Model Card++

Michael Boone | Trustworthy AI Product Manager

Agenda

Goal: Discuss how parts of NVIDIA’s Model Card++ could feed Voluntary Reporting Template

- What is a model card?
- Why do we need it?
- What is Model Card++?
- How is it different from Model Card?
- How does it help us?
- Demo
- Where is it?
- Questions to prod next steps

What are model cards?

Short documents to describe models

Model Cards on NVIDIA NGC™

Ethics Researchers created the original Model Card as a document to detail how machine learning models work.

Model cards provide information on:

- Performance
- Expected outputs
- License(s)

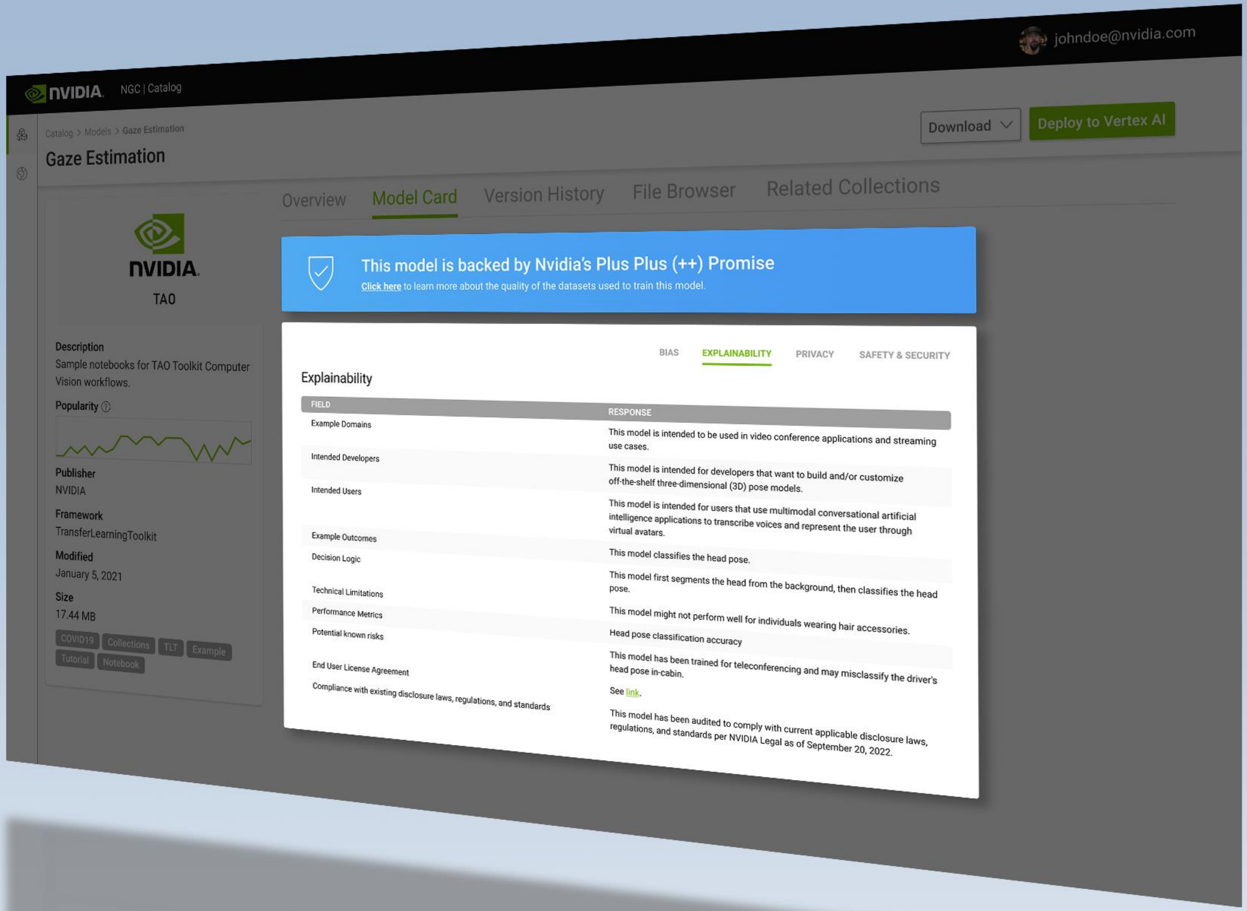
Model cards can help:

- Inform AI decision-makers and beneficiaries
- Educate customers
- Enable the AI community

Why do we need Model Card++?

Communicate AI transparently

- ✓Tie model info directly to model access
- ✓Provide clear and concise content
- ✓Structure sections and format responses consistently
- ✓Detail specific ethical considerations



Gaze Estimation

Description
Sample notebooks for TAO Toolkit Computer Vision workflows.

Popularity

Publisher
NVIDIA

Framework
TransferLearningToolkit

Modified
January 5, 2021

Size
17.44 MB

COVID19 Collections TLT Example Tutorial Notebook

This model is backed by Nvidia's Plus Plus (++) Promise
Click here to learn more about the quality of the datasets used to train this model.

BIAS | EXPLAINABILITY | PRIVACY | SAFETY & SECURITY

Explainability

FIELD	RESPONSE
Example Domains	This model is intended to be used in video conference applications and streaming use cases.
Intended Developers	This model is intended for developers that want to build and/or customize off-the-shelf three-dimensional (3D) pose models.
Intended Users	This model is intended for users that use multimodal conversational artificial intelligence applications to transcribe voices and represent the user through virtual avatars.
Example Outcomes	This model classifies the head pose.
Decision Logic	This model first segments the head from the background, then classifies the head pose.
Technical Limitations	This model might not perform well for individuals wearing hair accessories.
Performance Metrics	Head pose classification accuracy
Potential known risks	This model has been trained for teleconferencing and may misclassify the driver's head pose in-cabin.
End User License Agreement	See link.
Compliance with existing disclosure laws, regulations, and standards	This model has been audited to comply with current applicable disclosure laws, regulations, and standards per NVIDIA Legal as of September 20, 2022.

Includes:

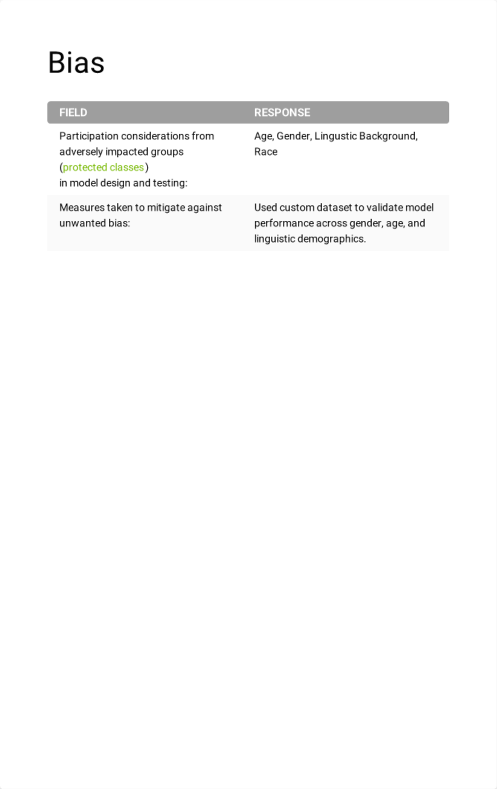
- NGC Model Card Characterizations
- ++ Promise (Triple "P")
- Four (4) Subcards

Introducing Model Card++

The Next Generation Model Card. Introduced 2023.

Model Card++

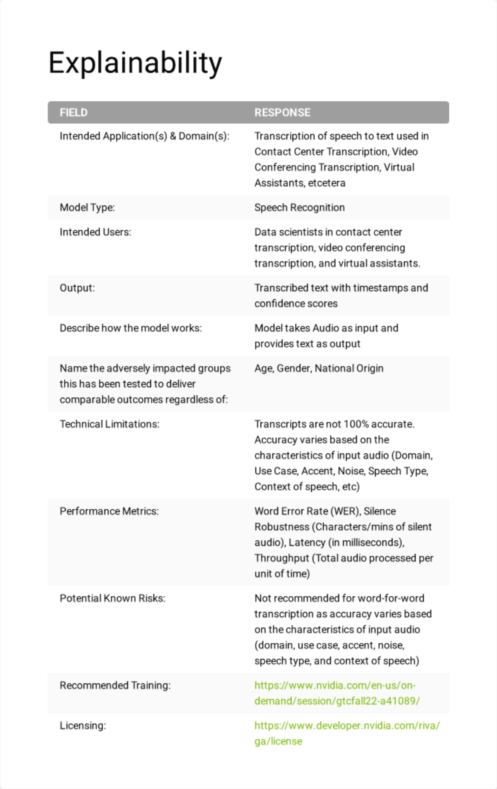
How NVIDIA improved the model card



Bias

FIELD	RESPONSE
Participation considerations from adversely impacted groups (protected classes) in model design and testing:	Age, Gender, Lingustic Background, Race
Measures taken to mitigate against unwanted bias:	Used custom dataset to validate model performance across gender, age, and linguistic demographics.

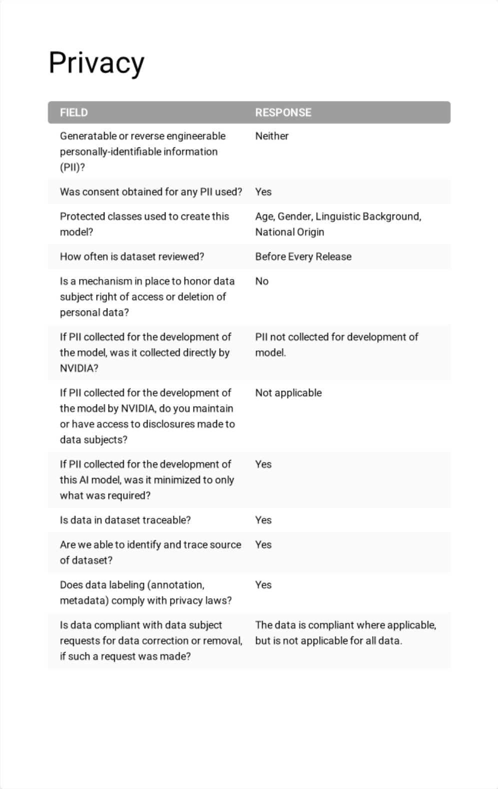
Bias Subcard



Explainability

FIELD	RESPONSE
Intended Application(s) & Domain(s):	Transcription of speech to text used in Contact Center Transcription, Video Conferencing Transcription, Virtual Assistants, etcetera
Model Type:	Speech Recognition
Intended Users:	Data scientists in contact center transcription, video conferencing transcription, and virtual assistants.
Output:	Transcribed text with timestamps and confidence scores
Describe how the model works:	Model takes Audio as input and provides text as output
Name the adversely impacted groups this has been tested to deliver comparable outcomes regardless of:	Age, Gender, National Origin
Technical Limitations:	Transcripts are not 100% accurate. Accuracy varies based on the characteristics of input audio (Domain, Use Case, Accent, Noise, Speech Type, Context of speech, etc)
Performance Metrics:	Word Error Rate (WER), Silence Robustness (Characters/mins of silent audio), Latency (in milliseconds), Throughput (Total audio processed per unit of time)
Potential Known Risks:	Not recommended for word-for-word transcription as accuracy varies based on the characteristics of input audio (domain, use case, accent, noise, speech type, and context of speech)
Recommended Training:	https://www.nvidia.com/en-us/on-demand/session/gtcfall22-a41089/
Licensing:	https://www.developer.nvidia.com/riva/ga/license

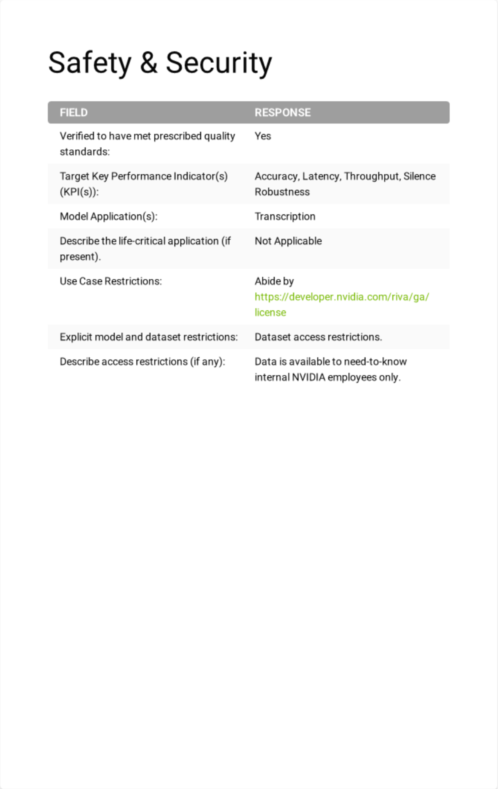
Explainability Subcard



Privacy

FIELD	RESPONSE
Generatable or reverse engineerable personally-identifiable information (PII)?	Neither
Was consent obtained for any PII used?	Yes
Protected classes used to create this model?	Age, Gender, Linguistic Background, National Origin
How often is dataset reviewed?	Before Every Release
Is a mechanism in place to honor data subject right of access or deletion of personal data?	No
If PII collected for the development of the model, was it collected directly by NVIDIA?	PII not collected for development of model.
If PII collected for the development of the model by NVIDIA, do you maintain or have access to disclosures made to data subjects?	Not applicable
If PII collected for the development of this AI model, was it minimized to only what was required?	Yes
Is data in dataset traceable?	Yes
Are we able to identify and trace source of dataset?	Yes
Does data labeling (annotation, metadata) comply with privacy laws?	Yes
Is data compliant with data subject requests for data correction or removal, if such a request was made?	The data is compliant where applicable, but is not applicable for all data.

Privacy Subcard



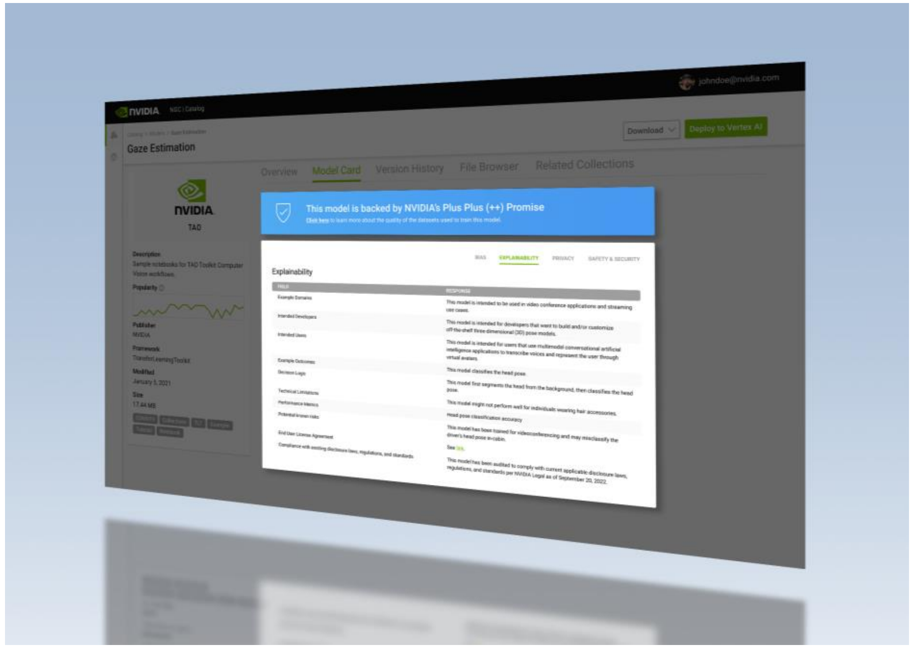
Safety & Security

FIELD	RESPONSE
Verified to have met prescribed quality standards:	Yes
Target Key Performance Indicator(s) (KPI(s)):	Accuracy, Latency, Throughput, Silence Robustness
Model Application(s):	Transcription
Describe the life-critical application (if present).	Not Applicable
Use Case Restrictions:	Abide by https://developer.nvidia.com/riva/ga/license
Explicit model and dataset restrictions:	Dataset access restrictions.
Describe access restrictions (if any):	Data is available to need-to-know internal NVIDIA employees only.

Safety & Security Subcard

Does Model Card++ help?

Inform and drive usage across domain and platform



Gaze Estimation

Description
Sample notebooks for TAO Toolkit Computer Vision workflows.

Popularity

Publisher
NVIDIA

Framework
TransferLearningToolkit

Modified
January 5, 2021

Size
17.44 MB

This model is backed by NVIDIA's Plus Plus (++) Promise

Click here to learn more about the quality of the datasets used to train this model.

Explainability

FIELD	RESPONSE
Example Domains	This model is intended to be used in video conference applications and streaming use cases.
Intended Developers	This model is intended for developers that want to build and/or customize off-the-shelf three dimensional (3D) pose models.
Intended Users	This model is intended for users that use multimodal conversational artificial intelligence applications to transcribe voices and represent the user through virtual avatars.
Example Outcomes	This model classifies the head pose.
Decision Logic	This model first segments the head from the background, then classifies the head pose.
Technical Limitations	This model might not perform well for individuals wearing hair accessories.
Performance Metrics	Head pose classification accuracy
Potential Known Risks	This model has been trained for videoconferencing and may misclassify the driver's head pose in-cabin.
End User License Agreement	See link.
Compliance with existing disclosure laws, regulations, and standards.	This model has been audited to comply with current applicable disclosure laws, regulations, and standards per NVIDIA Legal as of September 20, 2022.

Describe and clarify **"hard-to-find"** **"Top5"**:

- performance
- datasets
- terms of use/license
- intended applications and use cases
- known limitations and risks

Accelerates assessment & deployment, particularly for data scientists and management; **no need to read paper**

More descriptive model cards tend to have **higher usage** (Liang et al. 2024, 18) and are thereby **more discoverable**

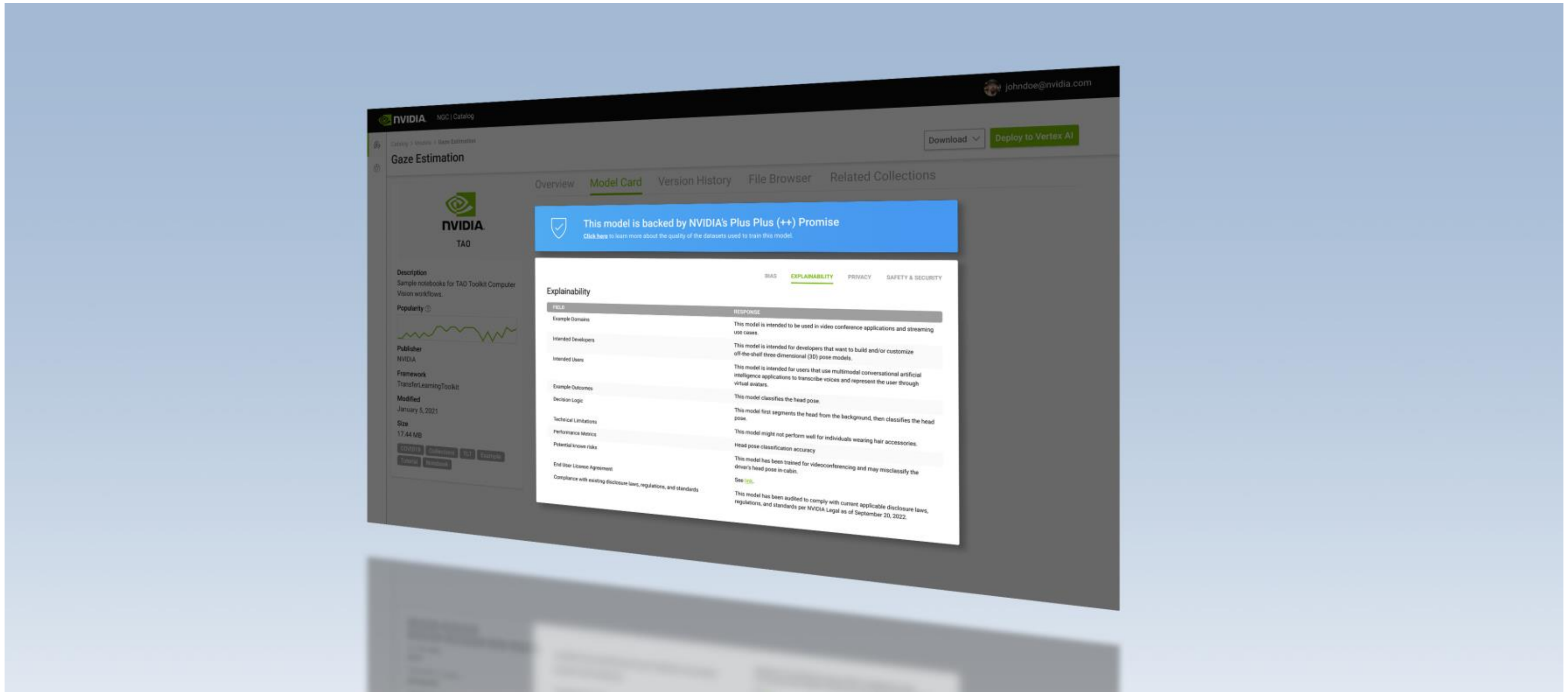
Model Card++ Demo

Mistral-NeMo Minitron 8B-8K Instruct

Small Language Model for Chatbot, Virtual Assistants, and Content Generation

[Model Card Link]

400+ Model Card++s Available Now



Gaze Estimation

Description
Sample notebooks for TAO Toolkit Computer Vision workflows.

Popularity

Publisher
NVIDIA

Framework
TransferLearningToolkit

Modified
January 5, 2021

Size
17.44 MB

This model is backed by NVIDIA's Plus Plus (++) Promise

Click here to learn more about the quality of the datasets used to train this model.

Explainability

FIELD	RESPONSE
Example Domains	This model is intended to be used in video conference applications and streaming use cases.
Intended Developers	This model is intended for developers that want to build and/or customize off-the-shelf three-dimensional (3D) pose models.
Intended Users	This model is intended for users that use multimodal conversational artificial intelligence applications to transcribe voices and represent the user through virtual avatars.
Example Outcomes	This model classifies the head pose.
Decision Logic	This model first segments the head from the background, then classifies the head pose.
Technical Limitations	This model might not perform well for individuals wearing hair accessories.
Performance Metrics	Head pose classification accuracy
Potential known risks	This model has been trained for videoconferencing and may misclassify the driver's head pose in-cabin.
End User License Agreement	See link.
Compliance with existing disclosure laws, regulations, and standards	This model has been audited to comply with current applicable disclosure laws, regulations, and standards per NVIDIA Legal as of September 20, 2022.

Automotive | Cybersecurity | Healthcare Manufacturing | Media & Entertainment | Retail Smart Cities | Telecommunications

Available wherever models are distributed externally; templates available here

Feedback

What could we do next?

- What resonates?
- Could a longer-term tiered or phased approach work?
- Could we prioritize fields of interest and get to a core set?

Questions?