What's this document
This is a hypothetical model card. We use a fictitious computer vision model in healthcare to give you a sense of how model cards document details about AI systems. Any similarity to actual persons, living or dead, or actual events or technology, is purely coincidental.

How to use it
The goal of model card templates is to help readers build a healthy mental model of your system so that can confidently say "If I do X with this model, Y will happen". Review this example of a model card to see how it supports a reader's decision.

How to write answers
When writing your answers, assume that the most important stakeholder reading this model card may not be wholly fluent in Machine Learning, Artificial Intelligence, or your Domain. Keep your answers simple and clear – and provide links where possible so readers can find more information as required by their use case.

Model Card Version: 0.0_YYYY
License: Apache 2.0

Model Name

Model: [Link to model]
Documentation: [Link to detailed documentation]
Model Card Authors: [Name and Contact]

Write a summary describing your model in a couple of sentences. Include information about the type of model and tasks, the motivation behind the models, and the problems or use cases it is suitable for. What is the benefit of using this model?

E.g. "Et-Track is a lightweight Deep Convolutional Neural Network model that classifies the presence and position of Endotracheal Tubes in chest radiographs. Et-track has achieved state-of-the art results with accurate attention explanations (CNN+CBAM) comparable to human decision-making . "

Model Snapshot

Model Overview

MODEL ARCHITECTURE	INPUT(S)	OUTPUT(S)

Describe the architecture of the model here.	*Provide a description (with necessary specifications) of the input data provided to the model for outputs.*	*Provide a description (with necessary specifications) of the output data from the model for given inputs.*
E.g. "TracheaNet_T Pre-trained on MedgraphNet image dataset fine-tuned for with data augmentation techniques used on training data. "	E.g. "grayscale chest radiographic images resized to a 256 × 256 matrix"	E.g. " **Presence/absence:** 0 to 1 confidence score **Low/Normal position:** 0 to 1 confidence scores, where normal is defined as the tube tip 3-7 cm above the carina. "

Usage

APPLICATION	BENEFITS	KNOWN CAVEATS
Where has this model been used, or where is it currently used? Include links for readers to learn more.	*Why might users choose to use this model, relative to others? Evidence your response with metrics or performance results*	*Are there any known and preventable failures about this model?*
E.g. "This model has or is being used in the following: • **Product:** VisNav-Track App used to monitor airways for ETT obstruction • **Production:** CareMed Tech's Radio-Track API (2020-2021) • **Research:** Advances in Ultrasonography at CareMed Tech. [Link to paper]"	E.g. "Et-Tack displayed significant improvement in AUC scores (0.989) over benchmark models, including: • PreTrained CardioNet (0.85) • Untrained CardioNet (0.625) • Untrained TracheaNet (0.604)"	E.g. "The model is trained on downsampled radiographic images in various orientations, with added noise and motion conditions and other augmentations. However: • Classification quality can degrade in extreme conditions • Classifiers are sensitive to physical obstructions in radiographic images. • Downsampling of training data may yield reduced accuracy"

Model Creators

MODEL CONTACT	MODEL AUTHOR(S)	CITATION
How can model owners be contacted for questions about the model?	*Write the names of all authors associated with the model. Provide the affiliation and year if different from publishing institutions or multiple affiliations, using the format Name, Title, Affiliation, YYYY:*	*If available, provide a citation to your model; else indicate unavailable.*
E.g. "Lea Kowalski, CareMed Research, leak@caremed.com" Group Email: E.g. "radio-ai@caremed.com" Website: "www.carened.com/radiology-AI"	E.g. " • Lea Kowalski, Staff Data Scientist, CareMed Research	E.g. "@inproceedings{krow2019endotracheal, title={Et-Track: Endotracheal tube detection and position chest radiographs using data augmentation},

	• Jules Zephania, Research Scientist, CareMed Research • Eric Alaniz, Doctoral Research Associate, Sidonia Institute of Technology”	author={Krowalksi, Lea and Zephania, Jules and Alaniz, Eric}, booktitle={Global Conference on Medical Images and Computer-Assisted Tracking}, pages={201--213}, year={2019}, organization={Spongers} }”

System Type

SYSTEM DESCRIPTION	UPSTREAM DEPENDENCIES	DOWNSTREAM DEPENDENCIES
Is this a standalone model, or intended to be used as part of a system with other models? Include links where necessary.	*If the model requires specific inputs, where should they come from? Are there any specific preprocessing steps that should be applied? Include links where necessary.*	*If the model’s outputs can be fed into another system, where should they go? Are there any specific post-processing steps that should be applied? Include links where necessary.*
E.g. “ **Standalone use:** Et-Track can be used as a standalone model for determining presence of ETT in chest X-rays. See detailed implementation for standalone use. **Sequential use:** Et-Track can be deployed as a part of a sequential system to enable airway monitoring. See detailed requirements for sequential configuration.”	E.g. “ **Standalone use:** Images should be cropped to a 256×256 matrix *or* square 1:1 aspect ratio **Sequential use:** Feature extraction endpoints from Track-Line pipeline are acceptable inputs to the model in sequential configuration.”	E.g. “ **Standalone use:** None known. **Sequential use:** Confidence scores from the model are converted into labels. Presence is classified at 0.97 threshold, and Normal is classified at 0.85 threshold. Labels and corresponding attention score are directly fed into the ActiveTrack model for airway monitoring within the Track-Line pipeline.”

Implementation Frameworks

HARDWARE & SOFTWARE FOR TRAINING	HARDWARE & SOFTWARE FOR DEPLOYMENT
Describe the hardware and software used for training the model.	*Describe the hardware and software used for deploying the model.*
E.g. “ • TensorFlow 1.0 • Trained on TPU v2 32 cores”	E.g. “ • TensorFlow 2.x • Deployed on TPU v2 2 cores”

Compute Requirements

COMPUTE REQUIREMENTS FOR FINE-TUNING*		COMPUTE REQUIREMENTS FOR INFERENCE*	
Describe the following compute requirements. Indicate unavailable if necessary. Do not delete any choices.		*Describe the following compute requirements. Indicate unavailable if necessary. Do not delete any choices.*	
Number of Chips	200	Number of Chips	100
Training Time (days)	7	Training Time (days)	2
Total Computation (floating pt operations)	2.91E+21	Total Computation (floating pt operations)	1.20E+5
Measured Performance (TFLOPS/s)	21.8	Measured Performance (TFLOPS/s)	10.1
Energy Consumption (MWh)	7.5	Energy Consumption (MWh)	0.5

**Modeled after Patterson, David, et al. "Carbon emissions and large neural network training." arXiv preprint arXiv:2104.10350 (2021).*

Model Characteristics

MODEL INITIALIZATION	MODEL STATUS	MODEL STATS
Describe how the model has been initialized. Include information about if the model trained from random initialization, or fine-tuned from a pre-trained model?	*Is the model static, or retraining on online data? If this model is trained and retrained, please include the update cadence, and the release date for the latest version.*	*What is the size of the model? Include attributes like number of weights and layers.*
E.g. " **Fine tuning or pre-trained TracheaNet:** The last layer set to random initialization of weights, and all other layers leared at a slower base learning rate (0.0001)."	E.g. "Static model trained on an offline dataset."	E.g. "This is a relatively small model, designed for on-device use."

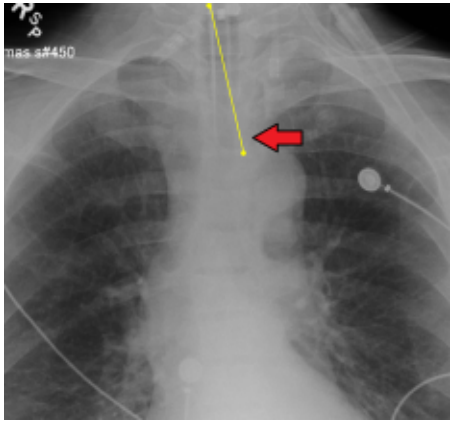
Training Epochs	90	**Dataset Name**	Et-Track(R)	**Size**	528 MB
Base Learning Rate	0.001	**Version**	2.0	**Weights**	138
Method	Stochastic Gradient Descent	**Release Date**	Unreleased	**Layers**	5 convolutional, 3 fully connected
Loss	Binary Cross-Entropy Loss using SoftMax	**Update Cadence**	Never	**Latency**	120/sec

PRUNING		QUANTIZATION		DIFFERENTIAL PRIVACY
Is your model pruned? If so, what is the level of sparsity of the deployed model?		*Is your model quantized? If so, what is the bit representation of the deployed model?*		*If any, describe the techniques implemented to preserve privacy?*
E.g. "TracheaNet_T was pruned and fine-tuned according to target-aware weight importance. In this way, we are able transfer a more suitable sub-structure applied during fine-tuning, improving final performance."		E.g. "We leveraged Half-precision floating point quantization to reduce the model's size by two times and decrease the inference latency.		E.g. "Et-Track uses differentially private stochastic gradient descent (DP-SGD) using the following hyperparameters: **L2 Normalized Clipping:** 1.2 to bind sensitivity to individual training points. **Random Noise Multiplier:** 1.8 to sample and add noise to gradients."
Methods	Magnitude Pruning	**Methods**	Half-Precision Floating-Point Quantization	
Structuring	Structured	**Pre-quantized Representation**	FP32 - int/float32	
Sparsity Level	50%	**End Bit Representation**	FP16 - int/float16	
Number of Params at Sparsity	12345	**Hardware**	Executed on GPU	
Accuracy at Final Sparsity after Training	80%			
Perplexity at Final Sparsity after Training	12%			

Data Overview

TRAINING DATASET SNAPSHOT	DATASET MAINTENANCE & VERSIONS	INSTRUMENTATION

Describe the dataset used to train the model. If a requested detail is inapplicable, following guidance on N/A. Include links to additional table(s) with more detailed breakdowns in the caption.		*Is the training data static, or updated/expanded? If so, what is the frequency with which this data is updated?*		*What instruments were used to collect or process the data? Describe any notable instrumentation requirements in the collection or preprocessing of data by customizing the table.*	
E.g. "HIPAA-compliant images obtained from our Picture Archive and Communications System (PACs), taken in posteroanterior (70%) and supine positions (30%). See Data Card for Et-Track(R)"		E.g. "The training data is collected yearly for the purposes of Radiology related AI Research at CareMed"		E.g. "All dataset images were captured using 28 x-ray machines at CareMed facilities. All machines had comparable optimized exposure settings that meet both state and CareMed Diagnostic Reference Levels (DRLs)."	
Dataset Size	123.45 MB	**Current Version**	2.0	**Instrumentation Criteria**	
Number of Instances	18000 [ETT presence] 18000 [ETT position]	**Update Cadence for Online Data**	Yearly	**Focal spot size**	1.8mm
Number of Fields	60024*	**Sampling methods**	Random	**Cooling method**	Air
Labeled Classes	2 [present/ absent/ unknown] [low/ normal/ unknown]	**Validation methods**	~60% of the images were validated by eleven board-certified radiologists and four cardiothoracic radiologist	**Avg Adult Effective Dose (mSv)**	0.02
Number of Labels	~70,000	**Processing methods**	Images underwent random cropping (277×277) and horizontal flipping	**Operational voltage range**	0 kV to 150 kV
Average labels per instance	1.94	**Annotation methods**	Manual (from patient data)		
Missing Labels	~2000 labels were marked unknown. No missing labels.	**Additional Notes:** The following data augmentation techniques were used to increase the size of the training data 12-fold: CLAHE, perpendicular rotations, rotations of ±5° and affine transformations using SWIRL			
Additional Notes: E.g. * Each pixel is converted into a single field with R,G,B values. Therefore, the number of fields is 60024 per instance.					
DATA PRE-PROCESSING		DEMOGRAPHIC GROUPS		EVALUATION DATA	

Describe any augmentation methods used during pre-processing to attain the requisite format. Are there any criteria that data points must satisfy to be included in the training set?	*Does the data contain any **labeled** groups**, or attributes that suggest demographic group membership? Describe any demographic groups considered when assessing distributions in the data.*	*Describe any notable factors about your final test set, including your train/test/dev split, any notable differences between the collection protocols for training & test data.*
E.g. "Normalization by ETT top-to-tip (TTT) distance is applied to unify the scale of the samples and is taken as 80%. CTT is calculated as the distance between coordinates of the entry point (top) of the ETT in the radiograph, to the murphy's eye (tip) on the distal tip. We do ignore 3D distance or length of the tube."  Above: Demonstrative image of TTT distance on an image sourced from wikimedia commons.	E.g. "To ensure broad geographic representation we include patient representation from the following regions: • Northern Africa (8%) • Eastern Africa (7%) • Middle Africa (7%) • Central Asia (6%) • Eastern Asia (6%) • South-eastern Asia (4%) All patients are adults (greater than 21 years of age at the time of the X-Ray taken) and have provided consent. We include the following self-reported sex characteristics distribution: • Male (42%) • Female (38%) • Intersex (16%) • Prefer not to say (4%)	E.g. " 6000 (20%) cases were held out for test The evaluation and tests sets used to report model performance **do not** contain any data augmentation" **Absence/Presence** Training/Eval: 60/20 **Low/Normal Position** Training/Eval: 60/20

***If there are groups that may be present, but are **not labeled** in the training data, please note this in the Ethical Considerations section below.*

Evaluation Results

Aggregate Evaluation Results

Document your aggregate or overall model performance evaluation.

EVALUATION PROCESS	EVALUATION RESULTS

Describe any notable factors in your process for evaluating your model's overall performance.	*Summarize and link to evaluation results for this analysis.*
E.g. " **Metrics:** Model performance for both predictions is measured using Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC). **Evaluation Set:** We use a randomly sampled split of Et-Track(R) as our test (20%) and validation (20%) datasets. No augmentation was performed on these. **Process:** We use pROC package (v1.7.3) to calculate performance using AUC-ROC and 95% confidence intervals on a holdout validation set for each model. We compared AUC results with and without data augmentation techniques using DeLong non-parametric method to determine significance."	E.g. "Model results compare models trained with and without dataset augmentation. Data augmentation significantly improved AUC scores. See the complete evaluation results." **With Augmentation** / **AUC** / **95% Confidence Interval** ET Tube Presence/Absence / 0.989 / 0.970 - 1.000 ET Tube Low/Normal / 0.809 / 0.697 - 0.921 **Without Augmentation** / **AUC** / **95% Confidence Interval** ET Tube Presence/Absence / 0.491 / 0.334 - 0.642 ET Tube Low/Normal / 0.657 / 0.513 - 0.800 **Significance** / **P value** ET Tube Presence/Absence / 0.216 ET Tube Low/Normal / 0.328 Note: *P value*> 0.05 is considered statistically significant

Subgroup Evaluation Results

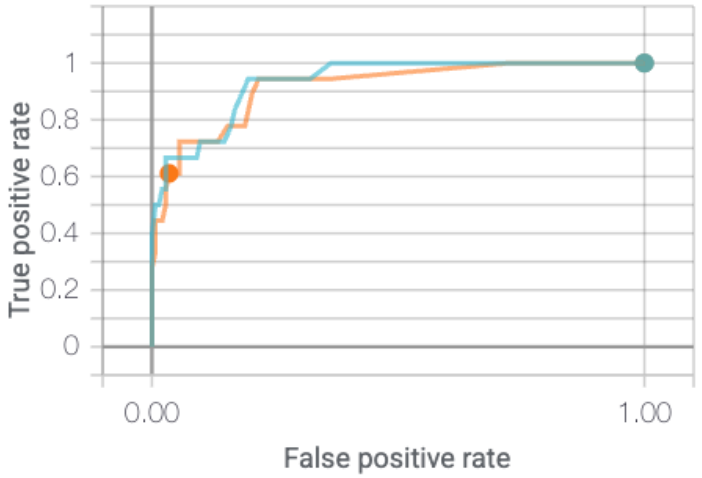
Document your disaggregated (e.g. fairness) evaluation. Duplicate this section (subgroup, evaluation process and data, evaluation results) for each subgroup evaluated.

SUBGROUP EVALUATED	EVALUATION PROCESS & DATA	EVALUATION RESULTS
Which subgroup was evaluated?	*Describe any notable factors in your process for disaggregated or sliced evaluation of model performance. Please include any assumptions made when disaggregating the data.*	*Are there any known and preventable failures about this model?*
E.g. " **Self-Reported Sex Characteristics** Sex characteristics were self-reported by patients whose x-rays were included in the validation set. "	E.g. " **Characteristic** / **Count** Male / 2520 Female / 2280 Intersex / 960 Prefer not to say / 240 Performance results are shown for categories that have more than 400 instances in the	E.g. "See the complete evaluation results. **Male (2520)** / **AUC** Presence/Absence / 0.93 Low/Normal / 0.92

validation dataset.

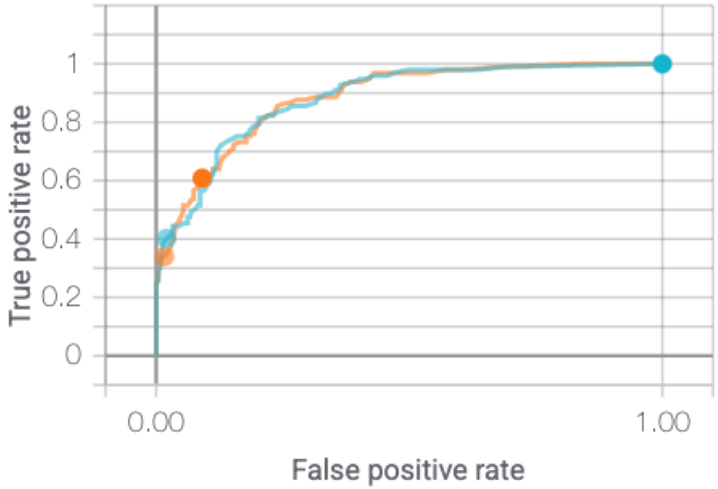
> **Note:** Sex characteristics refer to biological and physiological characteristics, and should not be conflated with gender identity. Further, these attributes do not indicate a medical diagnosis of gender dysphoria.

“

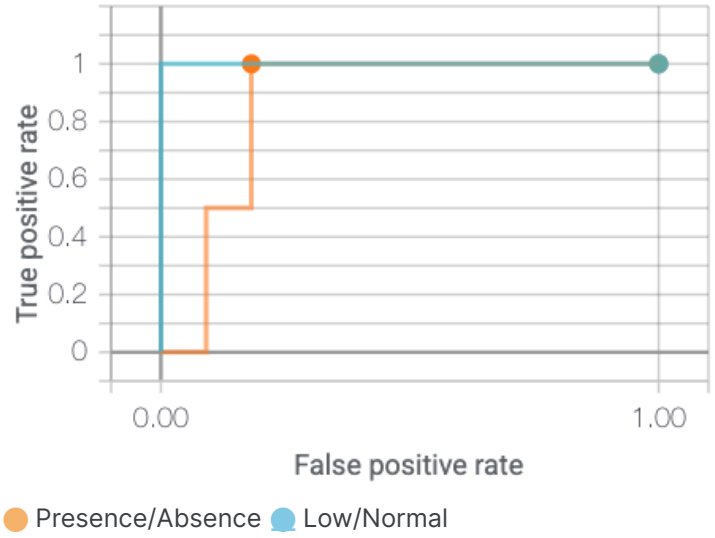


Presence/Absence Low/Normal

Female (2280)	**AUC**
Presence/Absence	0.88
Low/Normal	0.88

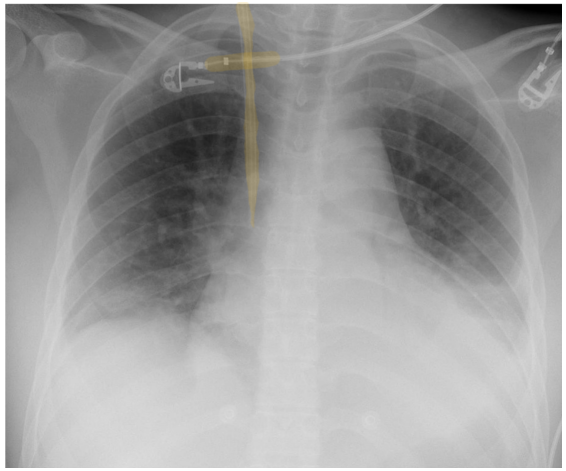


Presence/Absence Low/Normal

		Intersex (960) **AUC** Presence/Absence 1.00 Low/Normal 0.86 

Fairness Evaluation Results

FAIRNESS CRITERIA	FAIRNESS METRICS & BASELINE	FAIRNESS RESULTS
How did you define fairness? Describe the target fairness criteria you hoped to satisfy or optimize for before launch.	*Describe the metrics and the baseline for fairness against which you present your fairness results and how they are calculated.*	*Describe the results of your fairness analysis. Include any specific callouts or points that you would want to highlight for readers.*
E.g. “ **Et-track’s presence/absence predictions:** Fairness criteria are defined as predictions being equally accurate for patients in all subgroups defined under Self-Reported Sex Characteristics measured by equal positive predictive parity or predictive parity.”	E.g. “ **Et-track’s presence/absence predictions:** We compare sub-group performance to overall model performance with data augmentation techniques using sensitivity(%), specificity (%), PPV (%), and NPV(%), and we primarily focus on PPV. “	E.g. “ **Et-track’s presence/absence predictions:** Lower accuracy on PPV means alerts are less informative, and any predefined protocols may be less effective for patients in the subgroup; and can be remediated with additional clinician vigilance. See detailed recommendations. **Subgroup** **Sensitivity** **Specificity** **PPV** **NPV** Male 96% 7% 75% 10% Female 92% 12% 74% 12% Intersex 47% 42% 63% 18% ”

E.g. “ **Et-track’s low/normal predictions:** Fairness criteria are defined as the difference between CNN-CBAM attention salience and human annotation discrepancy measured in percentage of pixels (%px) *not important to prediction*. “	E.g. “ **Et-track’s low/normal predictions:** 18.76%px was obtained by measuring the average count of pixels that were of differing importance between human annotators (same as validators of the training set) and machine predictions on 120 samples. ”	E.g. “ **Et-track’s low/normal predictions:** In this table, we report on the %pixels that were important to annotators (I_A) and important to model (I_M) as a percentage of the entire radiograph. We then report on %pixels important to both annotator and model (IoU) , and not important to either annotator and model (Delta) as a percentage of all selected pixels (I_A + I_M).”
E.g. “  *Above:* Example of human annotation of important pixels, with ±5% error margin. Image source wikimedia commons.”	E.g. “  Above: Example of CNN-CBAM overlay. Yellow indicates pixels with importance >0.6 threshold to Et-Track. Image source wikimedia commons.”	E.g. “ **Subgroup** / **I_A** / **I_M** / **IoU** / **Delta** Male / 7% / 5.2% / 95 / 5% Female / 12% / 6.3% / 96 / 4% Intersex / 10% / 6.2% / 94 / 6% Reason for difference: We account for a 5% error margin since human annotators’s cognitive processing models are significantly different from computational models applied to measure salience. These results and follow-up studies show that Et-track learns features that are of importance to clinicians making the same prediction given a radiology image. Additional analysis here. “

Model Usage & Limitations

SENSITIVE USE	LIMITATIONS	ETHICAL CONSIDERATIONS & RISKS
Are there any use cases where deployment of this model would be considered sensitive?	*What factors might limit the performance of the model? What conditions must be satisfied to use the model?*	*What ethical factors did the model developers consider? Were any risks identified? What mitigations or remediates were undertaken? Where possible, link to additional documents.*

E.g. “ **Application:** This model is not a substitute for trained medical services. Using this model’s outcomes to confirm the position of ET Tubes *instead of* supporting clinical decisions is considered sensitive and should be subject to further scrutiny. **Pre-requisite training:** Any radiologists and medical staff who directly use the outputs of this model in standalone mode should receive onboarding or training from a qualified CareTech Relationship Associate. “	E.g. “ **Input conditions:** The model may not perform as well on low-contrast radiographs, radiographs with annotations, motion blurs, or with other possible real-world conditions. **Caveats:** The model does not predict tube types, and will therefore perform at the same efficacy for Endotracheal (ET) and Nasogastric (NG) tubes.”	E.g. “ **Research & Development:** Notably, • Subject matter experts in radiology, cardiothoracic imaging, and clinical practice were consulted throughout the Machine Learning (ML) model development. • Protected classes were designed around sex characteristics to mitigate Cohort Bias. • Model was trained accounting for Equal Performance and Informativeness Bias. • We collect yearly feedback from clinicians and patients See the entire TRIPOD report here. **Deployment:** Clinician interaction biases such as Automation Bias and Dismissal Bias are additional factors to consider when deploying Et-track in clinical care settings. We recommend training clinical staff before deploying Et-Track.”

Terms of Art

Concepts and Definitions referenced in this Model Card

✏️ *Use this space to include the expansions and definitions of any acronyms, concepts, or terms of art used across the Model Card. Use standard definitions where possible (e.g. MLCC Glossary). Include the source of the definition where indicated. If you are using an interpretation, adaptation, or modification of the standard definition for the purposes of your Model Card or model, include your interpretation as well.*

[TERM OF ART]	[TERM OF ART]	[TERM OF ART]
Definition: <write here> Source: <write and link> Interpretation: <write here>	Definition: <write here> Source: <write and link> Interpretation: <write here>	Definition: <write here> Source: <write and link> Interpretation: <write here>
[TERM OF ART]	[TERM OF ART]	[TERM OF ART]

Definition: <write here> Source: <write and link> Interpretation: <write here>	Definition: <write here> Source: <write and link> Interpretation: <write here>	Definition: <write here> Source: <write and link> Interpretation: <write here>
[TERM OF ART]	[TERM OF ART]	[TERM OF ART]
Definition: <write here> Source: <write and link> Interpretation: <write here>	Definition: <write here> Source: <write and link> Interpretation: <write here>	Definition: <write here> Source: <write and link> Interpretation: <write here>
[TERM OF ART]	[TERM OF ART]	[TERM OF ART]
Definition: <write here> Source: <write and link> Interpretation: <write here>	Definition: <write here> Source: <write and link> Interpretation: <write here>	Definition: <write here> Source: <write and link> Interpretation: <write here>

Reflections on the Model

✏️ *Use this space to include any additional information about the model that has not been captured by the Model Card.*

[Title]	Write notes here
[Title]	Write notes here
[Title]	Write notes here