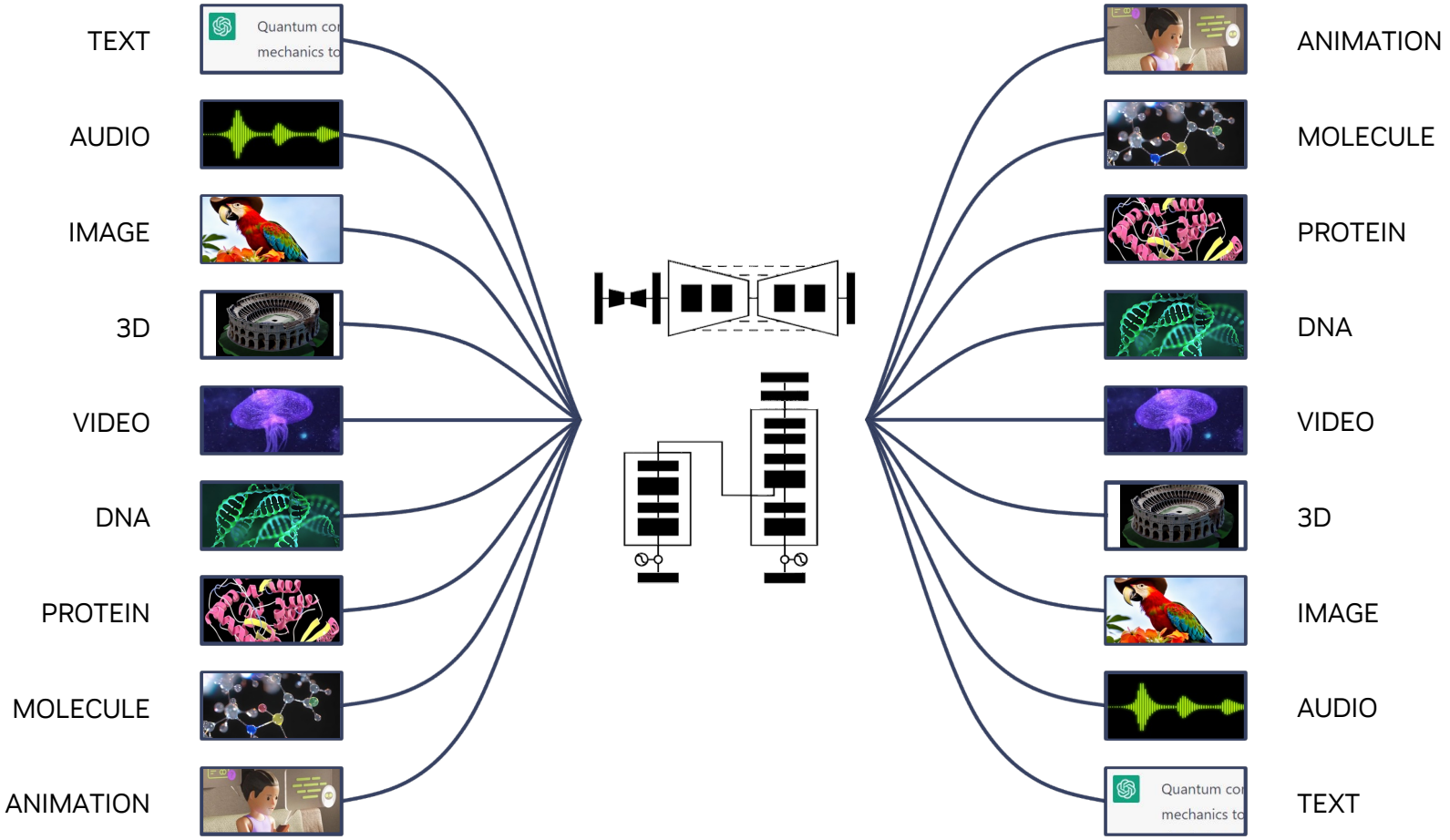
Energy-Efficient Generative AI: History, Challenges, and Mitigating Strategies

Scott McClellan

What is Generative AI?

Multi-Modal Creative Intelligence



The Arrival of Generative AI

For some, the future looks frightening…

- Fear for jobs, meaningful roles for humans in a future populated by talented and productive AIs
- Safety, bias, societal impact concerns
- “Paper clip anxiety” over hypothetical existential risks of superhuman AGI
- Concern AI advantages will accrue to privileged few or only giant companies, increasing inequity and reducing competition
- Worst-case extrapolations of energy use and environmental impact

The Arrival of Generative AI

For some, the future looks frightening…

Outside the scope of this talk…

- Fear for jobs, meaningful roles for humans in a future populated by talented and productive AIs
- Safety, bias, societal impact concerns
- “Paper clip anxiety” over hypothetical existential risks of superhuman AGI
- Concern AI advantages will accrue to privileged few or only giant companies, increasing inequity and reducing competition

- Worst-case extrapolations of energy use and environmental impact

Focus of today's discussion

Energy Use and Environmental Impact

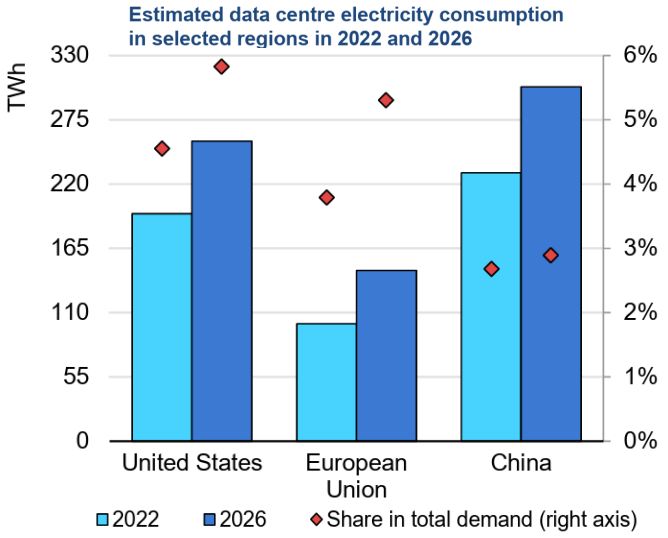
Common Modeling Errors Behind the Scary Headlines

- Overly simplistic assumptions and methodology
  - Incorrect initial conditions, e.g., actual scale of data center energy use
  - Unfair/inaccurate/missing perspective or comparison with other energy-consuming activities
  - Not accounting for closed-loop incentives driving efficiency
  - Failure to differentiate energy sources, actual efficiency of infrastructure and datacenter operations
  - Ignoring arc of AI technology and architecture improvement
  - Ignoring physical and practical limitations to rate and scale of growth
  - Poor understanding of actual Generative AI lifecycle and contributing factors to overall energy use
  - Focus on only largest, most expensive Generative AI models
  - Extrapolating from atypical, unrealistic, most-extreme deployment scenarios
- Result:
  - Propagation of (sometimes wildly) incorrect conclusions and implications of impending disaster
  - Unhelpful and inaccurate guidance for anyone wishing to employ Generative AI

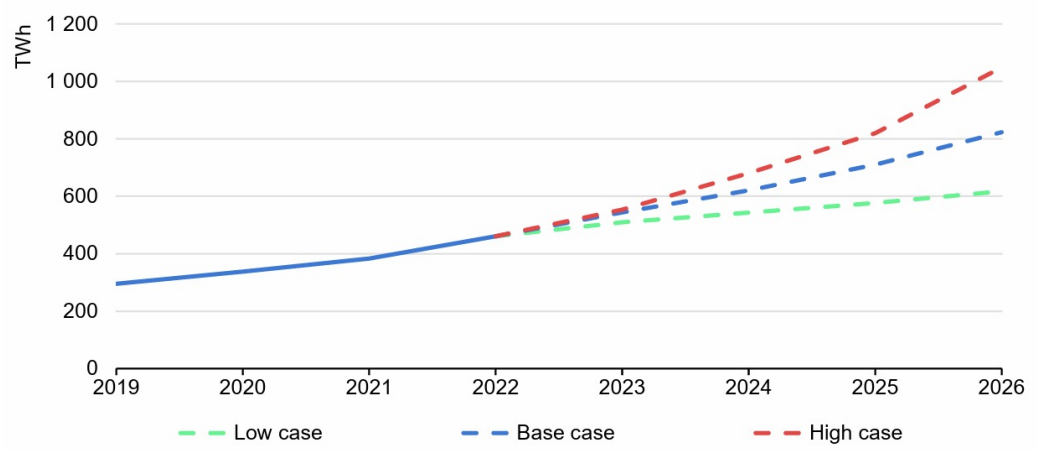
Modern Datacenter Energy Use

Current reality, projections, and perspective

- Datacenters, cryptocurrencies, and artificial intelligence (AI) consumed about 460 TWh of electricity worldwide in 2022
  - Almost 2% of total global electricity demand
  - Could double by 2026
- Datacenter efficiency gains absorbed much of workload and demand growth from 2015 to 2022
  - Datacenter workloads increased by 340%, global internet traffic by 600%
  - Energy consumed by datacenters increased by only 20-70%



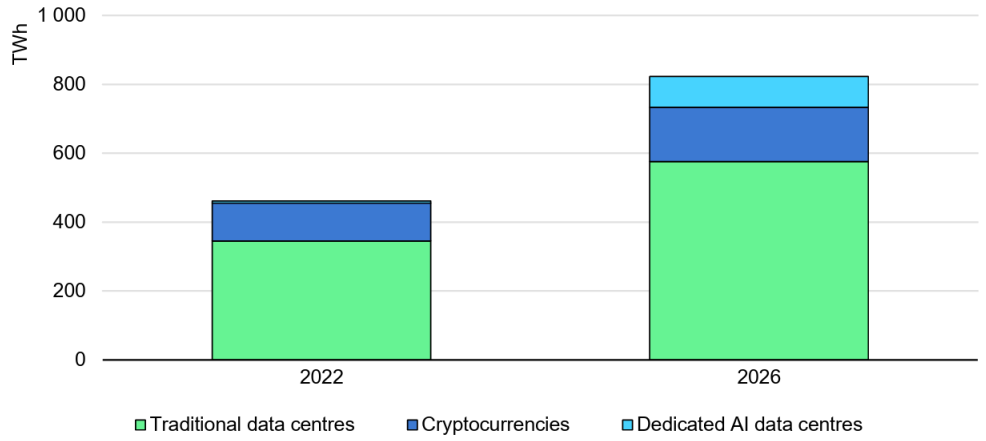
Global electricity demand from data centres, AI, and cryptocurrencies, 2019-2026



IEA. CC BY 4.0.

Notes: Includes traditional data centres, dedicated AI data centres, and cryptocurrency consumption; excludes demand from data transmission networks. The base case scenario has been used in the overall forecast in this report. Low and high case scenarios reflect the uncertainties in the pace of deployment and efficiency gains amid future technological developments.

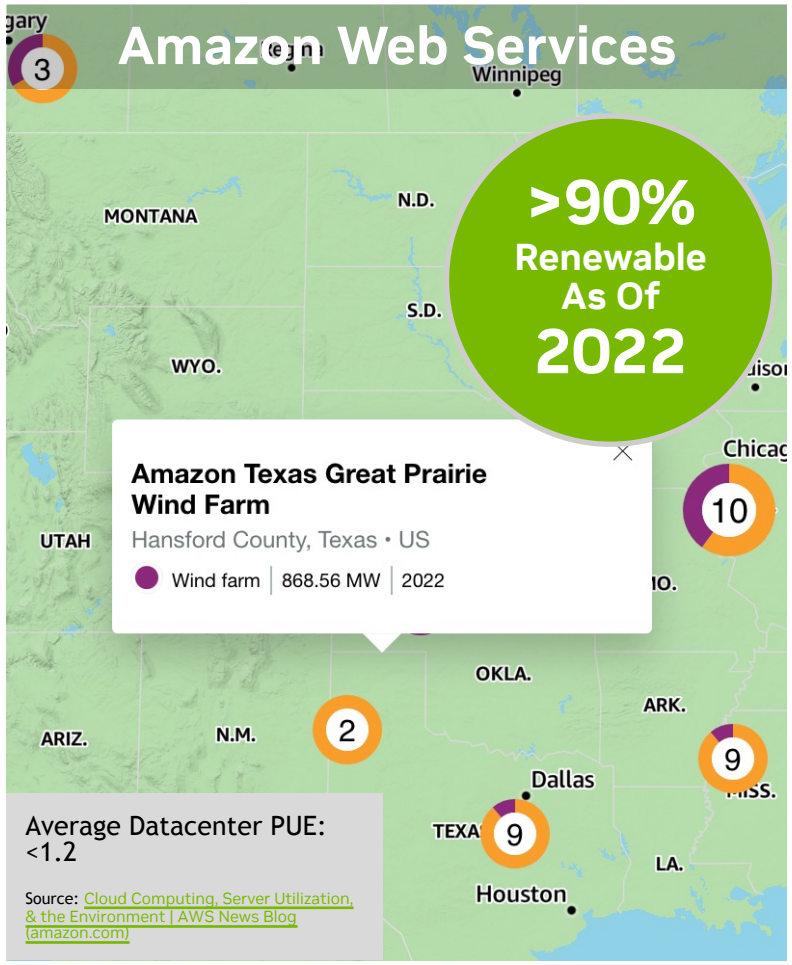
Estimated electricity demand from traditional data centres, dedicated AI data centres and cryptocurrencies, 2022 and 2026, base case

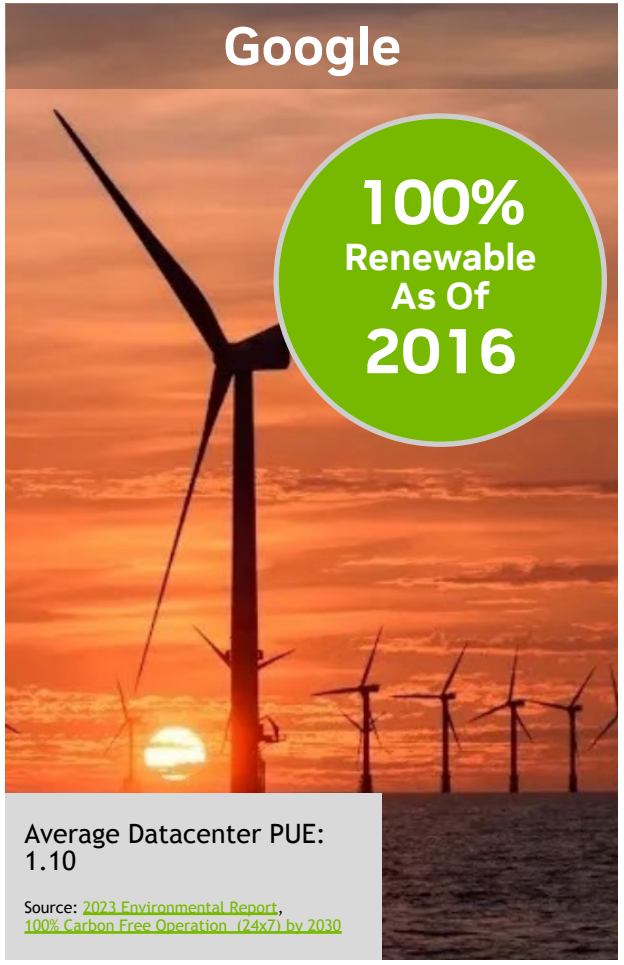


Sources: The Economist, "Data Centers Improved Greatly In Energy Efficiency As They Grew Massively Larger", 1/29/2024
International Energy Association, https://www.iea.org "Electricity 2024: Analysis and Forecast to 2026"

Cloud and Hyperscale Datacenters Leading the Way to Green Energy

Highest efficiency datacenters, massive investments in renewable energy sourcing and storage





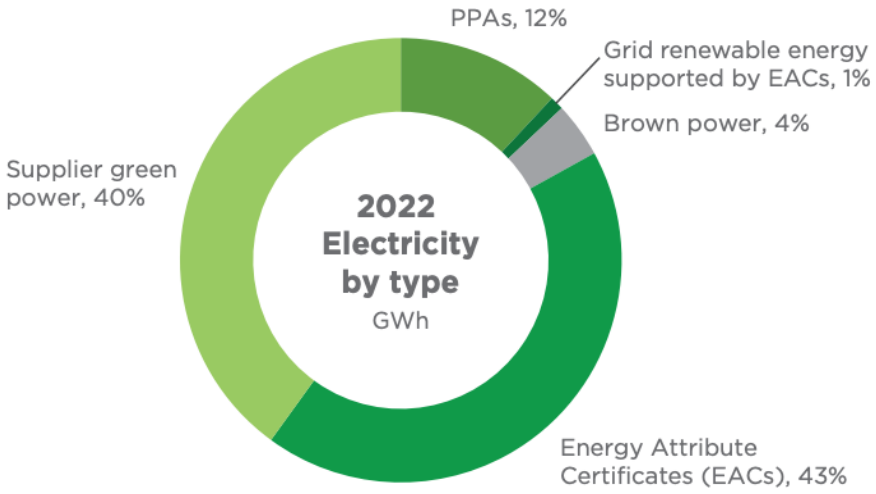
Business Rationale:

Supports rapid growth, resilience, and (most importantly) minimizes cost of energy and maximizes profits.

Non-Cloud Datacenter Progress

It us not just the CSPs that are focusing energy and environment! Example: Equinix

Renewable energy by type



$45M
invested in 2022 toward energy efficiency

Driving operational excellence and energy demand reduction across our global footprint

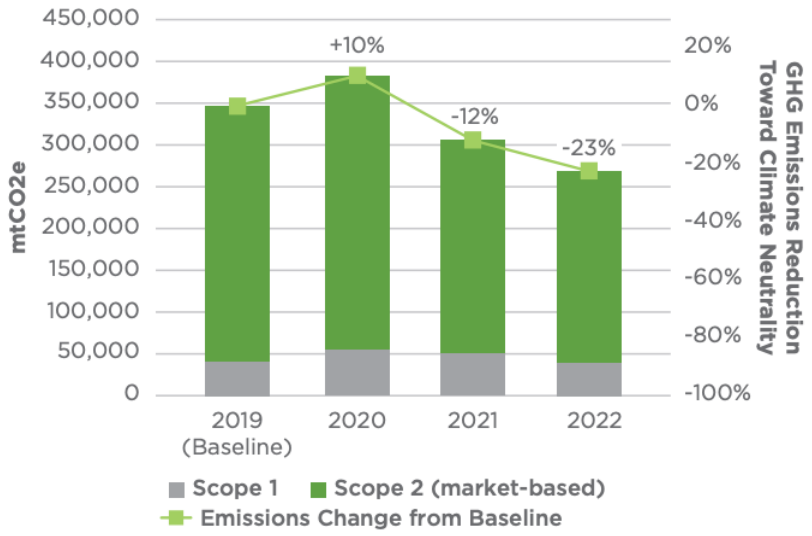
96%
renewable energy

Fifth year in a row with +90% global renewable energy coverage

1.46
average annual PUE

5.5% improvement in Power Usage Effectiveness from 2021

Operational emissions reduction toward climate neutrality

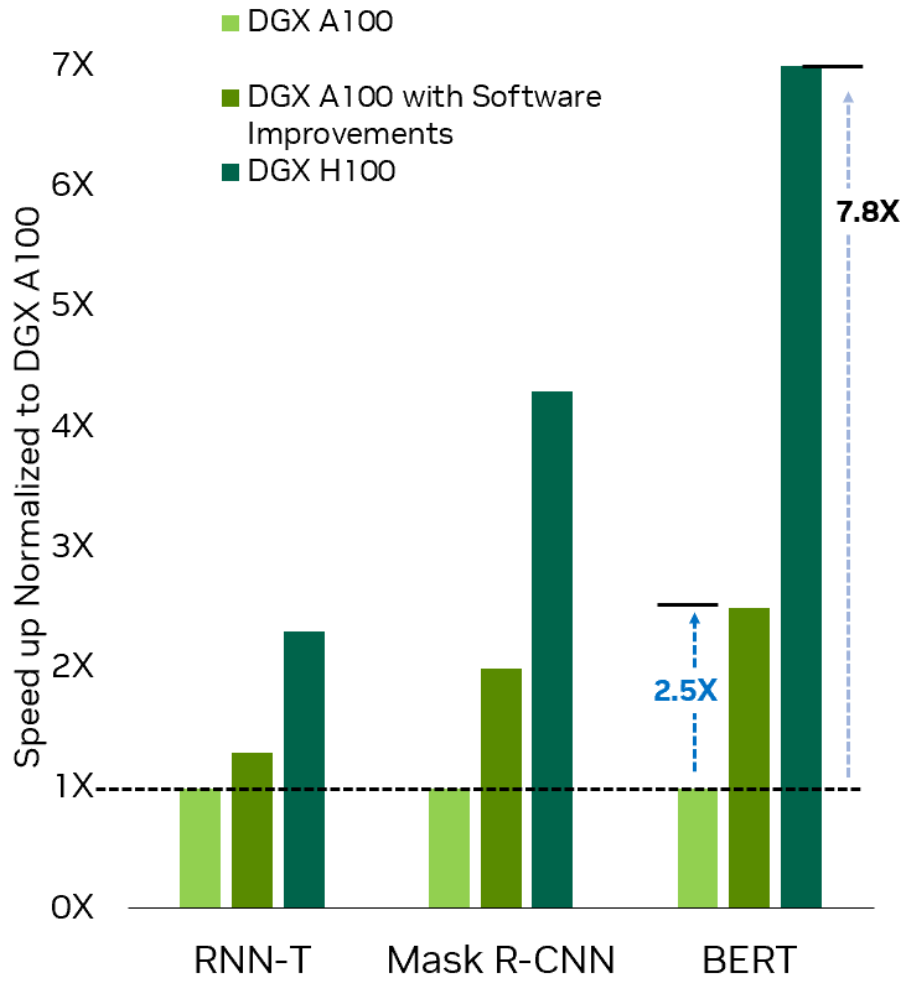


Source: Equinix 2022 Sustainability Report

The Arc of AI Computational Efficiency

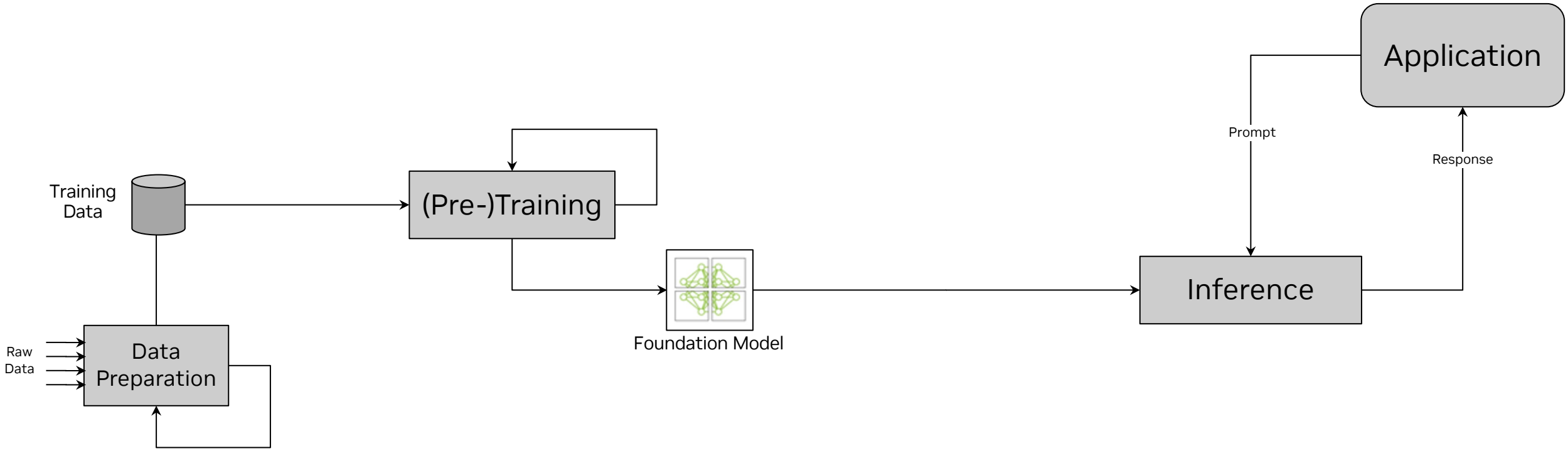
Energy Cost to Train a Model is Decreasing with Time

- Generation-over-generation HW advances
  - Mostly from architecture and packaging improvements, less from process
  - GPU examples: Tensor Cores, Transformer Engine, TMA, structured sparsity, quantization to lower precision and ints, etc.
  - Node examples: CPU-GPU C2C, EGM
  - System examples: Multi-node NVLink, Sharp, NDR, high-density LC packaging
- Continuous SW improvements
  - Same models, same HW, up to 2.5x faster (more efficient!)
  - Tools and management improvements
- Rapid AI algorithm and model architecture) improvements)
  - E.g., MoE, Transformer alternatives (Retention Network, Mamba, etc.
  - Better scaling, more efficient/faster training



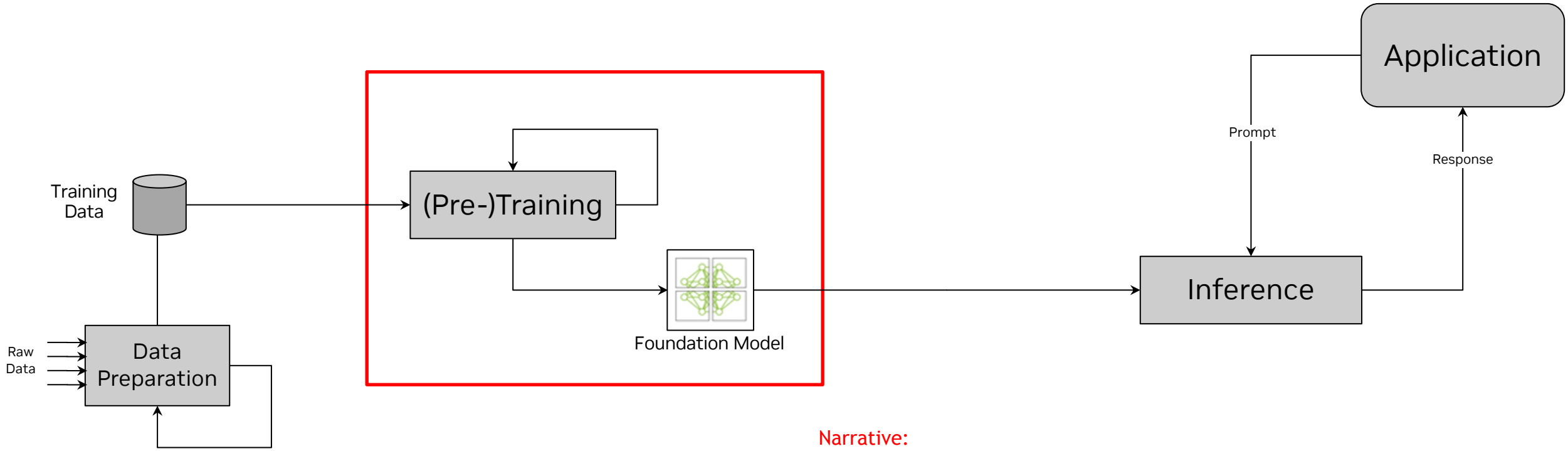
Simplest Generative AI Workflow

A starting point for the discussion…



Doomsday Projections – Focused On Training Cost

Amplified by this overly simplistic view of Generative AI



Build a simple model
GPT 4: 25000 GPUs = 3125 systems x 6.5kWh/system x 100 days =

50GWh

Narrative:

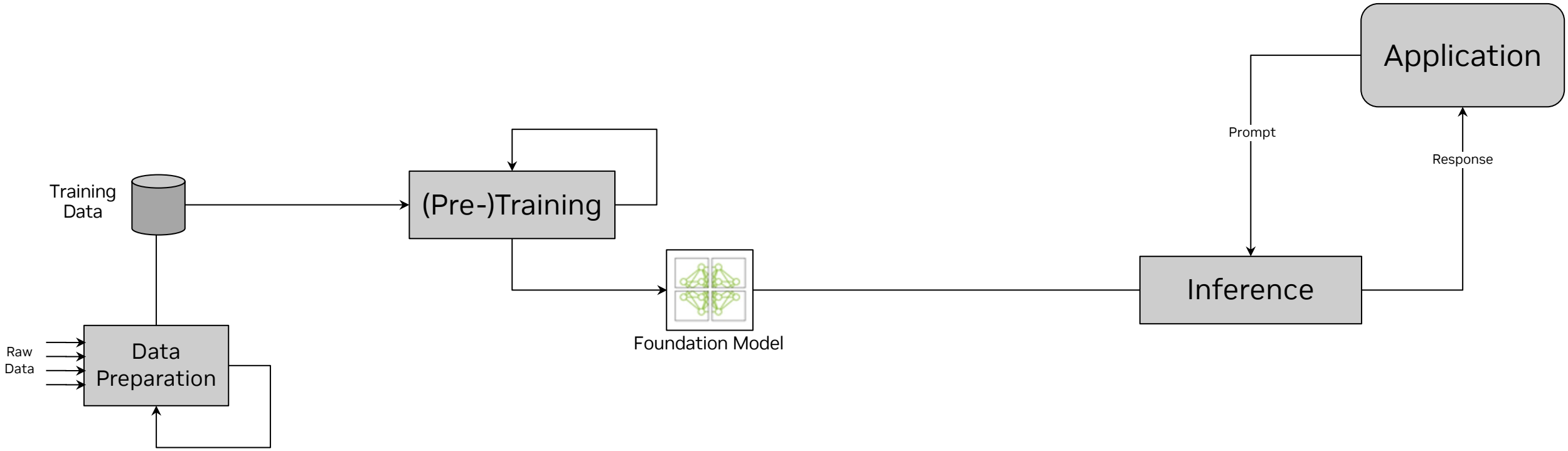
- 40x more than GPT3 (scary trend)
- 50GW would power 1000 US Households for 5-6 years (scary comparison)
- Under-estimated: Model only covers training (scary implication)
- Generative AI Adoption has “only just begun” (scary future)

??

Draw your own conclusion

Simplest Generative AI Workflow

A starting point for the discussion...



Too simplistic...

Applications generally do not use foundation models directly...

Developing a Generative AI Application

Data Preparation

Pre-processing raw data for use in the “next step” (training, fine-tuning, augmentation, etc.)

Model Development

Development of generative AI models including net-new models, refinement and lifecycle support and distribution for downstream use.

Model Customization

Fine-Tuning

Use of fine-tuning techniques to adapt a pre-trained model by adjusting its parameters on a smaller, task-specific dataset, enhancing its performance.

Augmentation

Use additional data to augment the output from a generative AI model. Also known as Retrieval Augmented Generation (RAG)

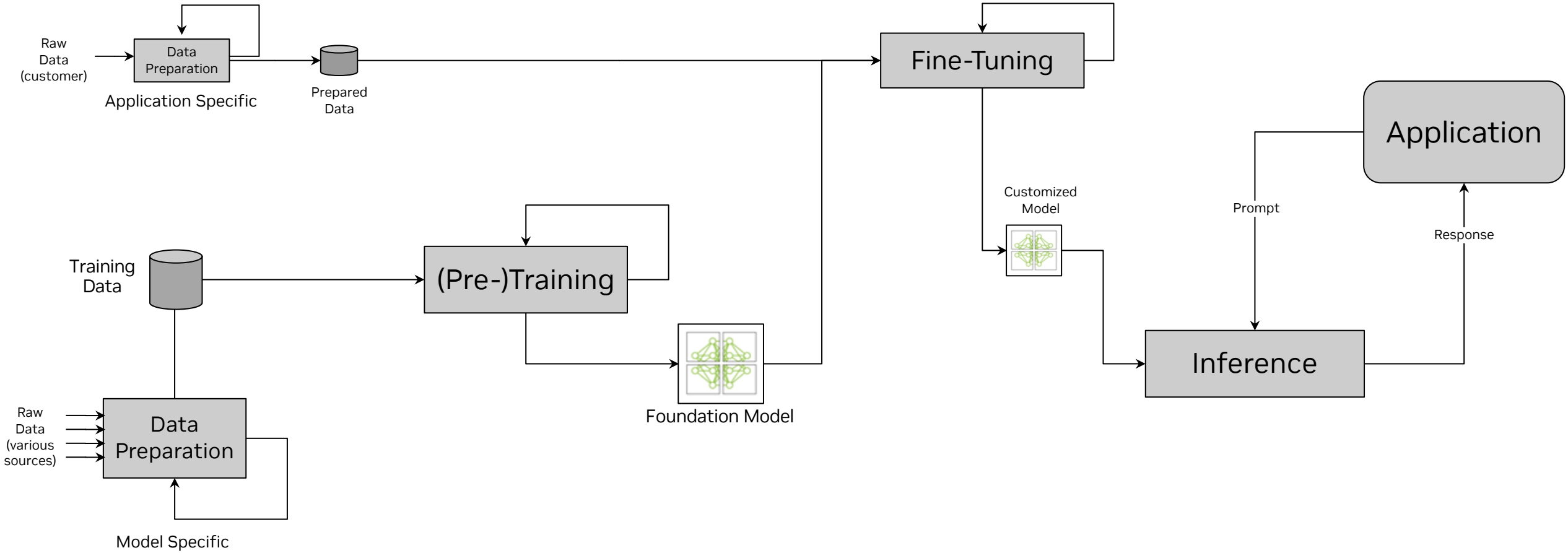
Optimization

Use of techniques like quantization (converting weights to lower precision) or distillation (transferring knowledge from a large model to a smaller one) to optimize a model for efficient use.

Model Use

Use of a model (or models) by one or more applications

Customization

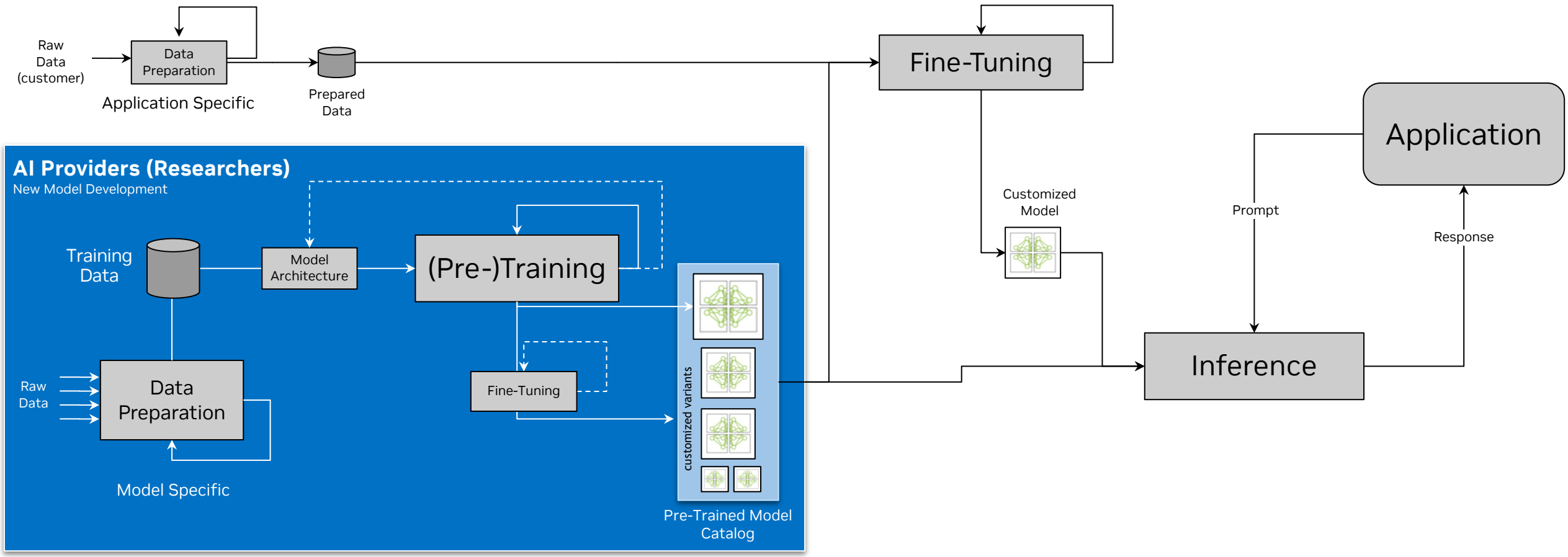


Still Too Simplistic

Need to separate the role of “model developer” from downstream users of pre-trained models.

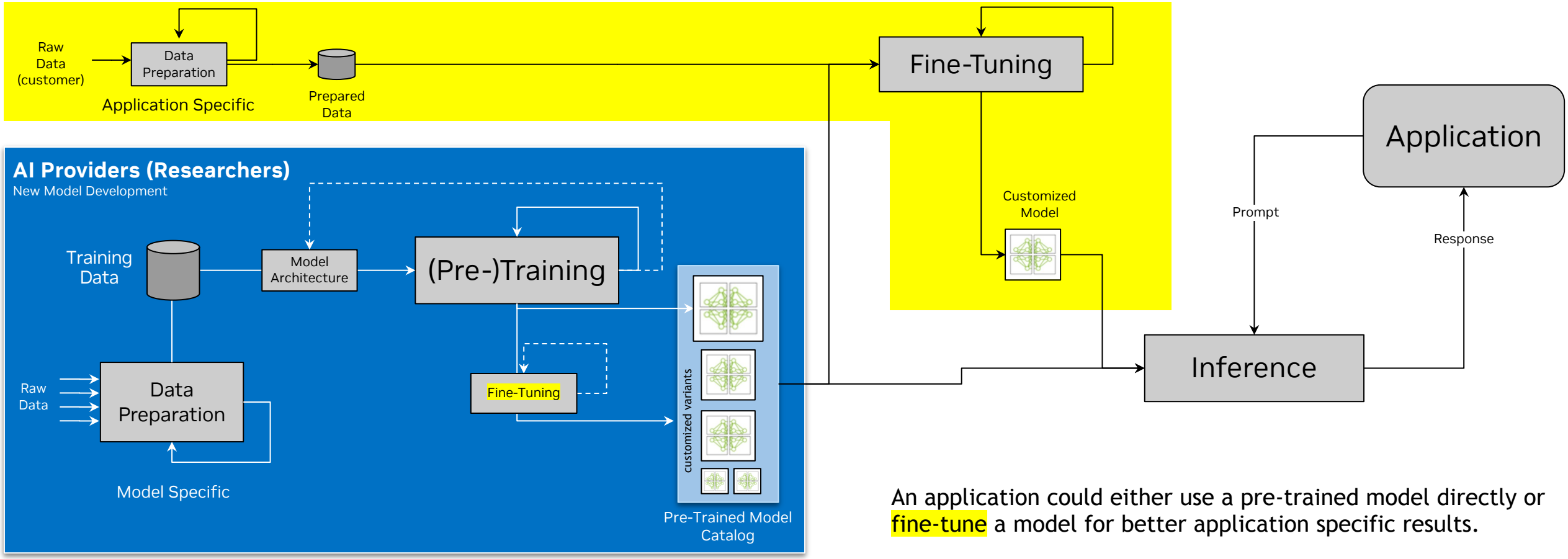
Separate Roles

Blue Box done by Model Developers (providers), Outside blue box done model user (customers)



Fine Tuning

Using additional data to fine-tune a model for a particular task



An application could either use a pre-trained model directly or fine-tune a model for better application specific results.

Note: fine-tuning could also happen upstream…

Fine-Tuning

Wide range of options for fine-tuning, ultimately using much less energy than training a model from scratch.

Fine Tuning:
Most or all of the parameters of the model are updated during the training process

Parameter-Efficient Fine-Tuning (PEFT):
Involves updating only a small subset of the model's parameters or adding a small number of trainable parameters to the model.

Full Model Fine-Tuning
Adjusting all the weights of the model on the new dataset.

Layer-wise Fine-Tuning
Gradually unfreezing layers of the model starting from the top (output) layers and potentially adjusting more layers as training progresses.

P-Tuning
P-Tuning takes Prompt Tuning further by introducing trainable embeddings, optimized to enhance the model's performance on specific tasks. P-Tuning involves learning these prompts (pattern templates) as part of the training process

Adapter Modules
Adding small trainable modules (adapters) between the layers of the pre-trained model while keeping the original weights frozen.

Prompt Tuning
Learning a set of embeddings (prompts) that are prepended to the input to steer the model's predictions without altering the underlying weights.

Low-Rank Adaptation:
Applying a low-rank factorization to modify a subset of the model's weights, impacting a small fraction of the parameters.

Pre Training

Uses More Energy

Uses Less Energy

Uses Least Energy

O(10-100)x + Savings vs Pre-Training

O(1000-10000)x + Savings vs Pre-Training

LLama2 Example

A more realistic/typical example of the energy cost for training a LLM and the power of re-using it downstream

	Reported			Calculated		
	Compute Time (GPU HRs)	Power Consumption (wH/GPU)	Carbon Emitted (tCO2eq)	Total Power (MWh)	MWh/ B Parameters	MWh/ T Tokens
Llama 2 7b	184,320	400	31.22	81.10	10.53	36.86
Llama 2 13b	368,640	400	62.44	162.20	11.34	73.73
Llama 2 34b	1,038,336	350	153.9	399.76	10.69	181.71
Llama 2 70b	1,720,320	400	291.42	756.94	9.83	344.06
total	3,311,616		539	1,400.00		
average	827,904		134.745	350.00		

Source: Llama 2: Open Foundation and Fine-Tuned Chat Models

1.4 GWh

4
pre-trained models

70 MWh

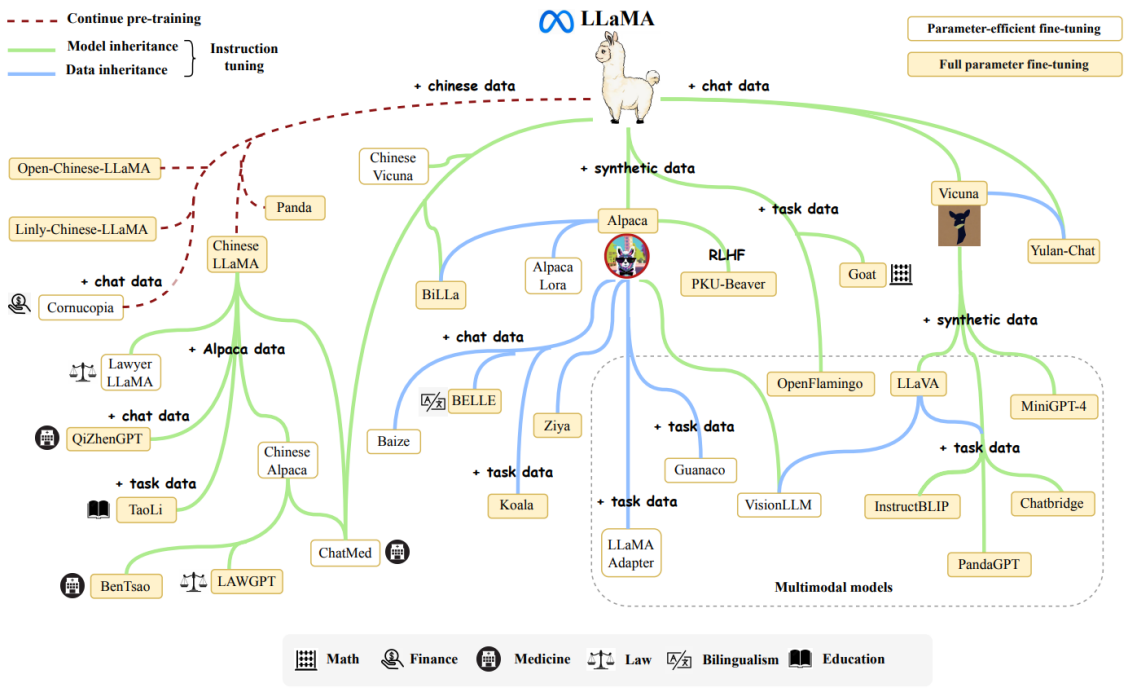
per 20
fine-tuned variants

350 MWh

per1000
PEFT tuned variants

O(7000) Llama Variants to date

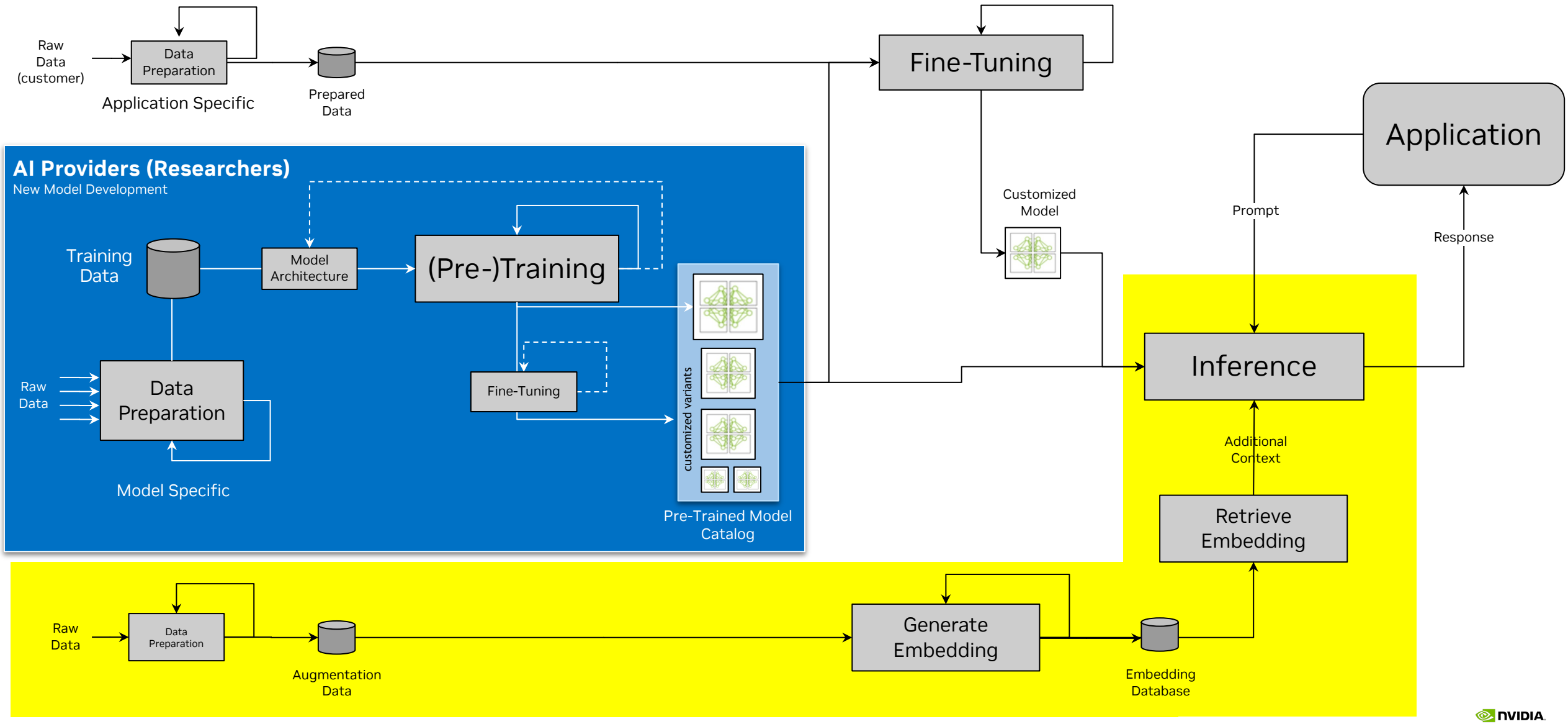
Released on Hugging Face: Link



Source: A Survey of Large Language Models

Adding Augmentation

aka: Retrieval Augmented Generation (RAG)



Retrieval Augment Generation

- Retrieval-Augmented Generation (RAG) combines the powers of retrieval (searching for relevant information) and generation (creating text) in machine learning models
- Objective: enable the model to produce more accurate, relevant, and contextually enriched responses by leveraging external information
- Useful in scenarios where the model needs to generate domain or application specific content that needs to stay updated with the latest information that may not be present in its original training data
- Effective method for Enterprises to use their own data to augment the model while maintaining full control of the data itself
- Like Fine-Tuning, RAG is energy efficient way to avoid training a generative AI mode from scratch, much more energy efficient than training, probably not as energy efficient as the Parameter Efficient Fine-Tuning techniques
- Will require some additional energy on the inference side; from the additional retrieval step, and small impact on throughput (potentially mitigated by caching)

Data Preparation

For (Pre-)Training A Generative AI Models

Data Collection

Data Collection: Collect data from various data sources.

Data needs to be ethically sourced insuring data collection complies with all relevant legal and privacy requirements.

For Pre-Training:
Raw data quantities are massive (10s TB - multiple PB)

Data comes from many sources (including crawling the internet, and potentially third party sources)

Data sourced by the model provider who has ultimate ethical and legal responsibility

Data Cleaning

Data Deduplication: Remove duplicate or near-duplicate data to prevent the model from being biased towards overrepresented data.

Data Annotation: Annotate data as needed. Model may need to understand specific entities or concepts, the data may need to be labeled accordingly.

Text Normalization: Converting text to consistent format. Examples; converting to lowercase, standardizing dates and numbers, and correcting misspellings, expanding abbreviations.

Tokenization: Split text into meaningful units, such as words, phrases, or symbols. Can be language or task specific.

Handle Sensitive Information: Identify sensitive information, remove or safeguard personally identifiable information (PII) or other sensitive content.

Data Augmentation (as needed) to enhance the diversity and volume of the training dataset.

Data Curation

Split Data: Divide into training, validation, and test data sets

Data Storage and Management: Store the prepared data (selecting storage solution based on scale, performance, security and cost considerations)

Ethical and Bias Consideration: Evaluate the dataset biases that could lead to unfair or unethical outcomes when used

Legal Compliance: Ensure compliance with all relevant laws and regulations, including copyright laws, data protection regulations, and any usage specific legal requirements

Data Preparation

For Fine-Tuning A Pre-Trained Generative AI Model

Data Collection

Data Collection: Gather additional (domain specific) data relevant to your specific task.

The quality and quantity of your data can significantly impact the model's performance.

For Fine-Tuning:
Modest quantities of data (10s GB - 10's TB)

Data is typically data that already belongs to the Enterprise (end-user) but needs some additional pre-processing to to use it for fine-tuning.

The main "governance" concern is protecting the privacy and intellectual property in the data.

Data Pre-Processing

Data Cleaning: Remove any irrelevant, duplicate, or corrupt data. Fixing typos, removing or correcting mislabeled examples, and handling missing values.

Data Annotation: For supervised learning tasks, ensure that your dataset is accurately labeled. Labels should be consistent, and follow a well-defined schema. This may involve manual labeling by experts.

Data Augmentation: Augmenting data to increase its diversity and volume (if necessary).

Tokenization: Convert text data into a format that the model expects. Typically using the same tokenizer that was used during the model's pre-training. It's crucial to maintain consistency with the pre-training tokenization process.

Sequence Length Adjustment: Adjust the length of your input sequences (truncating longer sequences or padding shorter ones to a fixed length).

Splitting the Dataset: Divide your dataset into training (for fine-tuning), validation (for hyperparameter tuning), and test (for evaluation) data sets.

Feature Engineering (if applicable): Extract features or perform additional preprocessing steps specific to your task's requirements.

Normalization (for non-text data): Ensure features are normalized or standardized to have a similar scale, to help model learn more effectively.

Data Loading and Batching: Implement data loaders to efficiently load and feed data into the model in batches (especially for large datasets).

Data Shuffling: Re-order the training data to prevent the model from learning any unintended patterns from the order of the data.

Handling Class Imbalance (if applicable): Address any class imbalances in dataset to prevent bias towards the majority class.

Data Preparation: Energy Implications

- Storing and processing data both take energy, but are relatively “cheap” compared to other steps in the Generative AI pipeline
- Data processing steps are generally not computationally intensive (the most computationally intensive steps can be GPU accelerated)
- For some models/steps, the datasets are fairly static (data preparation is infrequent)
- Many data processing steps can be done incrementally (only on new data)
- For customization and augmentation, Enterprise customers will (often) leverage existing data and data processing infrastructure

Data Prep For …	Done By Entity	Data Quantities	Infrastructure (range)	Trigger (Frequency x Duration)	VERY ROUGH Power Estimates
(Pre-) Training	AI (Model) Providers & Researchers	10’s TBs, multiple PBs	Small to Medium Cluster	When training/releasing a new model (~quarterly x ~day) < 200 node-hrs / quarter << 1000 node-hrs / year	<< 750MWh / yr per pre-trained mode
Fine-Tuning	End Customers	10’s GB, 10s TBs	Single node to Small Cluster	When significant/relevant new data is available (~monthly, few-hrs) ~5 node-hrs / month <<100 node-hrs / year	<<75 MWh / yr to fine-tune (per application) (similar for fine-tuning or augmentation)
Augmentation					

Training Energy

Many factors offset the “scary” implications found in most energy articles…

Key Considerations:

- Training a large, general purpose LLM, does take a significant amount of energy…
  - Training energy is roughly proportional to model size (within a model architecture) and training data set size.
  - Range is large (model sizes span 3 orders of magnitude!) – up to 50 GWh at largest scale (so far…)
  - Typical training energy cost is skewed toward the lower end of the range, MWh not GWh – O(1000x) lower than “hero models” in the news.
- Re-use of pre-trained models amortizes the initial training and up-stream fine-tuning energy across O(1000) – O(10000) applications
  - Customizing a pre-trained model is much more energy efficient, anywhere from 10x – 1000x cheaper than training from scratch

Inference Energy

- Inference energy is going to be larger than training energy over time
- Inference energy is proportional to model size
  - Small specialized models can be more accurate and O(100x) less energy than large general models
  - Quantization/distillation reduces model size
  - Model architecture improvements (e.g., MoE) reduce active memory size requirements
  - RAG increases skill w/o increasing model scale
- Model size affects number of GPUs required for inference -> directly relevant to energy
- Inference energy also function of demand
  - Throughput (number of users, queries/s, etc.), latency requirements, interaction/context length
- Energy also benefits from the arc of computational efficiency

Shrinking LLM Energy Use

Inference

- Many techniques slashing energy and cost while improving fitness
  - Pruning
  - Quantization (example: BitNet b1.58 vs. FP16)
  - Distillation
  - Model architecture improvements (e.g., MoE)

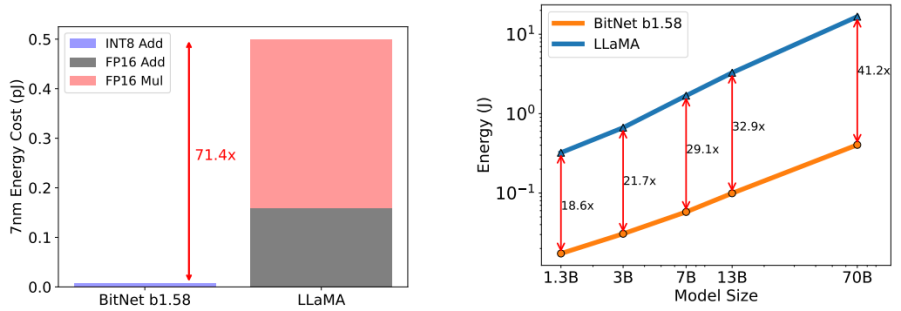
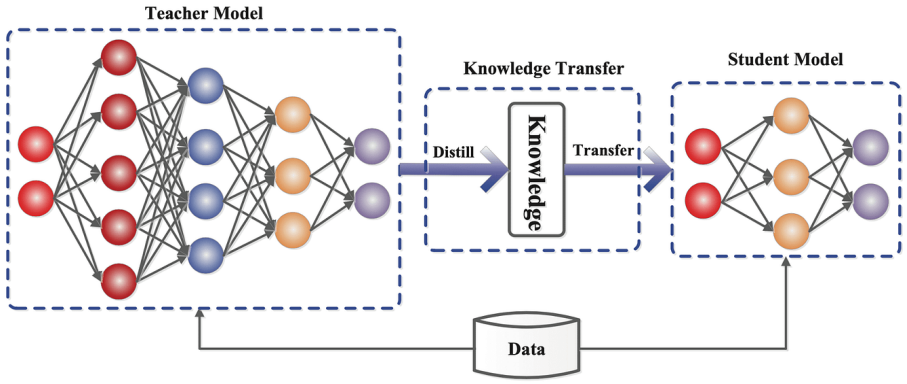
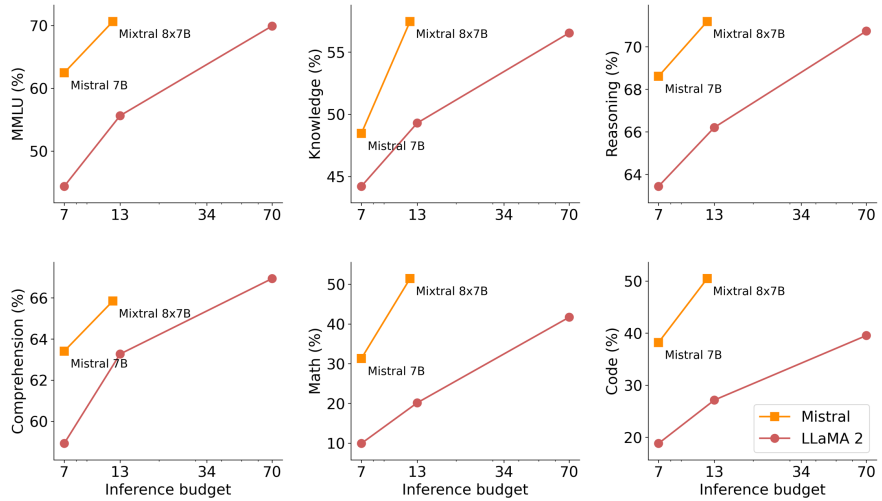


Figure 3: Energy consumption of BitNet b1.58 compared to LLaMA LLM at 7nm process nodes. On the left is the components of arithmetic operations energy. On the right is the end-to-end energy cost across different model sizes.

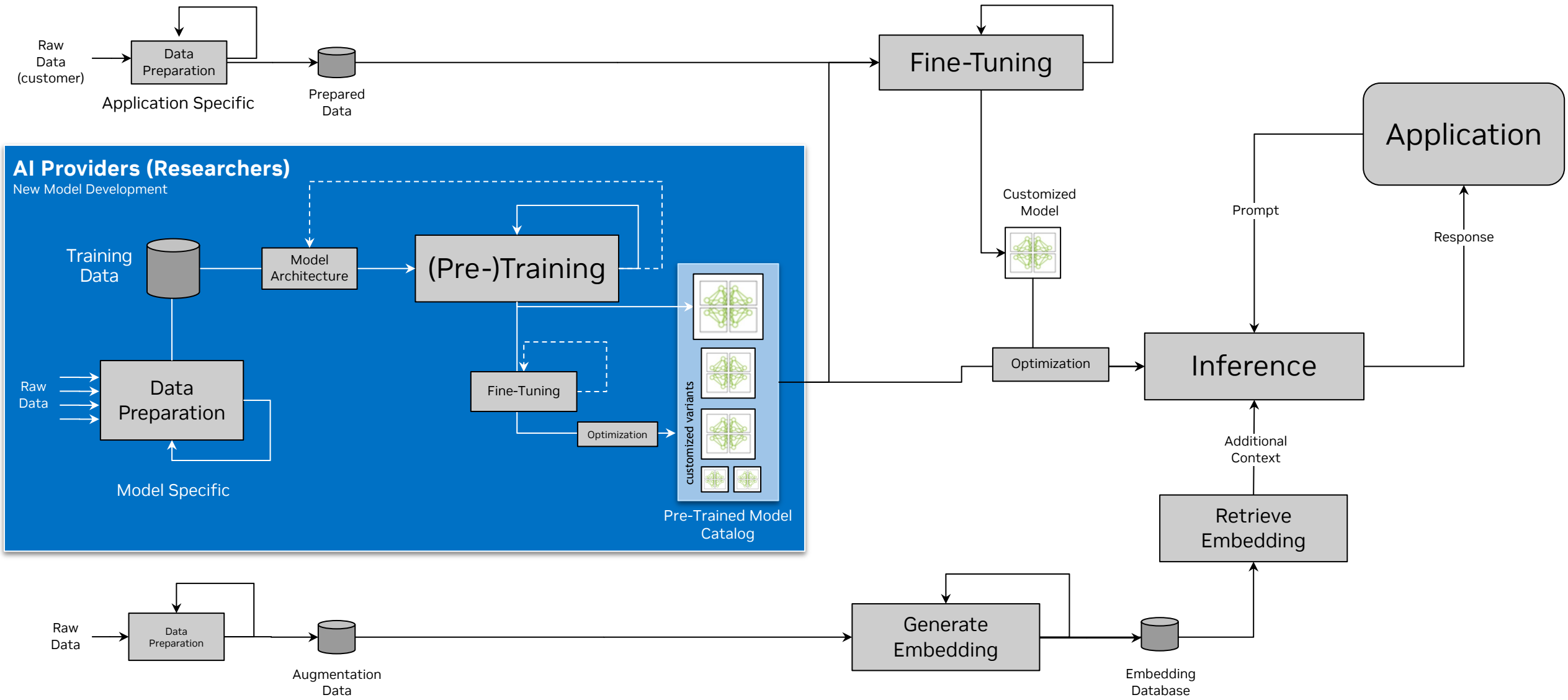
- Smaller and more specialized open source LLMs
  - Example: Mistral >300x less energy inference than GPT3 at similar in-domain accuracy





Optimization

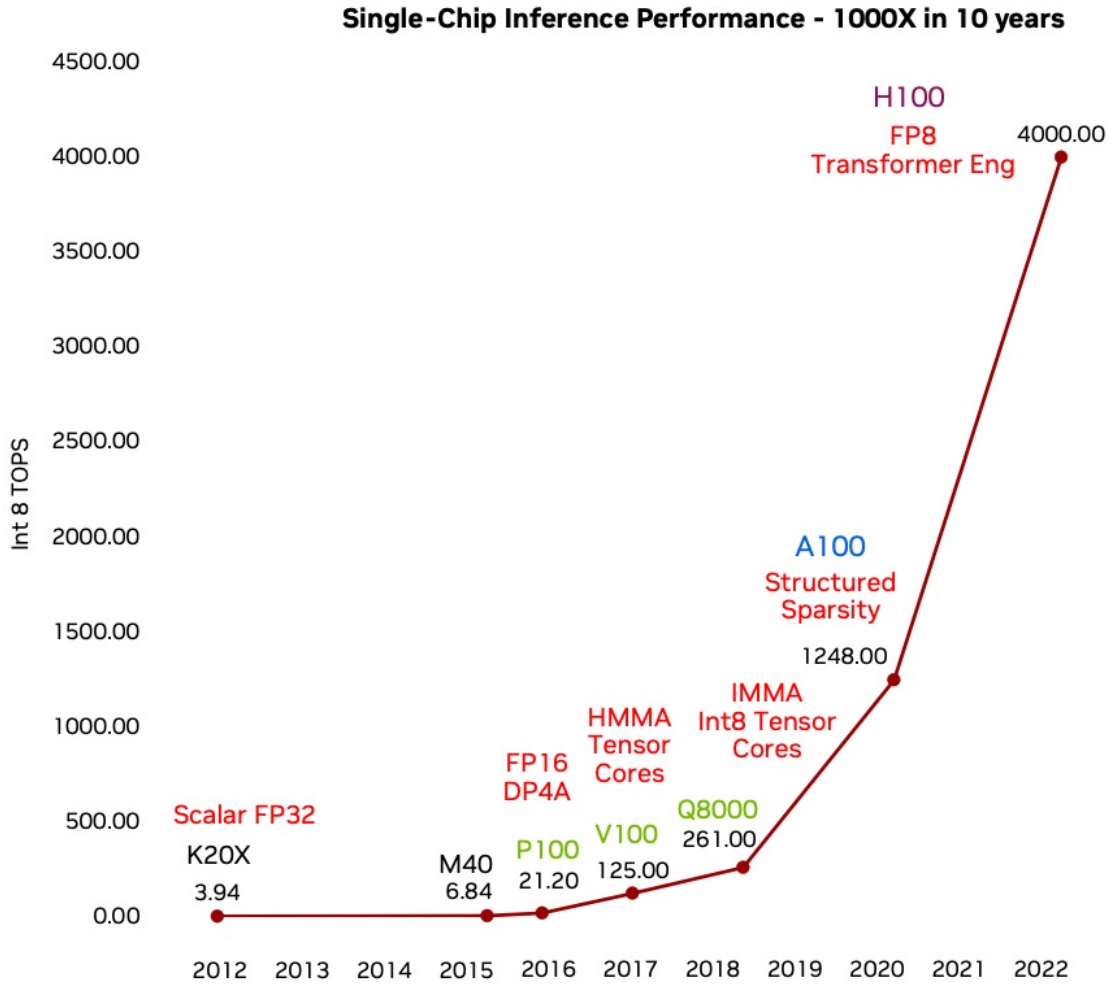
Reduce the size of the customized model, without compromising accuracy



The Arc of AI Computational Efficiency

Inference Energy Cost is Decreasing with Time

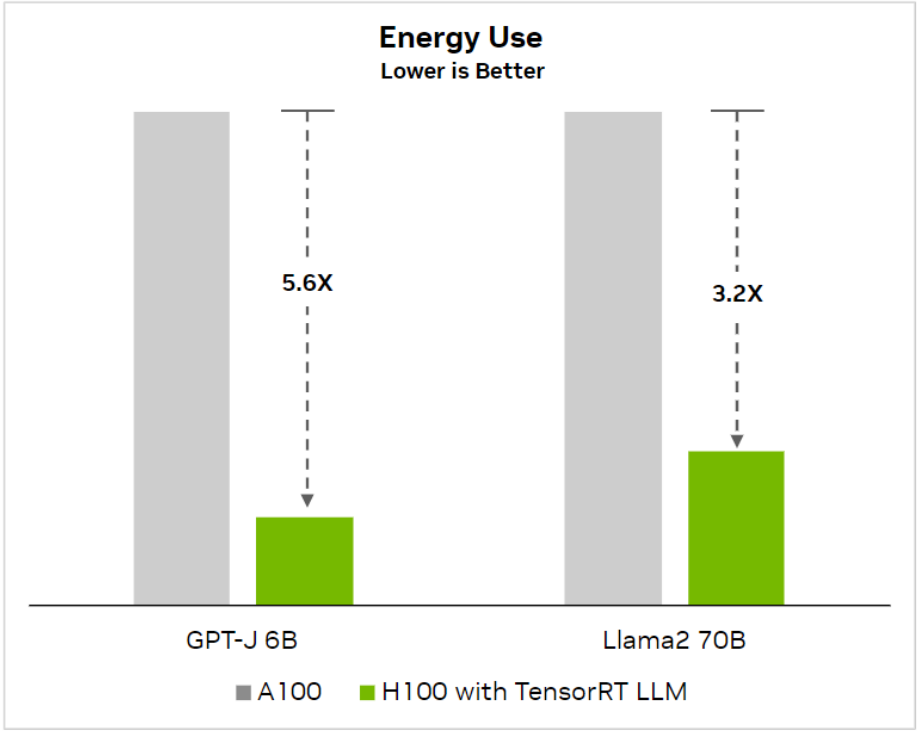
- 1000x GPU inference performance gain in 10 years
  - Only ~2x from process
  - Quantization (reduced precision, e.g., FP8, int8)
  - Tensor Core improvements
  - Structured sparsity
- Same node and system improvements as training



The Arc of AI Computational Efficiency

Inference Energy Cost is Decreasing with Time

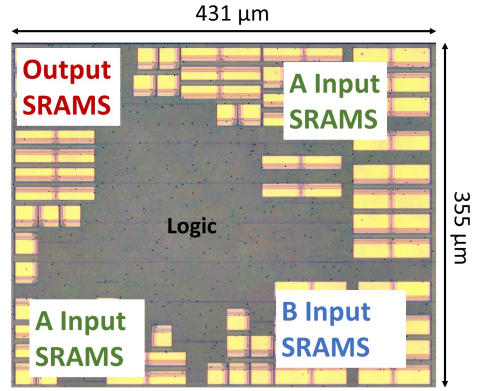
- 1000x GPU inference performance gain in 10 years
  - Only ~2x from process
  - Quantization (reduced precision, e.g., FP8, int8)
  - Tensor Core improvements
  - Structured sparsity
- Same node and system improvements as training
- Software efficiency and improved algorithms
- More to come!



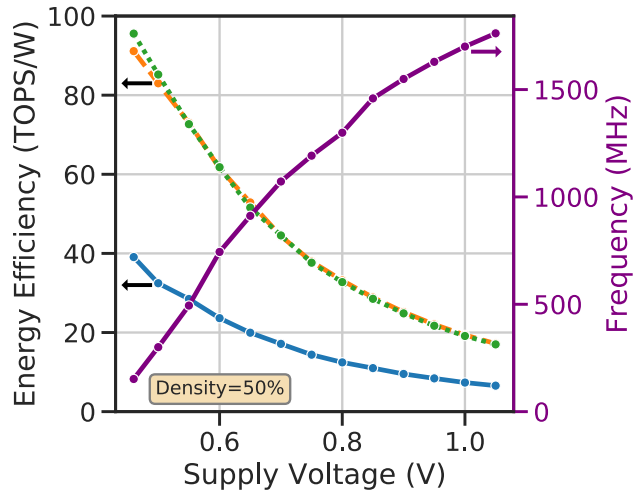
Energy-Efficient Transformer Inference Accelerator

NVResearch design demonstrates 95.6 TOPS/W

- Efficient architecture
  - MAGNet low-data-precision DL inference accelerator for Transformers
  - Multi-level dataflow improves data reuse and energy efficiency
- Low-precision data format: VS-Quant INT4
  - Hardware-software techniques to tolerate quantization error
  - Enable low cost multiply-accumulate (MAC) operations
  - Reduce storage and data movement
- Special function units
- ~10x lower energy than current SOTA (H100)



- TSMC 5nm
- 1024 4-bit MACs/cycle (512 8-bit)
- 0.153 mm² chip
- Voltage range: 0.46V – 1.05V
- Frequency range: 152 MHz – 1760 MHz



[Keller, Venkatesan, et al., "A 95.6-TOPS/W Deep Learning Inference Accelerator with Per-Vector Scaled 4-bit Quantization in 5nm", JSSC 2023]

NVIDIA

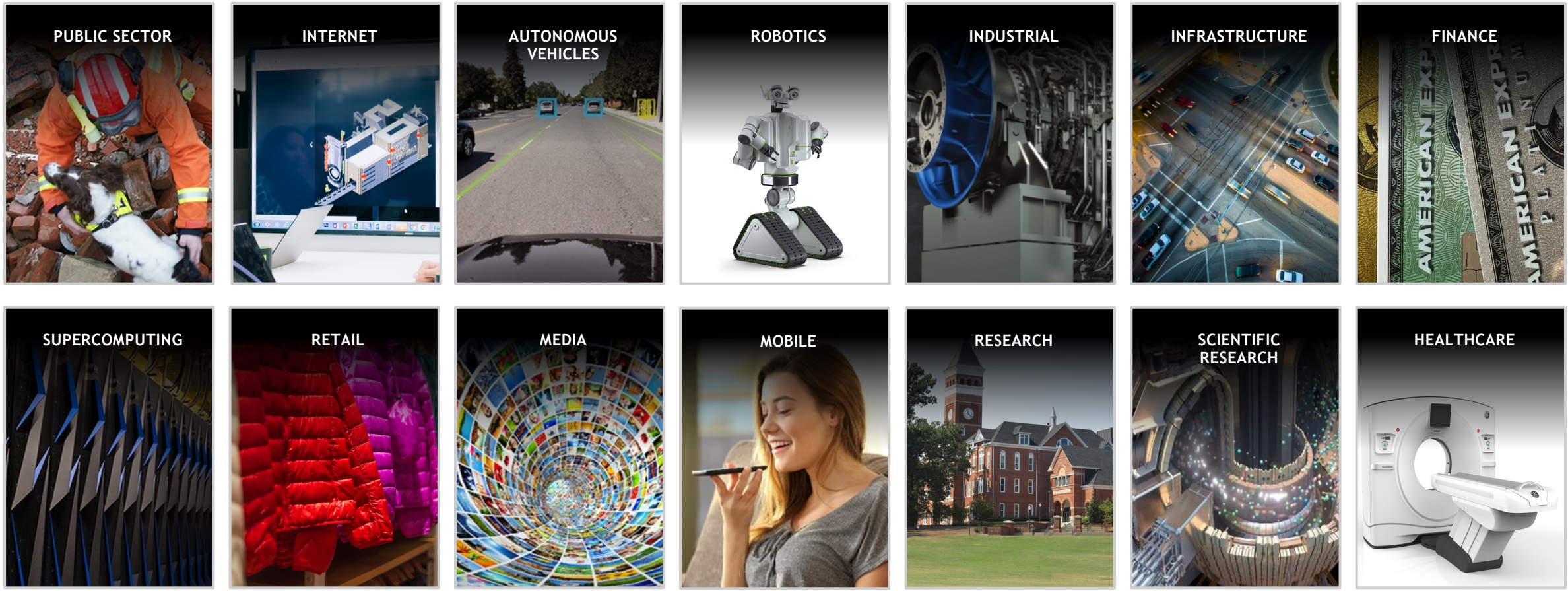
Generative AI Energy Use and Environmental Impact

Careful modeling and informed adoption is warranted, pessimism and fear are not

- Global datacenter scale and energy use have been growing, but efficiency has and continues to rapidly improve.
  - Even in AI-driven high-growth scenarios, datacenter fraction of total electrical energy is small **and** growth is largely green.
- Efficiency at every level of Generative AI stack (from algorithms to datacenter operations) is driven by strong incentives.
  - Renewable energy sourcing, efficient infrastructure and datacenter operations are good for bottom line.
  - Energy efficiency is first-order driver of AI technology and architecture improvement.
  - Every model architecture or algorithm improvement that improves training or inference efficiency increases economic value.
- Real-world Generative AI lifecycle parameters and choices affect energy efficiency by many orders-of-magnitude
  - Potential >10,000x difference in training energy (e.g., GPT-4-scale model from scratch vs. fine-tuned LLama 7B or similar).
  - Potential >10,000x difference in inference energy (e.g., GPT-4-scale model/100M users vs. optimized SLM w/RAG/10K users).

Generative AI Brings Extraordinary Value

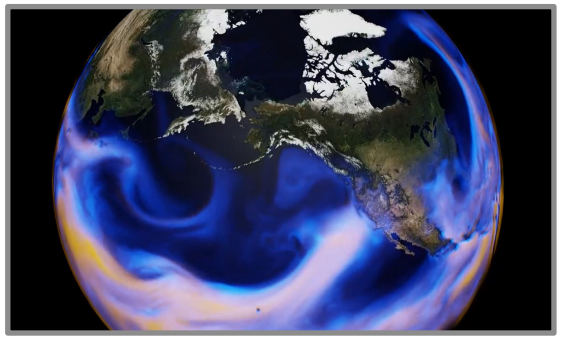
It’s worth getting right!



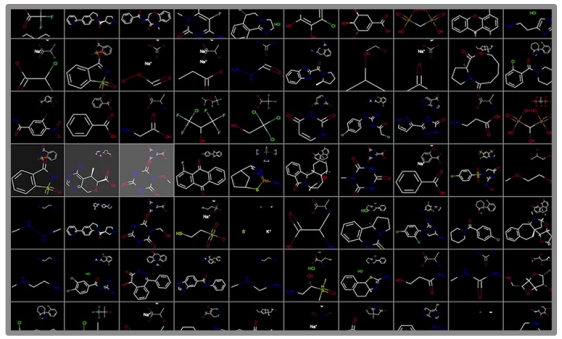

NVIDIA

Generative AI is a Revolutionary New Tool for Scientific Discovery

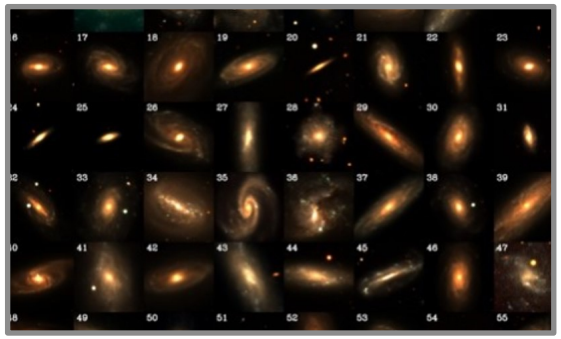
WEATHER FORECASTING
Forecasts Extreme Meteorological Disasters



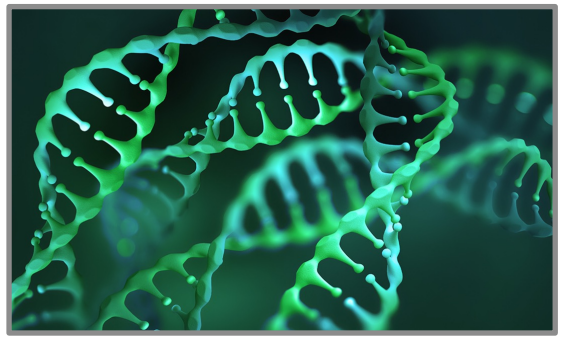
CHEMISTRY
Assists Researchers with Domain-Specific LLMs



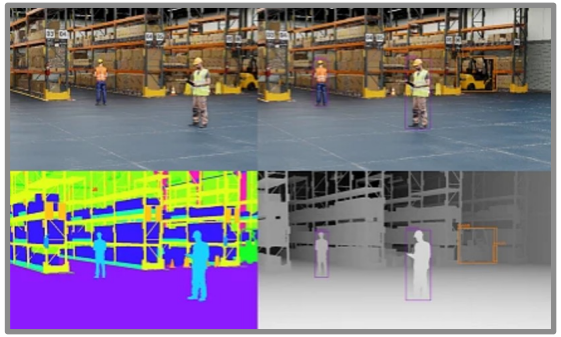
ASTRONOMY
Revives Corrupt Astronomical Images



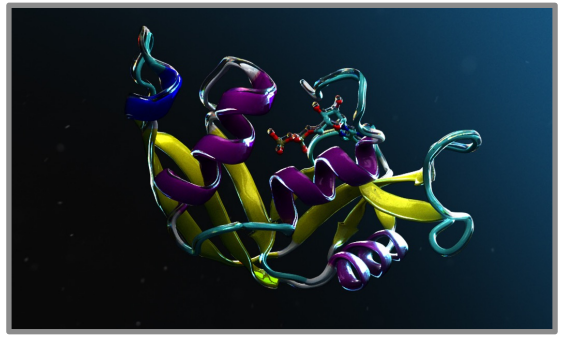
DNA SEQUENCING
Decodes the Language of Non-Coding DNA



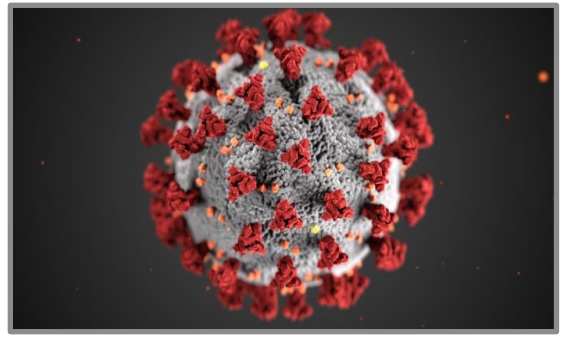
COMPUTER VISION
Reveals Image Features with Unsupervised AI

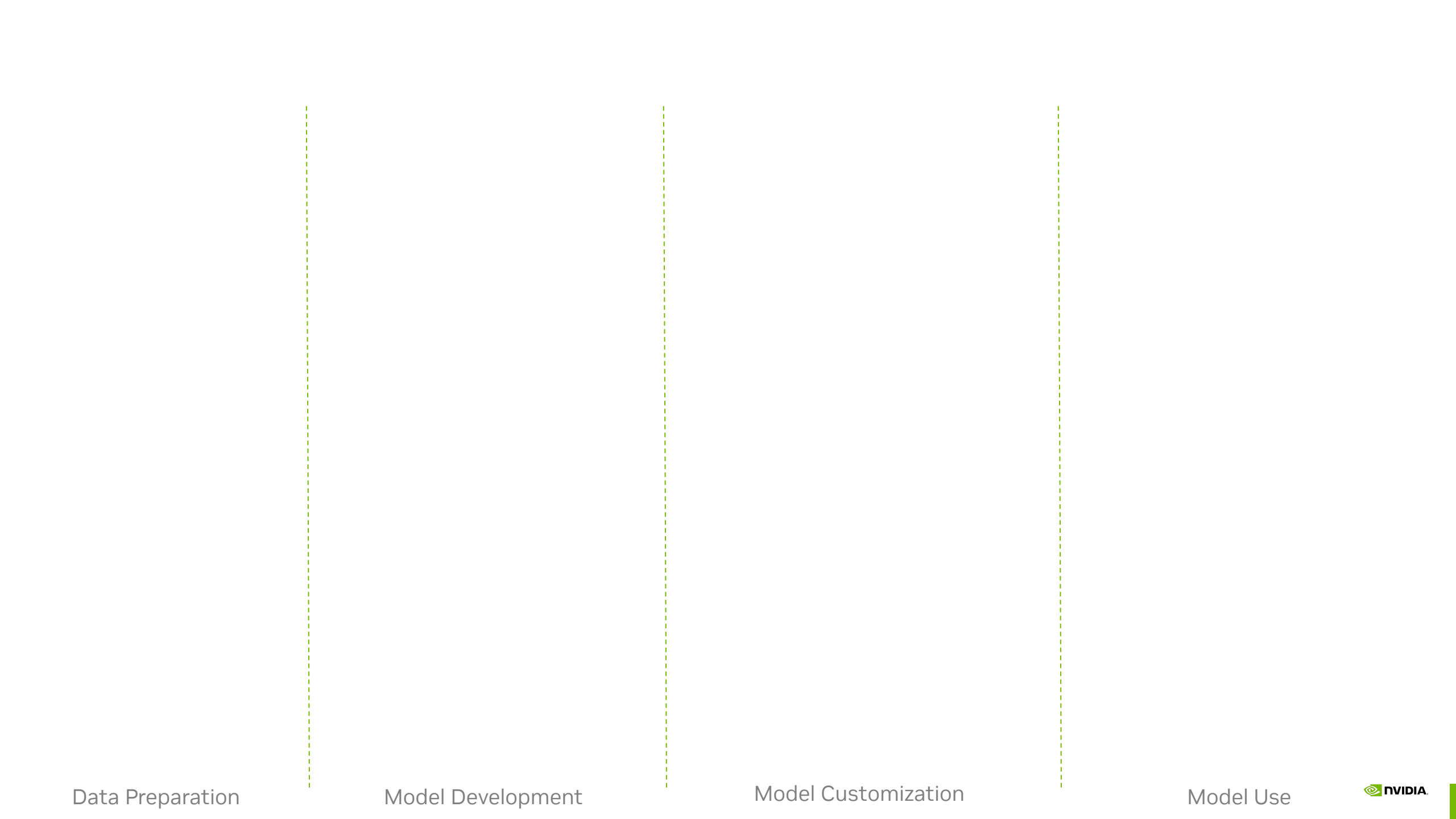


PROTEIN DISCOVERY
Generates New Protein Structures



DISEASE RESEARCH
Predicts Variants of SARS-CoV-2




Data Preparation
Model Development
Model Customization
Model Use