developers - https://arxiv.org/pdf/2310.00374
[33] O'Brien et al (2023). Deployment Corrections, An incident response framework for frontier AI models - https://arxiv.org/pdf/2310.00328
[34] METR (2024). Common Elements of Frontier AI Safety Policies - https://metr.org/assets/common_elements_of_frontier_ai_safety_policies.pdf
[35] RAND (2024). Evaluating Artificial Intelligence for National Security and Public Safety - https://www.rand.org/content/dam/rand/pubs/conf_proceedings/CFA3400/CFA3429-1/RAND_CFA3429-1.pdf
[36] TruthfulQA - https://github.com/sylinrl/TruthfulQA
[37] FEVER - https://paperswithcode.com/sota/fact-verification-on-fever
[38] GLUE - https://gluebenchmark.com/
[39] BBQ - https://paperswithcode.com/dataset/bbq
[40] BOLD - https://github.com/amazon-science/bold
[41] https://www.wmdp.ai/

- MBPP[42] measures code synthesis ability but would need adaptation to test for malicious code patterns.
- MoleculeNet[43] could be repurposed to determine whether the model can generate toxic compounds.
- ARC[44] can be adapted to detect if a model's presents capabilities beyond those it is intended or trained to have

AILuminate v1.0 from MLCommons[45] is one of the few benchmarks that is intended to evaluate frontier AI models across various dimensions of trustworthiness and risk. AILuminate broadens the scope to assess attributes such as robustness, fairness, explainability, compliance with ethical guidelines, and resilience to adversarial inputs. It aims to provide a more holistic view of a model's behavior and potential impacts in real-world scenarios.

As threats and vulnerabilities evolve, a benchmark may become outdated. Regular updates are needed to remain relevant, which can lead to versioning complexity and continuous re-benchmarking. Certain risks may also be hard to capture in a single, standardized framework. The benchmark might miss emergent, scenario-specific failure modes. Red teaming activities are used in conjunction with public benchmarks to address those limitations and capture those emerging risks that cannot be directly measured through benchmarking. In adversarial red teaming, expert human operators deliberately probe a frontier AI model's vulnerability and attempt to induce it to produce harmful, biased, or disallowed outputs. The red team also probes each guardrail component independently with targeted examples to identify weaknesses and improve performance in edge cases. NeMo offers an experimental red-teaming interface[46] that allows developers to run red teaming activities. These human adversaries are able to leverage domain knowledge, creativity, and context-awareness to simulate realistic attack strategies.

To help focus red teaming activities and respond to model vulnerabilities and weaknesses, we first need to be aware of them. In cybersecurity, vulnerability scanners serve the purpose of proactively checking tools and deployments for known and potential weaknesses. For generative AI, we need an analogue. NVIDIA runs and supports the Garak[47] LLM vulnerability scanner. This constantly updated public project collects techniques for exploiting LLM and multi-modal model vulnerabilities and provides a testing and reporting environment for evaluating models' susceptibility. The project has formed a hub with a thriving community of volunteers that add their upgrades and knowledge. Garak can test numerous scenarios rapidly, far exceeding the coverage possible with manual methods. Systematic exploration of model weaknesses can be repeated frequently, ensuring continuous oversight as the model evolves. NVIDIA takes advantage of this and uses Garak as a highest-priority assessment of models before release.

Optimizing Trustworthiness

The rapid advancement in AI development necessitates continuous monitoring and updating of risk frameworks to stay aligned with emerging capabilities and associated risks.[48] Rapid prototyping and evaluating AI models is also a key

[42] MBPP - https://paperswithcode.com/sota/code-generation-on-mbpp
[43] MoleculeNet - https://moleculenet.org/
[44] ARC - https://paperswithcode.com/sota/common-sense-reasoning-on-arc-challenge
[45] AILuminate v1.0 benchmark from MLCommons - https://ailuminate.mlcommons.org/benchmarks/
[46] NeMo Red teaming interface - https://docs.nvidia.com/nemo/guardrails/security/red-teaming.html
[47] Garak - https://github.com/NVIDIA/garak
[48] Department for Science, Innovation & Technology (2023). Frontier AI: capabilities and risks - https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/frontier-ai-capabilities-and-risks-discussion-paper

component to support the agile risk framework. NVIDIA offers a variety of tools that can optimize the speed of model evaluation.

NeMo Customizer[49] is a high-performance, scalable microservice that simplifies fine-tuning and alignment of LLMs for domain-specific use cases, making it easier to adopt generative AI across industries.

NeMo Curator[50] is a GPU-accelerated data-curation tool that enables large-scale, high-quality datasets for pretraining LLMs. NeMo Curator streamlines the process of curating large-scale datasets by filtering, annotating, and organizing data. By ensuring that only high-quality, vetted data enters the training/testing pipeline, it minimizes the risk of bias or malicious content being embedded in the model. NeMo Curator also maintains provenance records of dataset sources, transformations, and filtering criteria. This enables auditability of the data supply chain and supports regulatory compliance, fostering trust that the final model is built on reliable, transparent foundations.

NVIDIA NIMs[51] are containers that encapsulate the entire runtime environment in a self-contained package. Containers isolate application components from the host system and each other. This isolation prevents dependency conflicts, shields the model from external interference, and helps maintain compliance with security policies. Containers also simplify large-scale test orchestration by enabling reproducible deployments across clusters. They make it easier to spin up identical test environments, track configurations, and maintain audit logs, all of which contribute to a transparent and verifiable model lifecycle.

Accelerated computing on GPUs[52] makes large-scale, high-fidelity testing feasible. Thorough stress-testing and red-teaming for frontier AI models should be run at a relatively high frequency during development phases and can require a large amount of processing power. We've introduced a process that can minimize compute needed to run an assessment based on the desired level of confidence. During the development stage when models are refreshed at a higher rate then you could generate an assessment with 50% confidence to ascertain general trends in model performance. Then for data used in the model card an assessment with 95-99% confidence could be carried out.

Governance

Mitigating risks associated with frontier AI models presents a complex governance challenge for any organization, particularly for large companies developing a wide-range of diverse models across multiple industries. The breadth of applications and the dynamic nature of AI technologies make rigid, one-size-fits-all frameworks impractical. Instead, we have adopted a governance approach centered on oversight and adaptive risk management. This strategy allows innovation to flourish while ensuring that development processes remain accountable and transparent. Key to this approach is early detection of potential risks, coupled with mechanisms to pause development when necessary. NVIDIA's internal governance structures clearly define roles and responsibilities for risk management. It involves

[49] Nemo Customizer - https://developer.nvidia.com/blog/simplify-custom-generative-ai-development-with-nvidia-nemo-microservices/?ncid=pa-srch-goog-817720-brand
[50] NeMo curator - https://developer.nvidia.com/blog/simplify-custom-generative-ai-development-with-nvidia-nemo-microservices/?ncid=pa-srch-goog-817720-brand
[51] NVIDIA NIM - https://build.nvidia.com/explore/discover?&ncid=pa-srch-goog-898408-API-Build-Exact&_bt=719593723676&_bk=nvidia%20nim&_bm=e&_bn=g&_bg=169450950363&gad_source=1&gclid=EAIaIQobChMI2oTYkZisigMV7AKtBh3dQzMVEAAYASAAEgLc5vD_BwE
[52] NVIDIA GPU - https://www.nvidia.com/en-us/technologies/

separate teams tasked with risk management that have the authority and expertise to intervene in model development timelines, product launch decisions, and strategic planning.[53] This involves embedding risk-aware practices into the daily work of engineers, researchers, and product managers, supported by ongoing training and open dialogue on ethical considerations.

While our formal model evaluations provide quantitative data, model reviews and interviews with engineering teams reveal developers' intuitive understandings, early warning signs of risks, and internal safety practices. This qualitative approach offers a more nuanced perspective on AI capabilities and potential threats.[54] Establishing consistent communication channels with employees ensures that the correct stakeholders at NVIDIA remain informed about rapid advancements and can promptly address emerging concerns.[55] By integrating these processes into their development lifecycle, we can create a governance framework that is both flexible and robust. This enables responsible AI innovation while proactively managing the unique risks posed by frontier models, ensuring safer and more ethical deployment across various industry sectors.

[53] Centre for Long Term Resilience (2024). Transforming risk governance at frontier AI companies - https://www.longtermresilience.org/wp-content/uploads/2024/07/Transforming-risk-governance-at-frontier-AI-companies-CLTR-1.pdf

[54] Wasil et al (2023). Understanding frontier AI capabilities and risks through semi-structured interviews - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4881729

[55] Wasil et al (2023). Understanding frontier AI capabilities and risks through semi-structured interviews - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4881729