
Generative AI and Green Technology for Sustainability

葉信和 / Hsin-Ho Yeh

Founder & CEO / Software Engineer @ 信誠金融科技

hsinho.yeh@footprint-ai.com



Download Slides

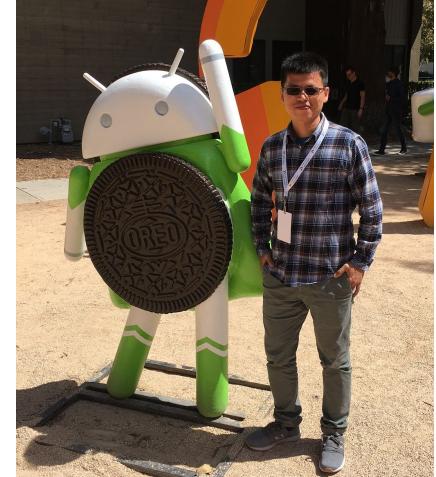
<https://bit.ly/48sGcff>



<https://github.com/FootprintAI/talks/tree/main/slides>

About me

- 2020 - Present at 信誠金融科技
 - **Tintin**: a machine learning platform for everyone
 - <https://get-tintin.footprint-ai.com>
- 2016 - 2020 at IglooInsure (16M+ in series A+ 2020)
 - Provide digital insurance for e-economic world
 - Funded in KUL, Headquartered in Singapore
 - First employee/ Engineering Lead / Regional Head/ Chief Engineer
- 2013 - 2016 at Studio Engineering @ hTC
 - Principal Engineer on Cloud Infrastructure Team
- 2009 - 2012 at IIS @ Academia Sinica
 - Computer vision, pattern recognition, and data mining
- CS@CCU, CS@NCKU alumni





Intro & timeline

- footprint-ai.com (信誠金融科技) is committed to providing Software-as-a-Service for serving AI/ML applications, specializing in:
 - Cloud-native architecture, MLOps, Green Tech, Internet-scale Data Analysis
- Practical experience in taking an idea (zero) to many products (one)
 - Singapore-based insurtech startup iglooinsure (18M USD/2020 Series A, 46M USD/2023 Series B).
- Business partner with IBM Taiwan (IBM.tw)
- Joined Nvidia Inception program in 2021.
- Moved into FinTechSpaces in 2023.



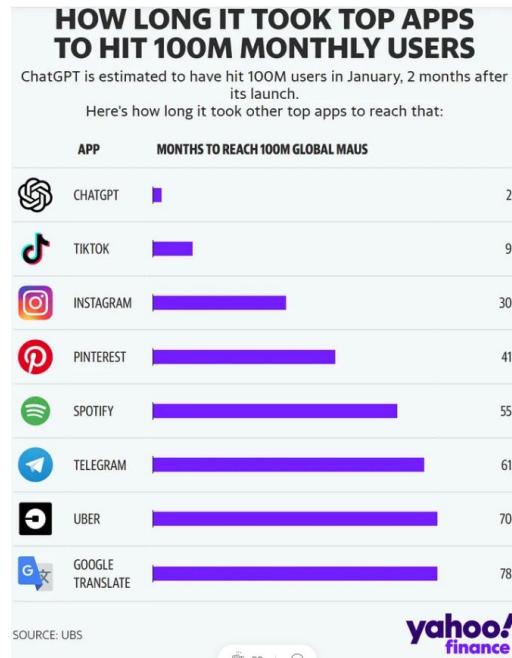
Agenda

- What is Generated Content AI (GCAI)? and how it impact our life?
- What is Green Technology / Green Software ?
- What is our effort on green economy?

ChatGPT



The golden age of NLP and distributed ML



What is standard language model?

Language models predict the next word(s) in a sequence

- Applications
 - Machine Translation
 - Text Classification
 - News Article Generation
 - Question Answering
 - Auto Completion
 - ...

$$P(w_1, \dots, w_m) = \prod_{i=1}^m P(w_i | w_1, \dots, w_{i-1})$$

HS

The Highest mountain in ____?



I would be happy to help you fill in the blank! Can you please tell me which region or country you would like me to find the highest mountain for?



What problem that ChatGPT wants to solve?

ChatGPT^[1] is a Generative Pre-trained Transformers that generate human-like text from a large corpus of text data.

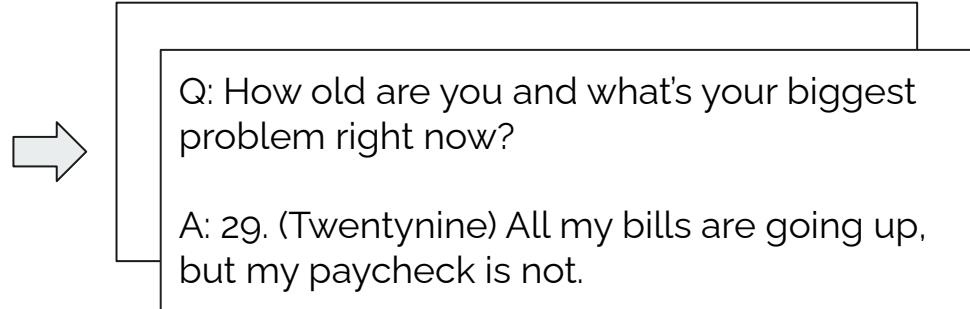
- Pre-training: Initial training process on a large text corpus where the model learns to predict the next word in a passage.
- Transformers [3]: attention mechanism to analyze input data, such as text, and extract important features or patterns

733 How old are you and what's your biggest problem right now?

2.8k Comments Share Save Hide Report 93% Upvoted

TRANSPARENT_Ink 3 hr. ago · edited 11 min. ago
29. (Twentynine) All my bills are going up, but my paycheck is not.
1.1k

SippingBinJuice 2 hr. ago
I feel this in my soul. I had the same salary between 2019 and 2022 and I just kept getting poorer and poorer, as the buying power of my salary dramatically decreased. In 2023 I got a \$10k raise, from getting a new job with 1000% more stress and responsibility. However, I only feel slightly better off than I was at the beginning of 2019. It's incredibly depressing.
186



Web-text corpus

[1] https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

[2] https://www.reddit.com/r/AskReddit/comments/11jhi8o/how_old_are_you_and_whats_your_biggest_problem/

[3] <https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>

Pretraining: Self-supervision

Self-supervision

1. The model's *only* objective is to learn co-occurrence patterns in the sequences it is trained on.
2. Alternatively: to assign high probability to attested sequences.
3. Generation then involves *sampling* from the model.
4. The sequences can contain anything.

The evolution of GPT [1]

Name	Year	Data Size	Model Parameters
GPT	2018	4.5 GB of text, from 7000 unpublished books of various genres.	0.12B
GPT-2	2019	40 GB of text from 8M web pages	1.5 B
GPT-3	2020	45T data from <ul style="list-style-type: none">● Common Crawl● Web Text● Books● Wiki	175 B (i.e. 800GB in memory)

How ChatGPT is trained.

Step 1

Collect demonstration data
and train a supervised policy.

A prompt
is sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



This data is used to
fine-tune GPT-3.5
with supervised
learning.

Step 2

Collect comparison data and
train a reward model.

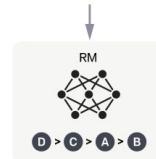
A prompt and
several model
outputs are
sampled.



A labeler ranks the
outputs from best
to worst.



This data is used
to train our
reward model.



Step 3

Optimize a policy against the
reward model using the PPO
reinforcement learning algorithm.

A new prompt is
sampled from
the dataset.



The PPO model is
initialized from the
supervised policy.



The policy generates
an output.



The reward model
calculates a reward
for the output.



The reward is used
to update the
policy using PPO.

(Recent) History: ChatGPT

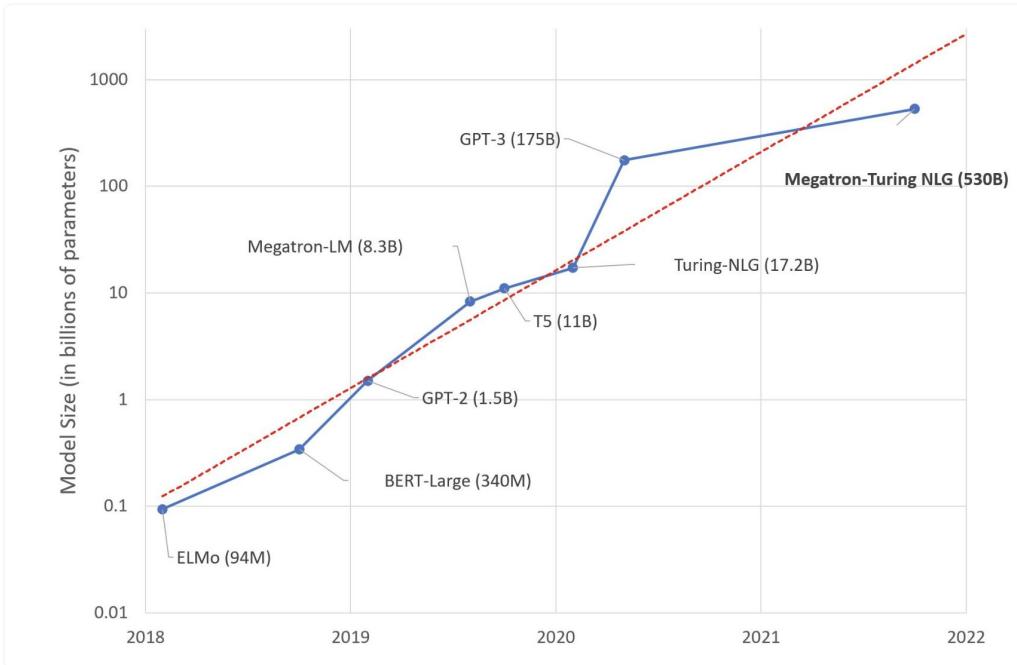
(rumor) about 10x spend on human annotation budget

(rumor) modification of RLHF training

Huge impact!



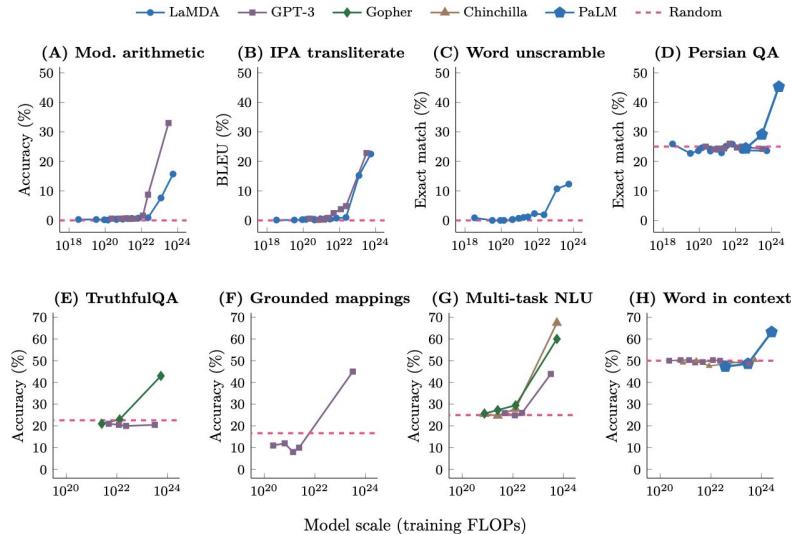
Large Language Models: A New Moore's Law?



Ref: <https://huggingface.co/blog/large-language-models>

The trend of LLM: Emergent Ability

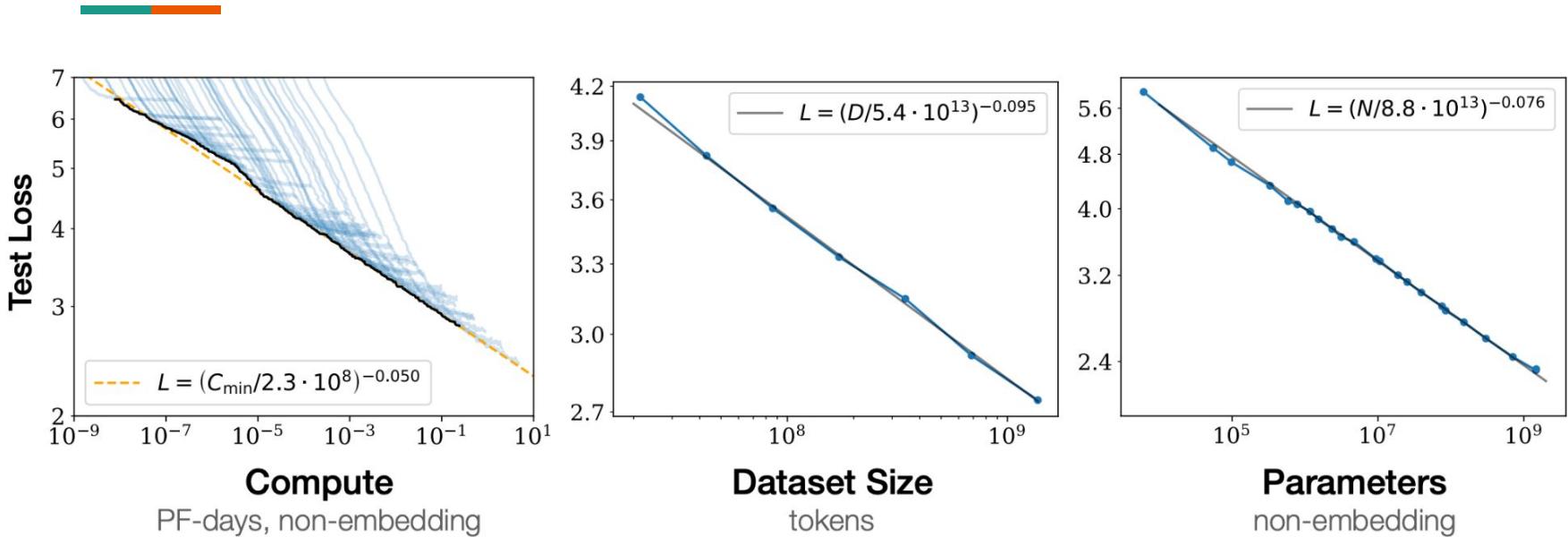
An ability is emergent if it is not present in smaller models but is present in larger models. - TMLR 2022



RTX 3090 is about $35 * 10^{12}$ FLOPs.

For 10^{24} , it needs
~ 10^{19} FLOP per day
And running for 10^5 days

The trend of LLM: Scaling Law for Natural language model



Kaplan, Jared, et al. "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020).

Supervision vs. In-context learning

Standard supervision vs. in-context learning

Standard supervision for
nervous anticipation

My palms started to sweat as the lotto numbers were read off.	<code>nervous anticipation = 1</code>
I took a deep breath as the curtain started to rise on my debut night.	<code>nervous anticipation = 1</code>
I couldn't shake a deep feeling of unease about the whole affair.	<code>nervous anticipation = 0</code>

“Few-shot in-context learning”

Hey model, here is an example of nervous anticipation: “My palms started to sweat as the lotto numbers were read off.”

Hey model, here’s an example without nervous anticipation: “...”

In-context learning: use prompting, instead of fine-tuning on each domain

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



An example on Few shots promoting

HS

please fill __ with meaningful words. The Highest mountain in __ is __ .



Sure, here's an example sentence:



"The Highest mountain in South America is Aconcagua."

In context learning: chain of thoughts

HS 10隻兔子有幾隻腳?

如果每只兔子都有4隻腳的話，那麼10隻兔子總共會有40隻腳。

HS 一隻兔子平均有兩隻腳，三隻兔子平均有兩隻腳。請問十隻兔子平均有幾隻腳?

這個問題有點棘手，因為兔子平均有兩隻腳和三隻兔子平均有兩隻腳這兩個條件似乎有些相互矛盾。

如果我們假設每只兔子都有兩隻腳，那麼10隻兔子平均會有20隻腳。

如果我們假設三隻兔子的總腳數是6，那麼一隻兔子的平均腳數就是2。因此，10隻兔子的總腳數將是10乘以2，等於20。

但是，這個問題的問題陳述似乎暗示每只兔子的腳數可能會有所不同，因此我們無法確定答案。

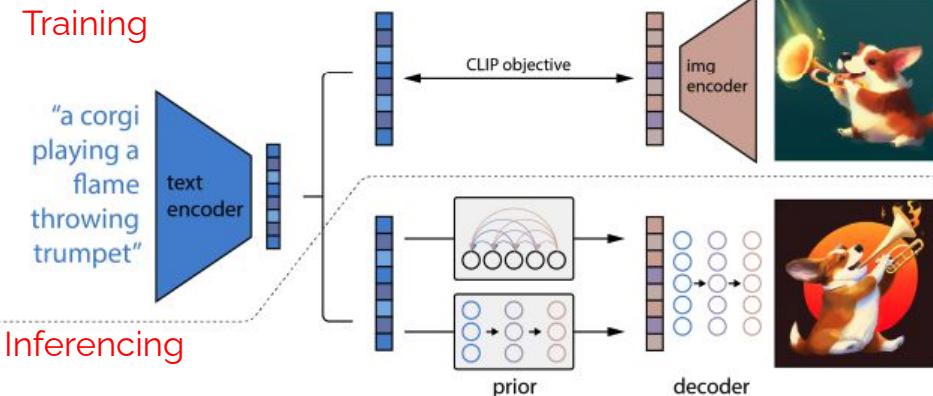
HS Q: 一隻兔子平均有兩隻腳 A: 一隻兔子有兩隻腳，所以平均是2/1等於兩隻腳。Q: 三隻兔子平均有兩隻腳 A: 三隻兔子有六隻腳，所以平均是6/3等於兩隻腳。請問十隻兔子平均有幾隻腳?

根據這些條件，我們可以計算出每隻兔子平均有兩隻腳，無論是單獨的一隻兔子還是三隻兔子一起算。因此，根據平均數的性質，10隻兔子平均也會有兩隻腳。所以答案是2隻腳。

Teach ChatGPT how to think by prompting

Text-to-Image Generation

Joint correlation between text-embedding and image-embedded



> a panda family with five (two parents with three kids) wear hats, with Chinese New Year style, saying "Happy New years"



Not only generative text, but also audios

Napoleon Crossing the Alps - Jacques-Louis David



"The composition shows a strongly idealized view of the real crossing that Napoleon and his army made across the Alps through the Great St Bernard Pass in May 1800." By [wikipedia](#)

0:06 / 0:30

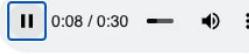


The Starry night - Vincent van Gogh



"The Starry Night (Dutch: De sterrennacht) is an oil-on-canvas painting by the Dutch Post-Impressionist painter Vincent van Gogh. Painted in June 1889, it depicts the view from the east-facing window of his asylum room at Saint-Rémy-de-Provence, just before sunrise, with the addition of an imaginary village." By [wikipedia](#).

0:08 / 0:30



Not only generative text, but also videos

> a panda family with five (two parents with three kids) wear hats, with Chinese New Year style, saying "Happy New years"



New challenges while AI/ML is booming

ML model is **complexing**.

Cost spent on ML infrastructure (GPU) are huge.

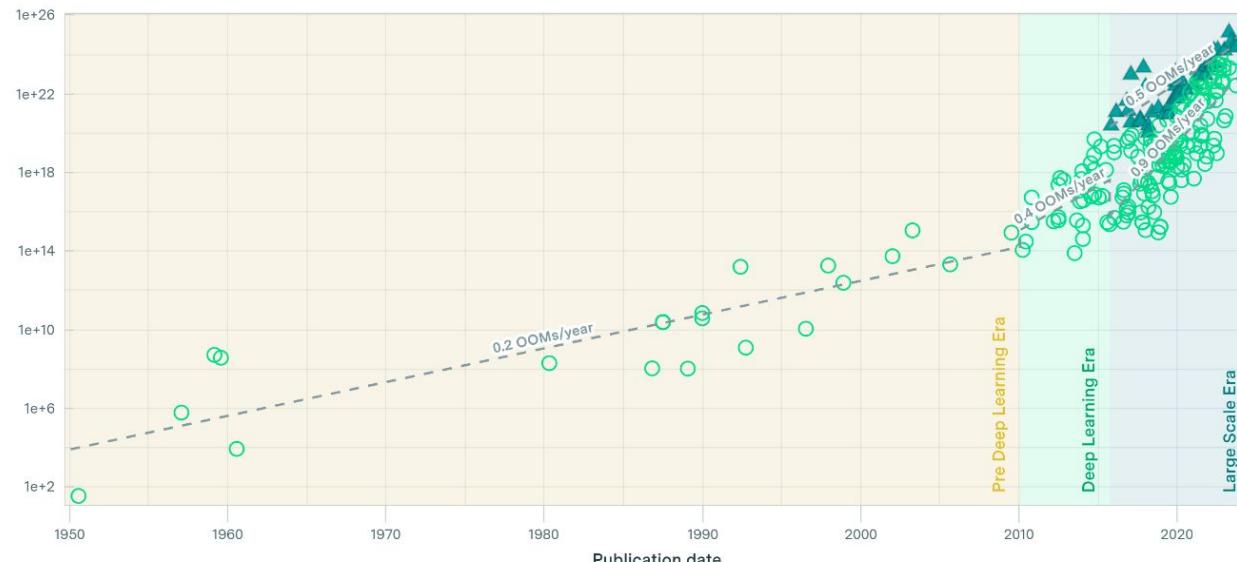
The proliferation of AI/ML is significantly **driving up** carbon emissions.

ML Model Growth history

Training Compute of Notable Machine Learning Systems Over Time

≡ EPOCH

Training compute (FLOP)



The gas emission is also accelerating ...

Common carbon footprint benchmarks

in lbs of CO₂ equivalent

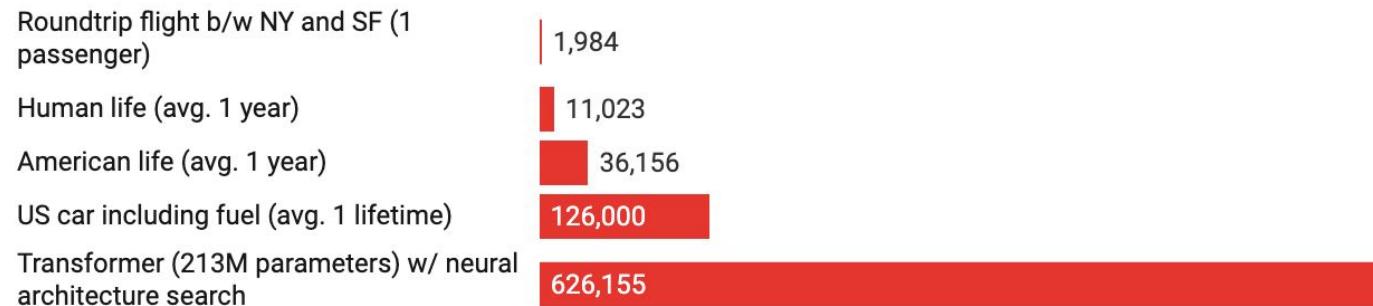
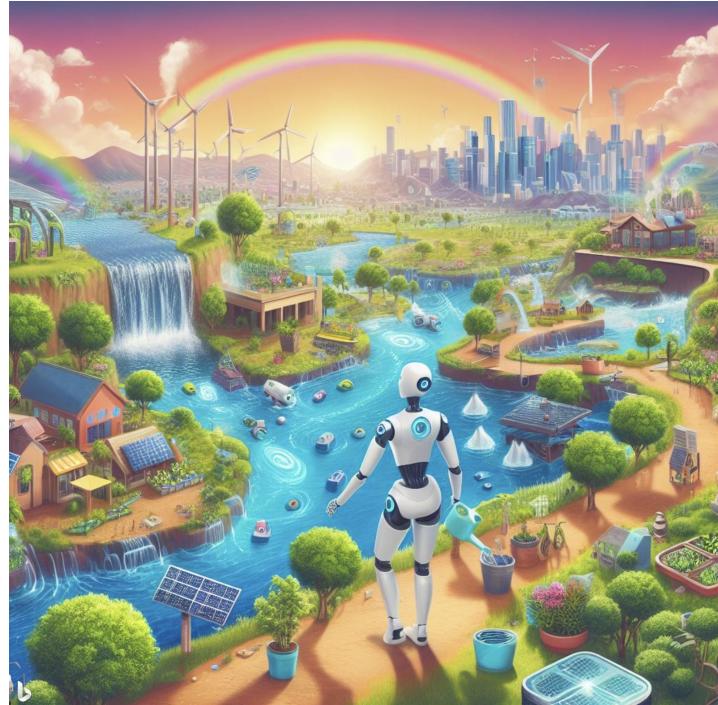


Chart: MIT Technology Review • Source: Strubell et al. • [Created with Datawrapper](#)



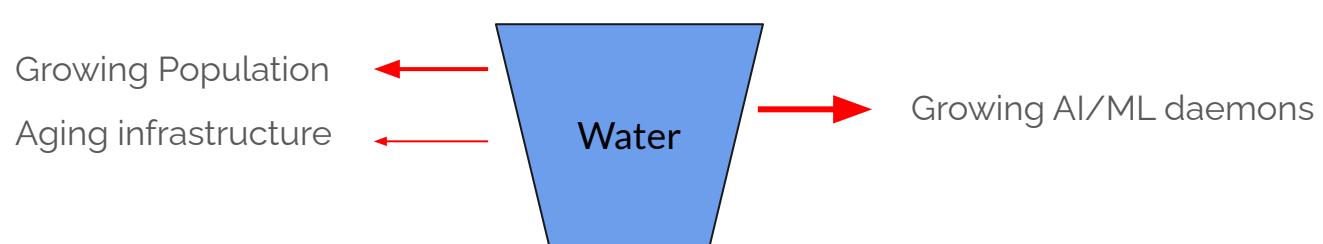
<https://www.bing.com/images/create/people-don27t-care-about-carbon-emission-and-water-/654fe07a3f8545e099187f40b34fded9?id=JNCSWj7OssHkQcitT2C9yQ%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay>

<https://www.bing.com/images/create/create-a-photo-where-an-ai-is-embrace-sustainability/654fe07a3f8545e099187f40b34fded9?id=Y4ngLElWrhJD%2fMcqQJXcg%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay>

If we do not make a change from now

We need AI/ML to accelerate our daily tasks and automation, but ...

- Data Centers (DCs) in US are consumed **2%** electricity use.
- ChatGPT needs to “drink” a 500ml bottle of water for every simple 20-50 questions and answers (and GPT-4 is even thirstier) [1]



[1] <https://arxiv.org/pdf/2304.03271.pdf>

“The gains we get from hardware, in terms of output per megawatt, will be dwarfed by the gains we get from software”

**Paul Calleja,
director of research computing services
University of Cambridge**



<https://www.bing.com/images/create/the-gains-we-get-from-hardware2c-in-terms-of-output/65517760b1c24629aef5e094f85edb0?i=yO85O15ojMiiKQdM5WD62w%3d%3d&vi=ew-detailv2&idpp-genimg&FORM=GCRIDP&mode=overlay>

What is Green Software? (1/4)



Green software engineering minimizes environmental impact by integrating sustainable practices into software development, creating **responsible** systems that are **energy-efficient** and **resource-optimized**

<https://engineering.leanix.net/blog/sustainable-green-software-engineering/>

What is Green Software? (2/4)



- Architecture/design decision
 - **Static code analysis** for estimating carbon emission
 - Real time power consumption monitoring with **dynamic code analysis**.
 - **Serverless computing** shares infrastructure resources by executing functions only on demand.

What is Green Software? (3/4)

- Green Coding Practice
 - Choose statically-typed language over scripting language.

Time & Memory	Energy & Time	Energy & Memory	Energy & Time & Memory
C • Pascal • Go	C	C • Pascal	C • Pascal • Go
Rust • C++ • Fortran	Rust	Rust • C++ • Fortran • Go	Rust • C++ • Fortran
Ada	C++	Ada	Ada
Java • Chapel • Lisp • Ocaml	Ada	Java • Chapel • Lisp	Java • Chapel • Lisp • Ocaml
Haskell • C#	Java	OCaml • Swift • Haskell	Swift • Haskell • C#
Swift • PHP	Pascal • Chapel	C# • PHP	Dart • F# • Racket • Hack • PHP
F# • Racket • Hack • Python	Lisp • Ocaml • Go	Dart • F# • Racket • Hack • Python	JavaScript • Ruby • Python
JavaScript • Ruby	Fortran • Haskell • C#	JavaScript • Ruby	TypeScript • Erlang
Dart • TypeScript • Erlang	Swift	TypeScript	Lua • JRuby • Perl
JRuby • Perl	Dart • F#	Erlang • Lua • Perl	
Lua	JavaScript	JRuby	

What is Green Software? (4/4)



- Moving to **Cloud** or **Edge** Computing.
 - Approaches like virtualization, energy-efficient data centres, multi-tenancy, etc., enable cloud computing to reduce energy usage and carbon emissions.
 - Edge computing allows you to redistribute computation near the users to lower costs
- Building the mindset
 - Trade-off between business and environmental goals.

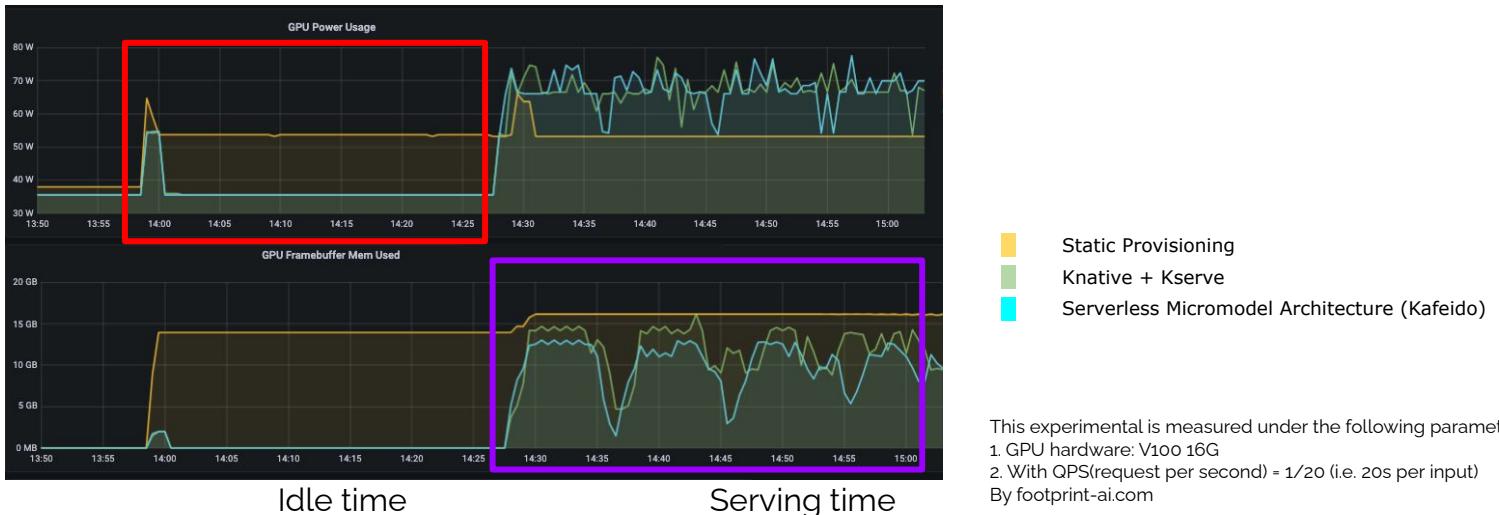
How cloud-native architecture can do to save our environment?

From Kserve and Fluid Perspective

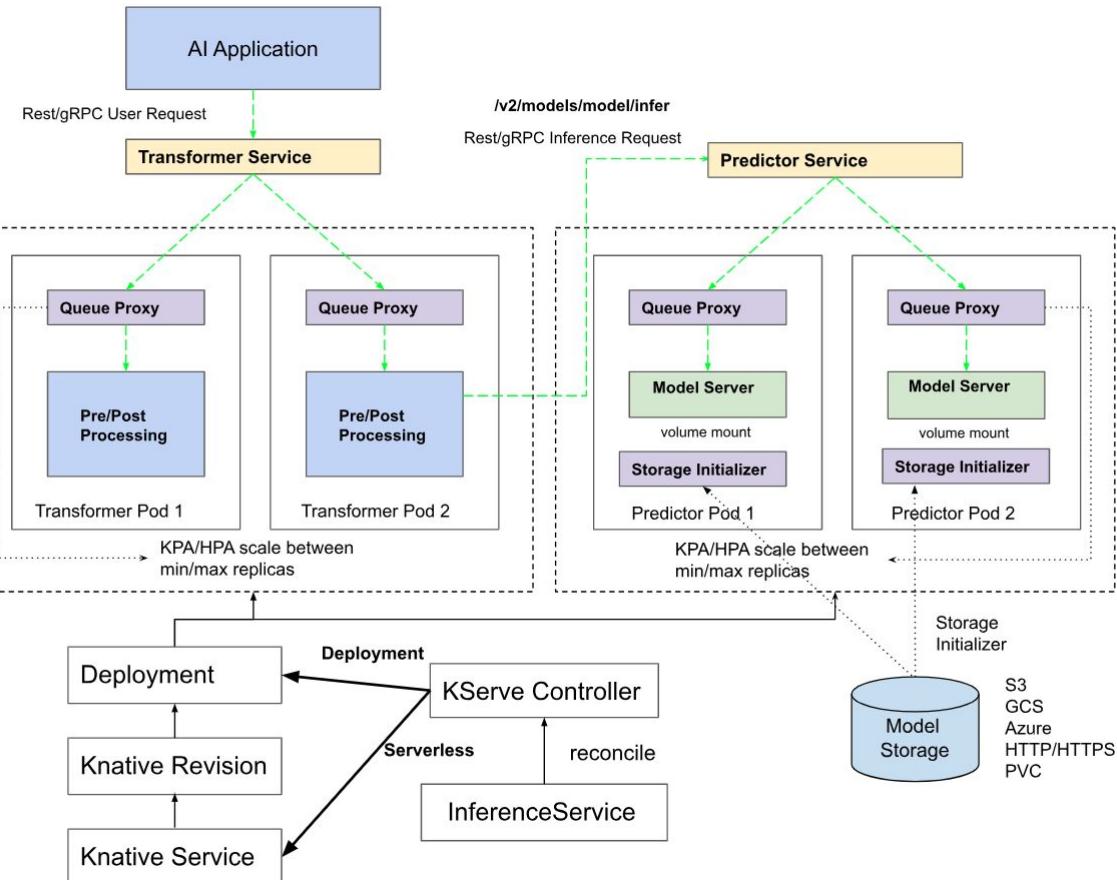


<https://www.bing.com/images/create/how-cloud-native-architecture-could-make-our-envir/1-65827acbf4314c3abf6448d31e46ae81?id=050jKSTaUdG7Pe1c5z1KA%3d%3d&view=detailv2&idpp-genimg&FORM=GCRIDP&mode=overlay>

How Kserve Serverless can save your cost?



Kserve Architecture





Kserve Example

```
1  apiVersion: "serving.kserve.io/v1alpha3"
2  kind: "Service"
3  metadata:
4    name: "tensorflow-gpu"
5  spec:
6    predictor:
7      tensorflow:
8        storageUri: "gs://kfserving-examples/models/tensorflow/flowers"
9        resources:
10          limits:
11            nvidia.com/gpu: 1
12        nodeSelector:
13          cloud.google.com/gke-accelerator: nvidia-tesla-k80
```

Kafeido: machine learning platform for green economy

Machine Learning Platform For Green Economy

Kafeido : Machine Learning Platform For Green Economy

Our one-step platform equips both the serverless and the exclusive micro-model deployment architecture to achieve real-time machine learning model deployment for the green environment. Kafeido can not only save your personnel costs on model operations but also avoid the excessive expenses on hardware and electricity.

Features highlighted

- POINT 01** Kafeido is Your Best Choice
- POINT 02** Developing with heterogeneous frameworks
- POINT 03** Serverless Architecture and Horizontal Expansion Advantages
- POINT 04** Micro-model Green Deployment Architecture

How does Kafeido work?

Applicable places: community, school, hospital, shopping mall, factory, corporate
Deployment plan: On-prem, SaaS

- Step 01** Select a model from our model zoo or upload your own model
- Step 02** Generate the model deployment
- Step 03** Select your data source
- Step 04** Automatic management model inference

Customer success story : Sustainable Smart City Monitoring Center

- Existing Challenges**
The traditional command and control center (or called ICCO) simply collects all information from each camera to one control panel. The security personnel can monitor the entire city conveniently and provide adequate assistance as needed. However, when the monitoring scope increases (e.g. from CCT to drone shooting, from single point to field monitoring...etc.), long-term dependence on the security personnel not only increases personnel costs dramatically but also fails to operate in high quality.
- ESG Awareness**
Environmental monitoring (e.g. air pollution, factory waste gas emission...etc.) is another challenge—how to import the green machine learning platform to keep a city green while minimizing carbon emissions also becomes a hot topic.
- Kafeido Accomplished**
Combining multiple data sources from drones and CCTV cameras and various machine learning models, Kafeido triggers warning events based on the model inference results to notify related personnel via SMS and/or email and achieves a 24/7 decentralized monitoring process. With Kafeido, you can easily scale out your data sources for monitoring and state-of-the-art machine learning models to guard valuable properties for you and for our next generation.

Our professional software architecture and green-oriented solution help your business apply much easier and more affordable AI technologies!

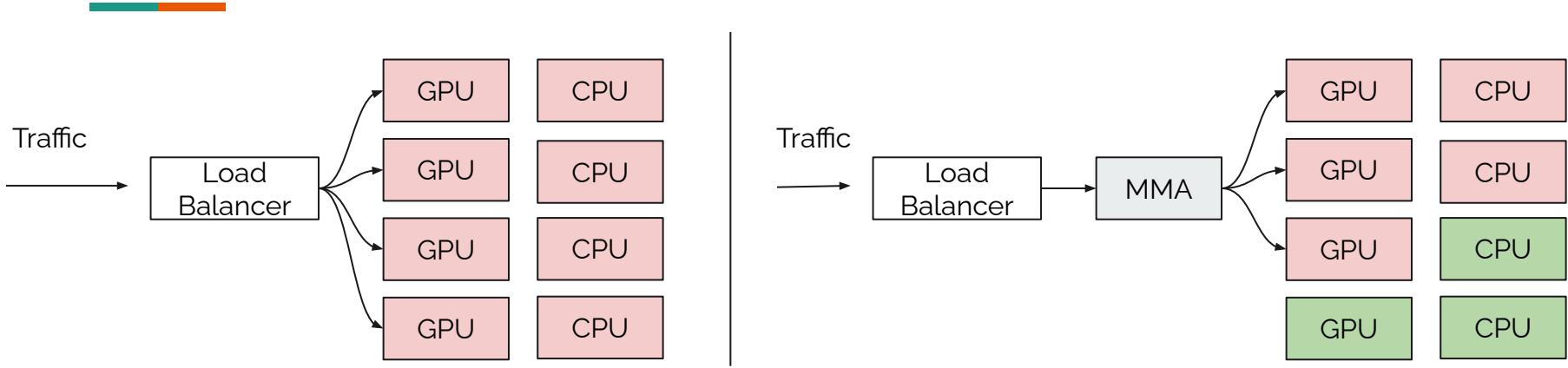
Contact Us
信誠金融科技股份有限公司 (footprint-ai.com) dedicates in developing machine learning platform and providing AI-oriented software services. We are expert in machine learning platform, data middle platform, and their customizations.
Address : No. 287-2, Sec. 3, Chengde Rd, Taipei City 103, Taiwan.
Email : kafeido@footprint-ai.com

Tintin
Machine Learning Platform For Everyone

NVIDIA.
INCEPTION PROGRAM

XINCHEN FINTECH CO., LTD

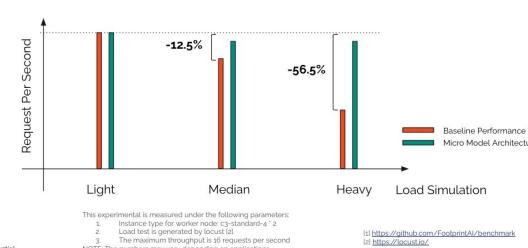
Increasing utilization with micro model architecture



Energy Cost Efficiency - GPU Inference Case Study



Energy Cost Efficiency - CPU Inference Case Study



Machine Learning Platforms for Green Economy

Kafido : Machine Learning Platform for Green Economy

How does Kafido work?

Customer success story - Sustainable Smart City Monitoring Center

Features Highlighted

Feature Accelarated

Development with heterogeneous frameworks

Services Architecture and Horizontal Scalability Advantages

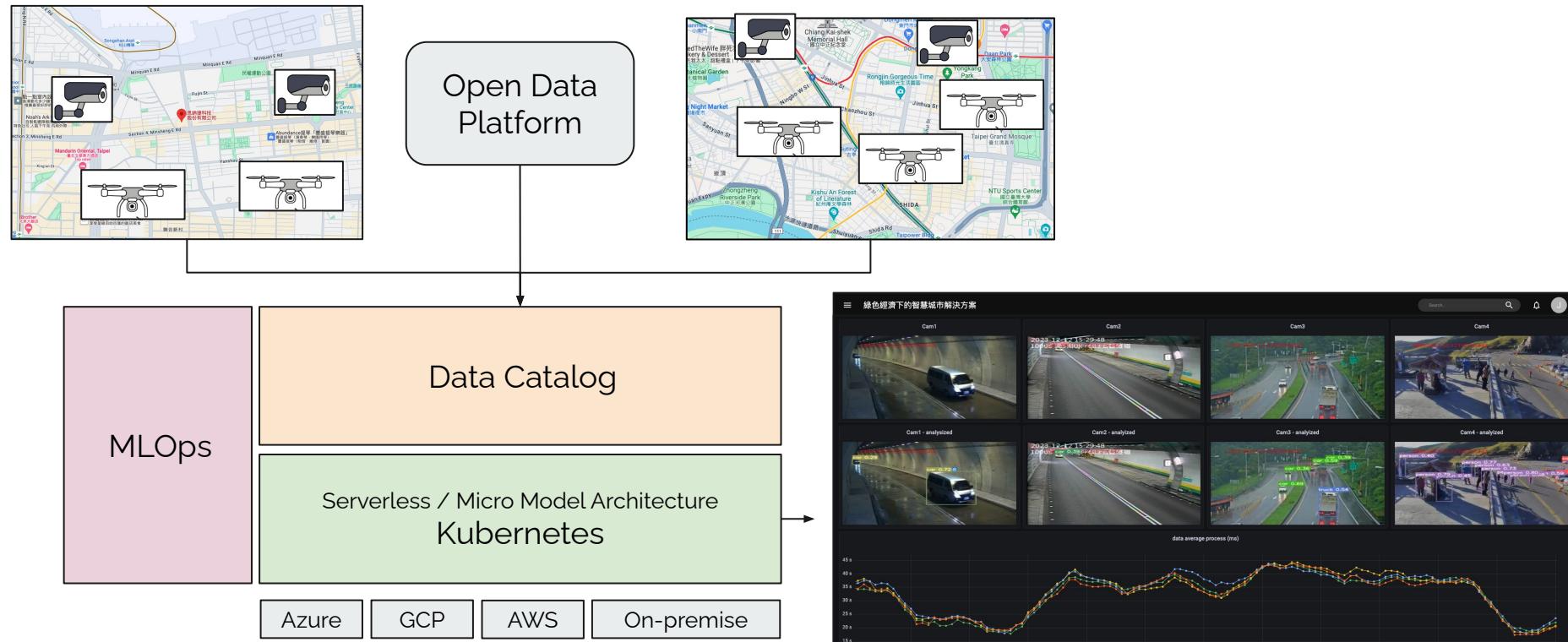
Micro-model Green Deployment Architecture

Contact Us

© 2023, Xinchen fintech Co. LTD Confidential.



Application I - Smart Surveillance Solution for Green City [2022Q4, launched]



Parametric Insurance

For protect uncertainty when we move to sustainable solution.



<https://www.bing.com/images/create/parametric-insurance-for-protect-uncertainty-when-/654ff2b8b6f148aaaae103722a52b9b7?id=MwAEAGt%2bwrmf4Kn5vgEpaPg%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay>

Parametric Insurance

- One variable insurance
 - Agriculture insurance where the farmer wants to protect his/her yields from rainfall over a time period
- Payout is implemented when a certain event happened in a certain time period
- I.e. $f(\text{time period}) > \text{threshold}$, where f could be
 - Weather condition
 - Bitcoin pricing
 - Cloud service status



Existing Parametric Insurance

- Weather-based index insurance
 - For protecting farmers loss due to extreme climate change.
- Renewable Energy Insurance
 - For mitigating construction and operation risk, such as Offshore wind power and photovoltaic.
- Energy Efficiency Insurance
 - For easing risks from business interruption, material damage, and so on.

Parametric Insurance for NEXT

- We are designing a new parametric insurance for guarding our virtual properties.
-
- Feel free to reach out if you have idea and **We are hiring.**



<https://www.bing.com/images/create/c...?id=vYRditgOsNkH6GtTwudwFg%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay>

Key take a way

- GCAI has changed the way how we work, new opportunities come with new threats.
- Software, not hardware, plays a key role the way of combating and preserving environment.
- Fostering sustainability improves efficiency, a win-win approach.



Thank You

and follow us on Facebook



Contact Us: partnership@footprint-ai.com

careers@footprint-ai.com