香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Towards Edge-Native Foundation Models:

A Paradigm Shift in AI Construction, Deployment and Governance at the Edge

郭嵩教授

FCAE, FIEEE, MAE

计算机科学与工程系

香港科技大学

电话：+852-2358-8833

电子邮箱：songguo@cse.ust.hk

PERVASIVE INTELLIGENCE

Pervasive Edge Intelligence Lab 普適邊緣智能實驗室

Research Interest

Artificial Intelligence

Cloud/Edge Computing

Blockchain

Internet of Things

Big Data

Books

- Edge Learning for Distributed Big Data Analytics Theory, Algorithms, and System Design
- Machine Learning on Commodity Tiny Devices

Team

40+ researchers:

- **2** Research Assistant Professor
- **8** Postdocs
- **20+** PhD students
- **5+** Research Assistant

Honor & Award

- 2024 Edward J. McCluskey Technical Achievement
- Gold Medal in 2023 Geneva Inventions Expo
- Gold Award in 2023 AsiaWorld-Expo

Agenda

01
Background
The Rise of Edge AI

02
Research
Edge-Native Foundation Models

03
Impact
Research and Development

04
Prospect
Open Issues

The Rise of Edge AI

01

AI computing migrates to the edge side

From Cloud Intelligence to Edge Intelligence

Long latency

Insufficient bandwidth

High energy consumption

Privacy exposure

Cloud Intelligence

Cloud

Model training & Inference

NETFLIX amazon

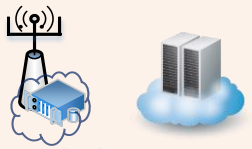
Various Web 2.0 services

Edge Intelligence

Cloud

Model training & Inference

Edge



Model training & Inference

Smart city

Autonomous driving

Industrial 4.0

Real-time Analytics and Decision

s-level to $ms/\mu s$-level

Communication **Cost Saving**

847 ZB vs. 19.5 ZB Per Year

Store **Sensitive Private** Data Locally

Evolution of Edge Intelligence

NVIDIA Jetson TK 1

NVIDIA Jetson TX 2

NVIDIA AGX Xavier

Huawei Atlas 200

Huawei Atlas 300

Huawei Atlas 500

Huawei IES

Huawei Atlas 200 DK A2

NVIDIA

HUAWEI

2009 — 2012 — **2014** — 2017 — **2018** — 2020 — 2022 — **2024**

"Fog Computing and its Role in Internet of Things" (Bonomi et al.)

CISCO

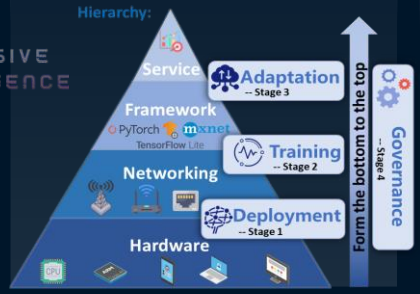
"Emergence of Edge Computing" (Satyanarayanan)

PEILab Established

PERVASIVE INTELLIGENCE

Research Grants Council of Hong Kong 香港 研究資助局

Edge Learning, HK Research Impact Fund



Textbooks

The Next Decade of Edge AI: From Analysis to Generation

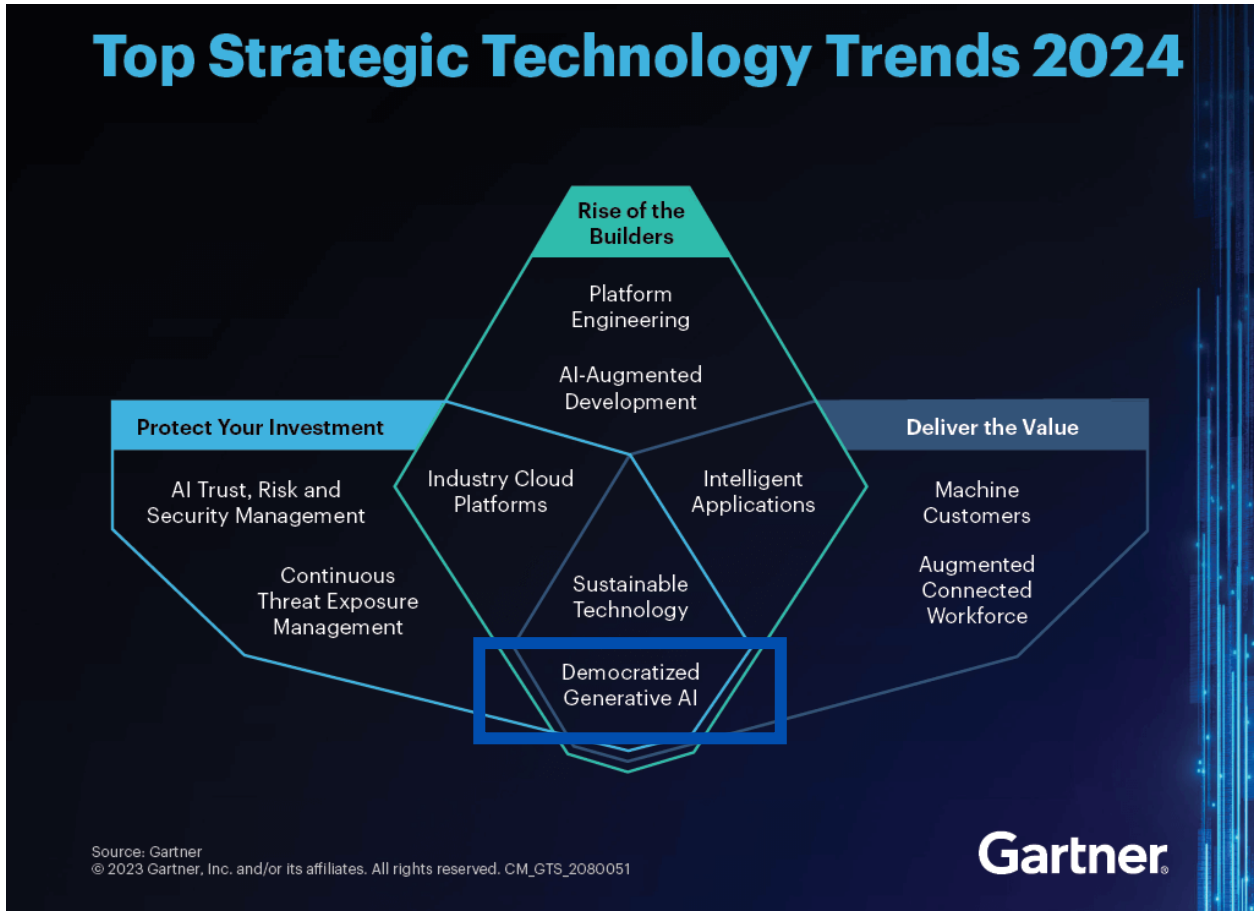
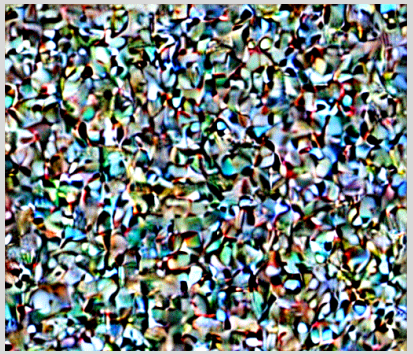
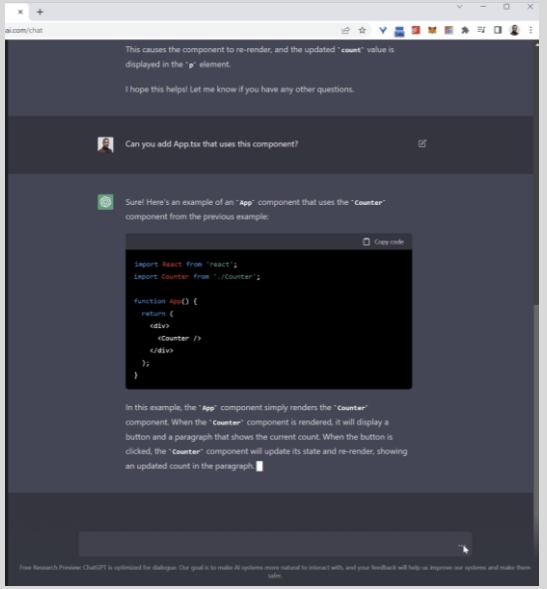


Image (Dall-E 2, **Stable Diffusion**)



Video (**Sora**, Microsoft X-CLIP)

Text (**ChatGPT-3**, DeepMind Gopher)



Others (DreamFusion, Tabnine, Stability.ai, ...)

3D	Speech	Game
Metaverse	Music	...

By 2025, Gartner expects generative AI to account for 10% of all data produced, up from less than 1% today.

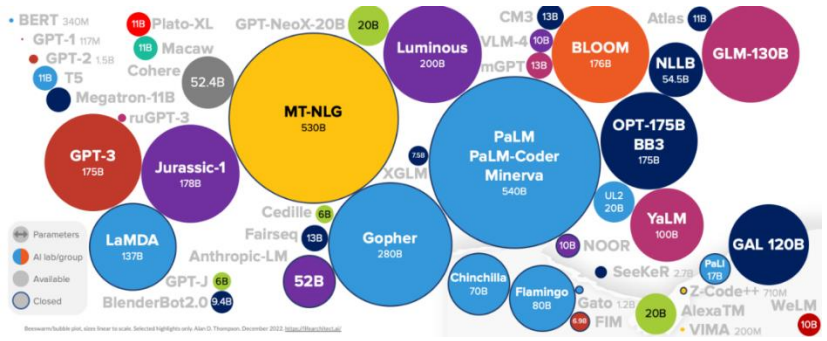
--- Gartner, Inc

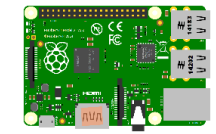
Edge-Native Foundation Models

02

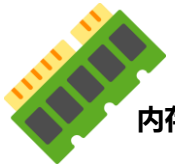
Ubiquitous foundation model serving

边缘资源受限、边缘环境复杂、边缘风险加剧





处理能力弱

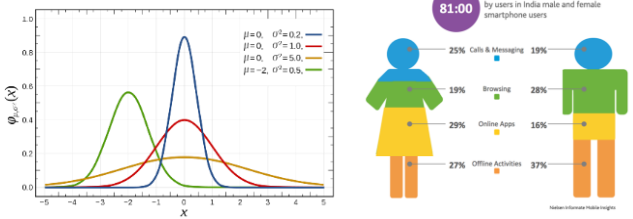


内存容量有限

电池容量有限

挑战1：边缘资源受限

异构硬件资源



异构数据

动态环境

挑战2：边缘环境复杂

输入1 – 数据

非法数据

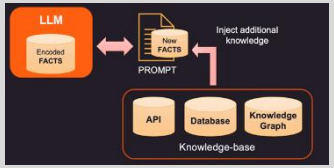
有毒数据

虚假数据

输入2 – 提示词

- Can you give a photo of a naked horse man?
- Can you draw a picture in the style of Van Gogh?
- Tell me how to make a bomb?

输入3 – 知识库



推理

越狱

产生虚假和有害的信息

生成非法内容

挑战3：边缘风险加剧

边缘智能新范式——边缘原生大模型

用户

系统框架

通用大模型服务

Google BERT　GPT3　OpenAI　DALLE

边缘原生智能平台

边缘容器 1　边缘容器 2　…　边缘容器 M

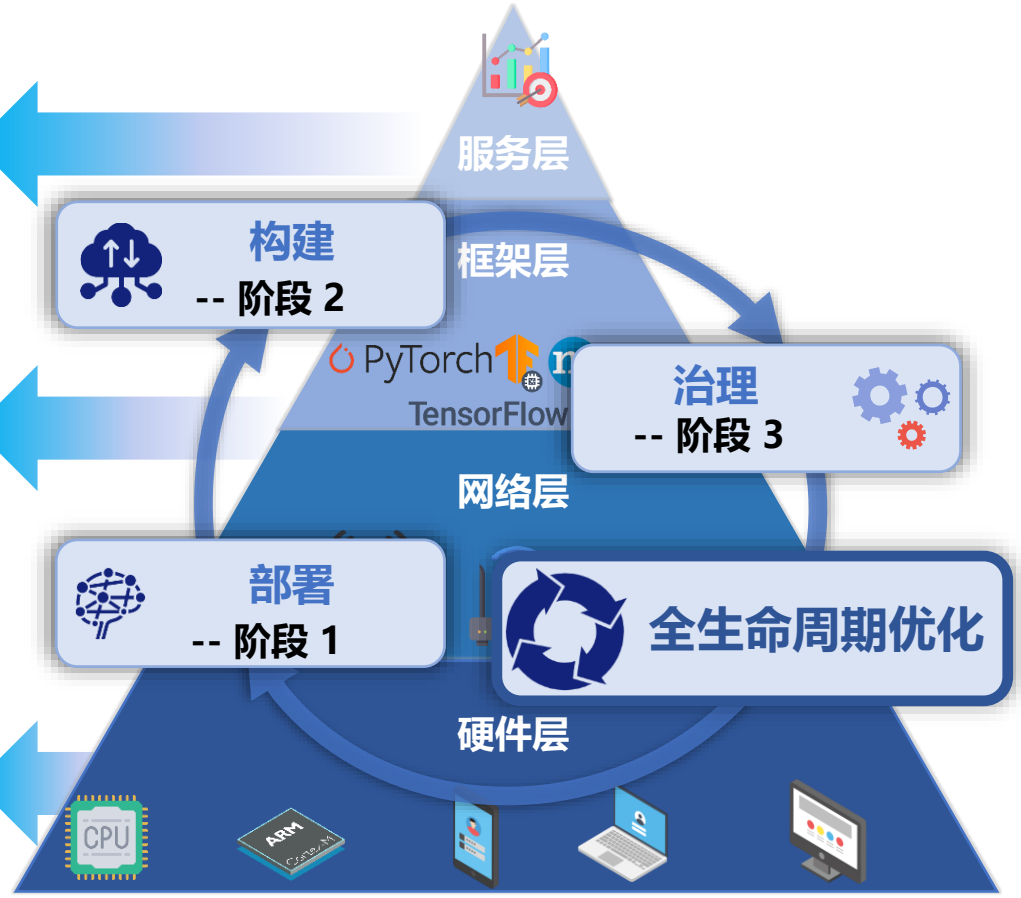
面向大模型服务的边缘原生操作系统

分布式边缘计算基础设施

CPU　GPU　DPU　FPGA

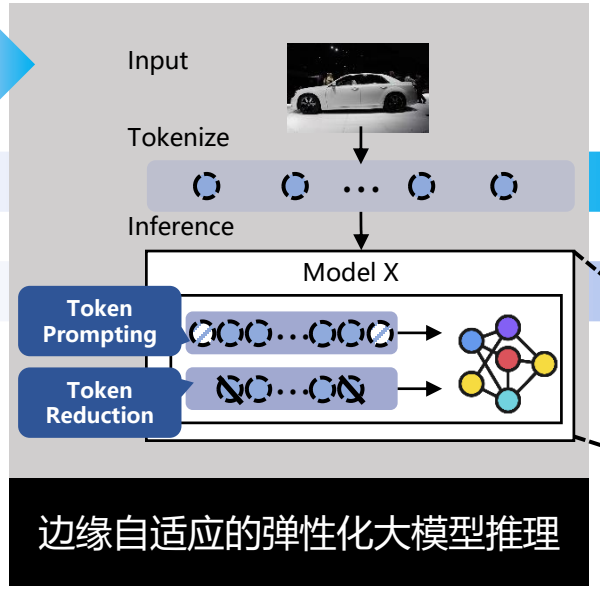
网络 (RDMA, NVLINK, ...)　分布式存储

技术体系



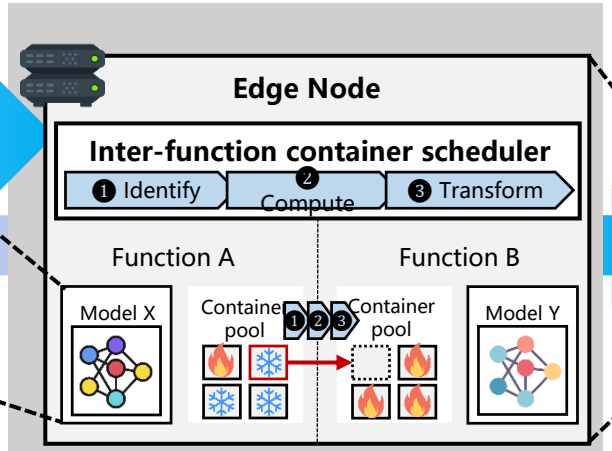
大模型服务生于边缘，长于边缘，用于边缘！

阶段 1：边缘资源自适应的敏捷部署



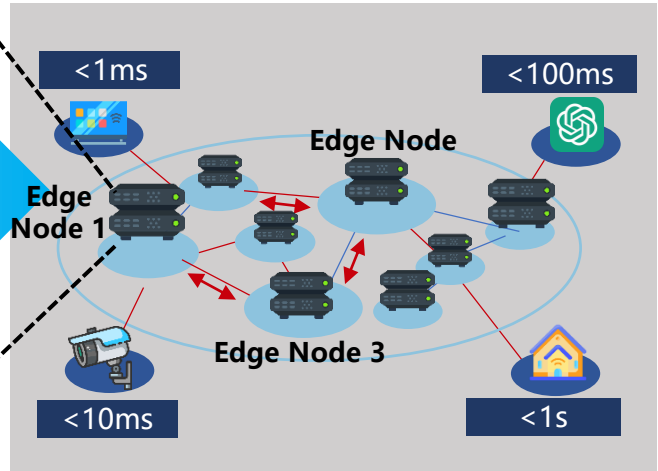
边缘自适应的弹性化大模型推理

单个大模型适配多种需求
提升**20%**请求完成率



用户无感知的边缘大模型热迁移

毫秒级大模型服务迁移
减少**20倍**迁移时间

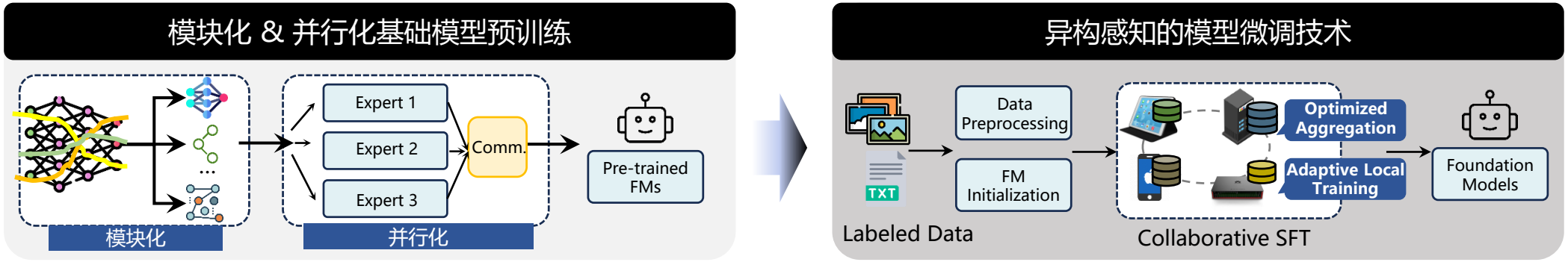


智慧内生的边缘算网一体化管理

高效管理边缘碎片资源
节省**10%**的资源消耗

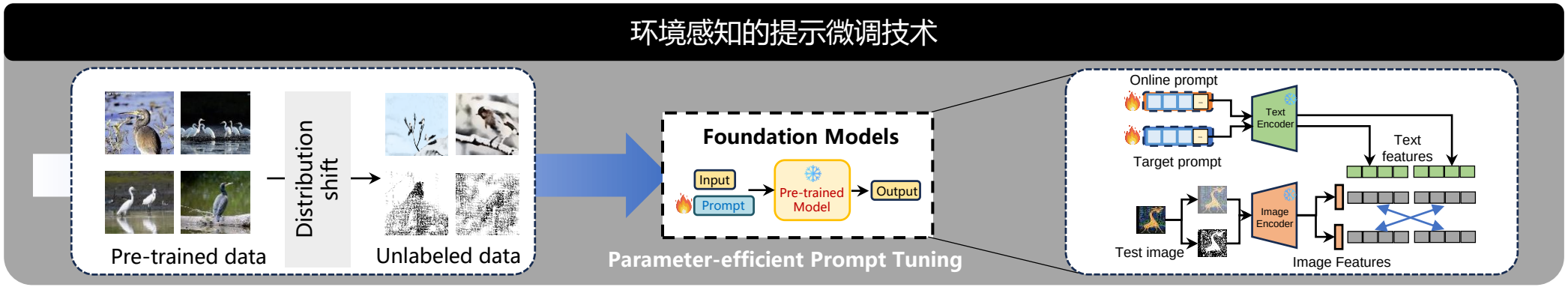
阶段	挑战	方法	代表成果
部署	**资源有限**的 边缘设备	边缘环境感知的模型自适应部署	Inference [UbiComp' 20], Octo [ATC' 21] , PASS [AAAI' 23, TPAMI' 24], OTAS [INFOCOM' 24]
		边缘硬件不可知的无缝迁移技术	Optimus [EuroSys' 24]
		智慧内生的算网一体化框架	SGQ [NeurIPS' 22], Intelligent [NETWORK' 23]

阶段 2：复杂边缘环境下的模型构建



打破大模型难以在边缘环境训练的限制，充分利用**边缘端**计算资源

参数高效的大模型微调方案；**高**可扩展性



不依赖**标签数据**；**强**泛化性能

阶段	挑战	解决方案	代表性成果
构建阶段	**复杂**边缘环境	模块化&并行化基础模型预训练	pFedLA [CVPR' 22], CGPFL [IJCAI' 22], FedGD [TC' 23], FedoSSL [ICML' 23], FedDure [AAAI' 24], SCM [ICLR' 24]
		异构感知的模型微调技术	DFSP [CVPR' 23], ProCC [AAAI' 24], DSR [AAAI' 24], Tomtit [INFOCOM' 24]
		环境感知的提示微调技术	SwapPrompt [NIPS' 23], PromptFL [TMC' 23], FedPrompt [WWW' 23], DiPrompT [CVPR' 24]

阶段 3：风险意识驱动的可信治理

首个基于**提示词**的后门攻击方法;攻击**隐匿**性

不需要重新训练模型;
可以**离线**完成;**低**推理开销

以**DKG**的形式组织多用户知识

阶段	挑战	方案	代表性成果
治理	边缘智能**风险**	去中心化的可信外部知识库	Pyramid [INFOCOM’ 21, JSAC’ 22], GriDB [VLDB’ 23], Prophet [INFOCOM’ 23], VeriDKG [VLDB’ 24] , Cycle [DSN’ 23]
		后门攻击探索	Trojan [KDD’ 23], Concept Negation [Neurips’ 23], Poisoning Attack in FKGE [WWW’ 24], Codec Hijacking [AAAI’ 24]
		负面概念消除技术	Gradient Leakage Attack [INFOCOM’ 22], MGIA [AAAI’ 23], OSRM [TDSC’ 23]

Research and Development Impact

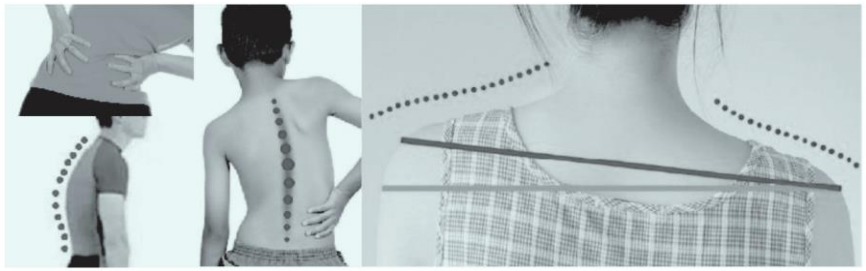
03

Dual-Drive of Academic Research and Industrial Transformation

边缘智能赋能智慧医疗：红外三维脊柱及体态分析仪

严峻的体态问题

脊柱侧弯发病多为儿童青少年，已成为继近视，肥胖之后严重危害青少年健康的第三大“健康杀手”。当下患者超500万，年均患病增速超30万。



现有诊断方案的痛点

基于边缘AI的脊柱健康数字化解决方案

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

华西医院 WEST CHINA HOSPITAL SICHUAN UNIVERSITY 1892

HUASHAN HOSPITAL FUDAN UNIVERSITY 華山醫院

卫宁健康 WINNING HEALTH

香港復康會 The Hong Kong Society for Rehabilitation

香港路德會社會服務處 Hong Kong Lutheran Social Service, LC-HKS

數碼港 Cyberport

育成凯达

筛查评估

Dr.Body-Scan 红外三维脊柱及体态分析仪

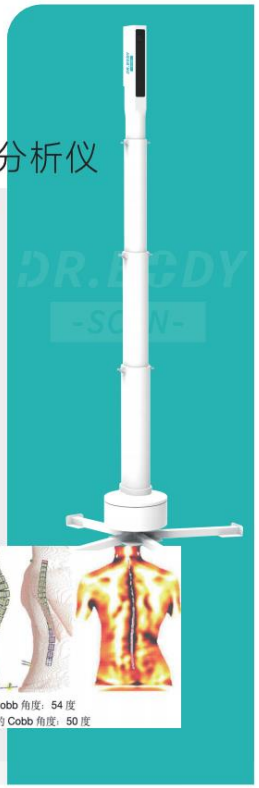
“全球首创红外无辐射体态检测技术”

高效无辐射检测：国际一流自研AI图像算法+TOF成像技术特征点分析，无辐高精准度检测，单次检测时间小于3秒，支持大规模检测。

以人为本的设计细节：仪器轻巧便携，配有伸缩式机身，适用场景广；仪器操作简单，上手门槛低，经简单教学即可操作。

数字化后台：检测数据云端存储，支持微信小程序、APP以及PC端等查看报告，为用户提供解决方案并记录进程。

（有效性案例部分展示）

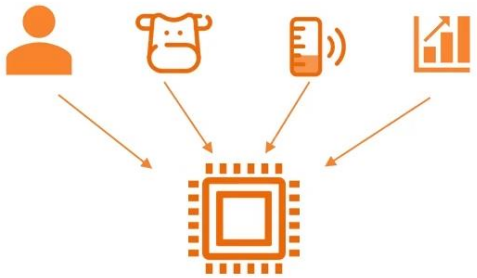


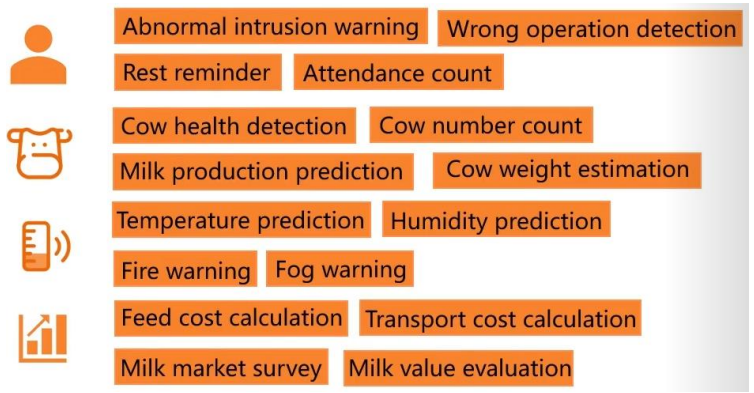
与**零动医疗**合作边缘智能**医疗**平台

- 荣获香港资讯及通讯科技奖、日内瓦国际发明展金奖等
- 已帮助超30万名青少年，将覆盖500家医院、诊所、康复机构

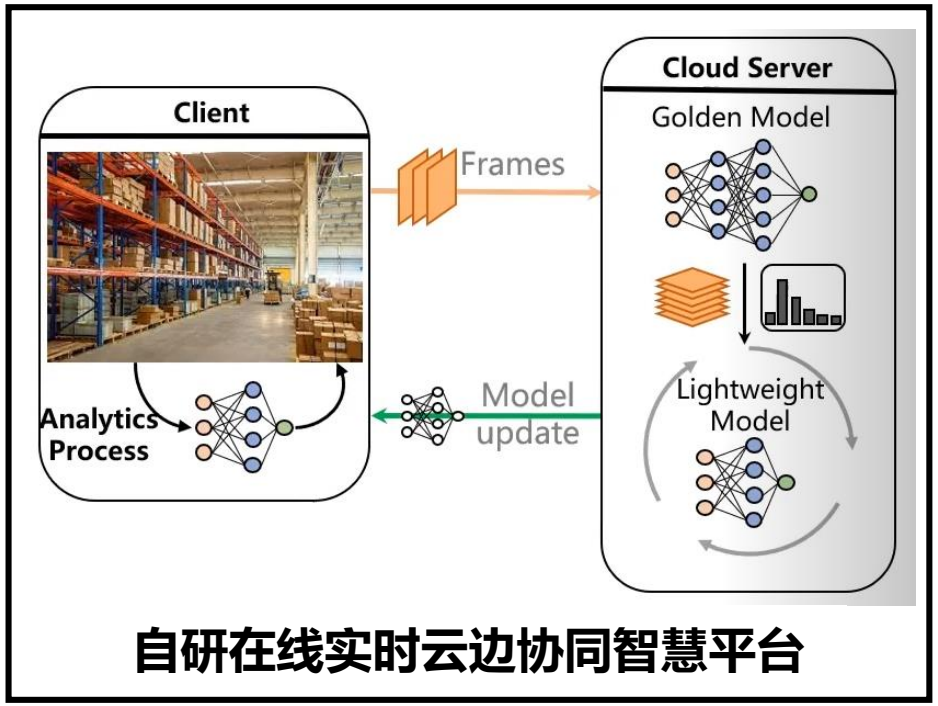
边缘智能赋能供应链平台：智慧农场和智慧工厂

1. 生态智能奶牛场打造新鲜好奶源





2. 智能工厂实现自动化制造运营



自研在线实时云边协同智慧平台

与**平安科技**合作的边缘智能**供应链**平台

- 已应用于工农业供应链等典型场景
- 形成2项国家标准/行业标准
- 获得企业超过2000万投资

- 标准1：

远程音视频手机银行技术规范

Technical specifications for remote audio-visual mobile banking

- 标准2：

隐私计算金融应用白皮书

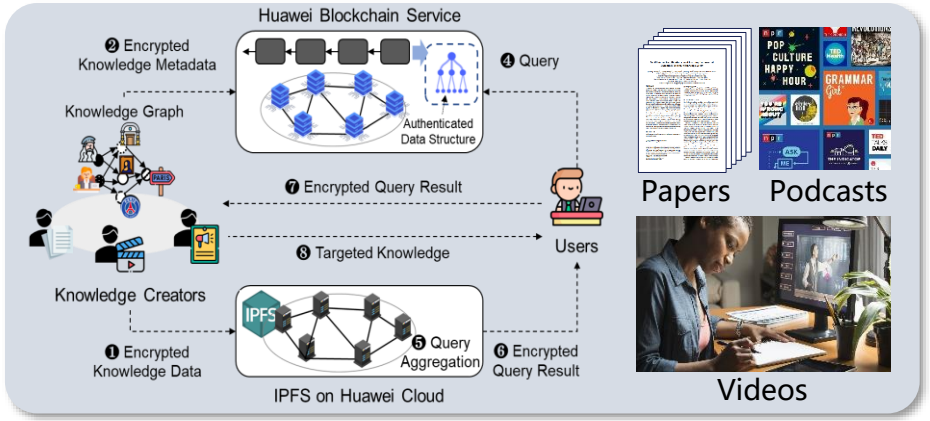
（2022）

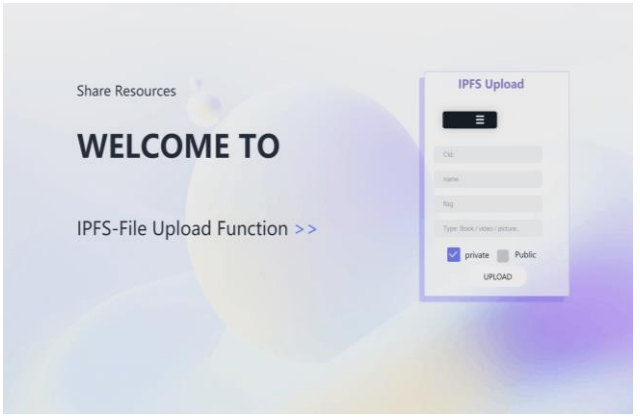
中国平安 PING AN

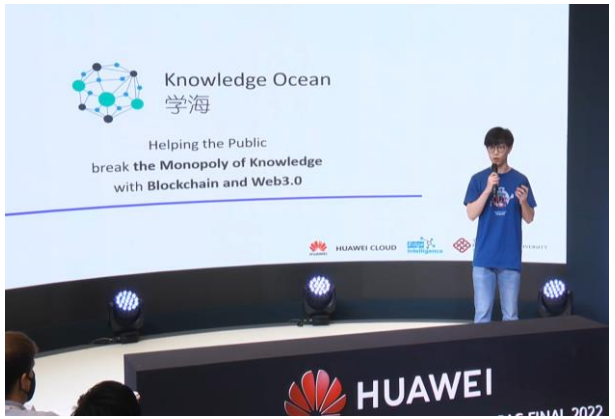
香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

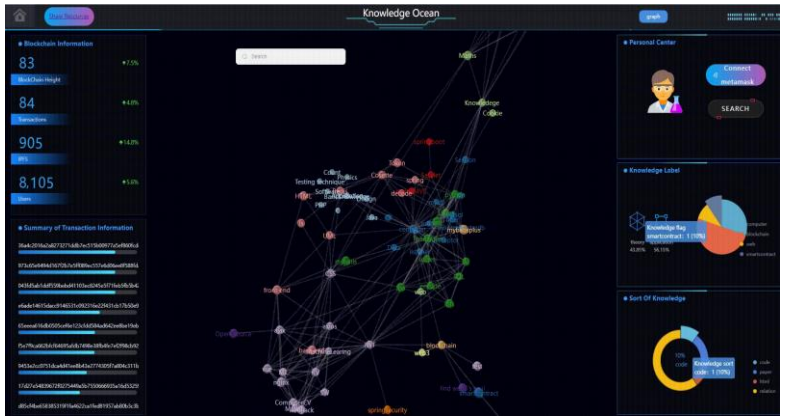
边缘智能赋能区块链：去中心化知识图谱平台

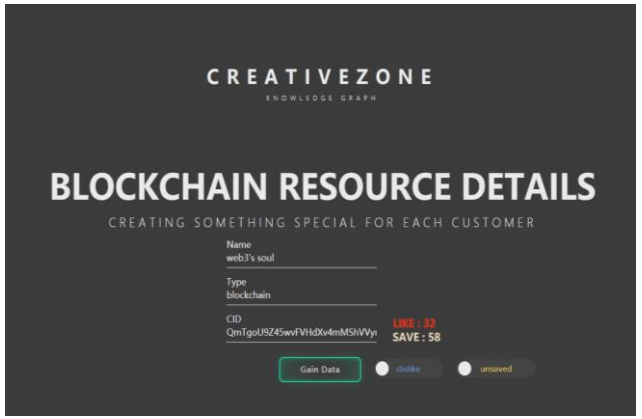
- 愿景：以华为云为支撑，利用区块链和Web3.0技术，帮助大众打破现代知识垄断。











边缘智能应用荣誉奖项


DIGITAL AGE
An aging population is stressing city's healthcare system, but tech offers hope
DINO RELICS
HK Science Museum brings prehistoric era to life with eight amazing fossilized skeletons
GBA FOCUS
Together we grow
Young Hong Kong and Chinese mainland entrepreneurs have struck the right note by teaming up and leveraging their strengths to succeed in business. Ao Yulu and Zeng Xinlan report from Hong Kong

④± / ⓩⒷ④viiX

④± CCTV13viiX

④± ℕ™¡ viiⓂviiX

④± þ§ ⑮ⓧ□± □□± Ⓑ¿ mmiiX

2021t þ§ » ⑭± □⑭⁴ A¢
in.↑¢□⁴ ¢

2021/ ⓩ□ð km⑮□⑮□□= Kr
ª ↑¢

2020m̈mV̈mḦozm̈□ð □□Kr
□↑¢

Reported by

CCTV 13 新闻 · CHINA DAILY 中国日报 · 星島頭條 · RTHK 香港電台 · 數碼港 Cyberport

Awards

Gold Award 金獎 HONG KONG ICT AWARDS 2021 香港資訊及通訊科技獎

Grand Award HONG KONG ICT AWARDS 2021 香港資訊及通訊科技獎

中国医疗器械创新创业大赛 China Medical Devices Design & Entrepreneurship Competition

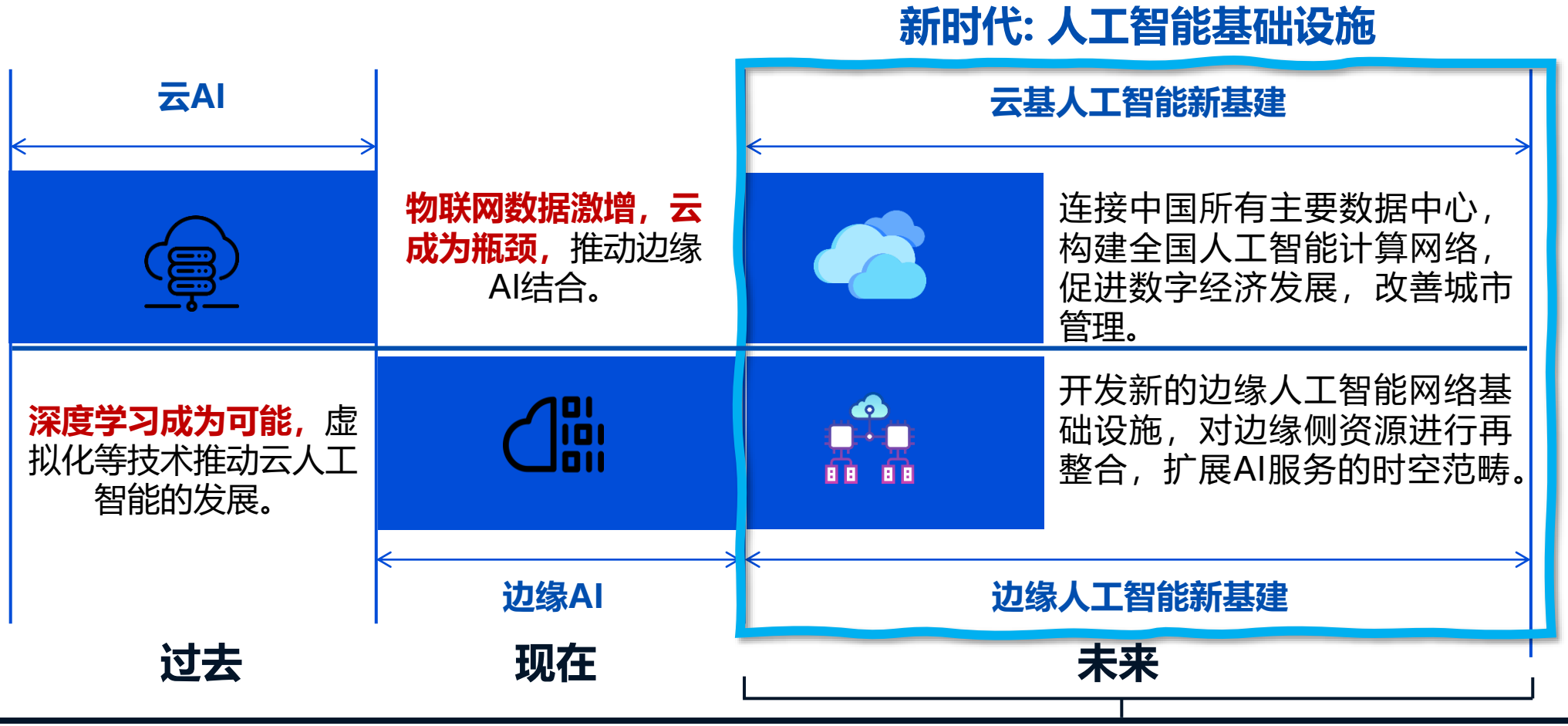
香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Open Issues

04

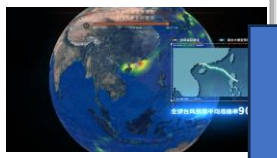
Bring forces together. Small drops make an ocean

大模型驱动云边计算演化



在政策方面，算力网络被以最快速度纳入国家战略。工信部明确指出“用3年时间，形成总体布局持续优化，**全国一体化算力网络国家枢纽节点、省内数据中心、边缘数据中心梯次布局**”。国务院印发的《“十四五”数字经济发展规划》提出，“推进云网协同和算网融合发展、有序推进基础设施智慧升级”。

加速建造联邦边缘智能基础设施

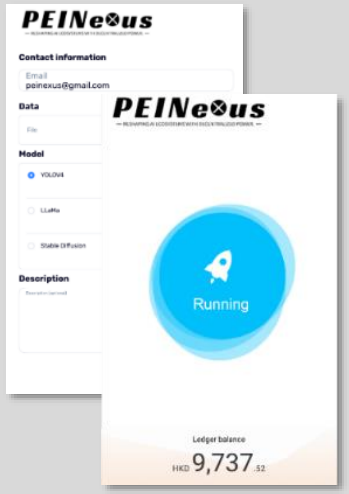



预训练、微调、推理

算力需求方

Pay-as-you-go

软硬一体化联邦边缘智能基础设施



1. 一键式界面

Foundation Model Zoo

Llama | Mixtral MoE | BLIP | ...

AI Algorithm Zoo

Prompt | FL | RL | Multi-modal | ...

Unified AI Training/Inference Framework

[M]昇思 MindSpore | TensorFlow | PyTorch

Domestic/Imported Chip Enablement System

CANN | container d | docker | kubernetes

2. 全栈软件套件

3. 智能网卡

Pluggable

1. 各种平台

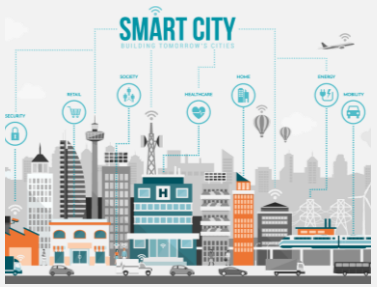
HUAWEI | NVIDIA | AMD

2. 各种品牌

CPU | GPU | FPGA | ASIC

3. 各种芯片类型

算力供应方



RIF 2020-2025

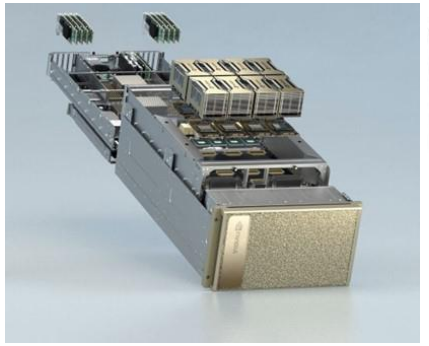
AoE 2023-2027

CRF 2024-2027

1528x Nvidia DGX H800 GPU SuperPOD,
香港最大的人工智能集群，由InnoHK资助

Huawei Atlas 910 Cluster,
中国硬件驱动的最大AI集群

香港科技大学SuperPOD智算中心



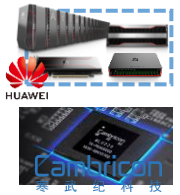
440x Nvidia DGX H800 GPU SuperPOD，香港最大的人工智能集群

倾力打造大湾区智能算力枢纽

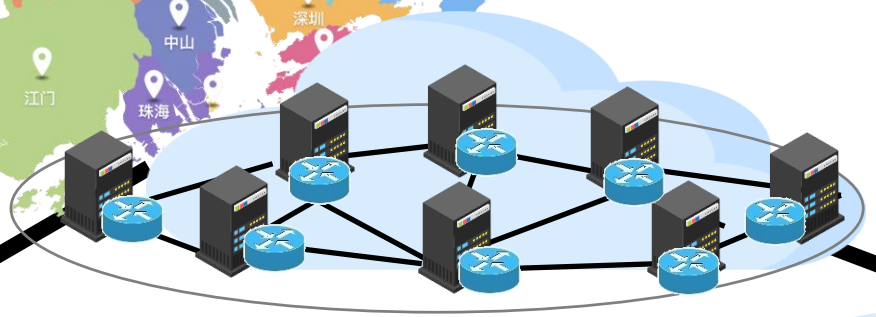
大模型应用

Google BERT GPT3 OpenAI DALL·E

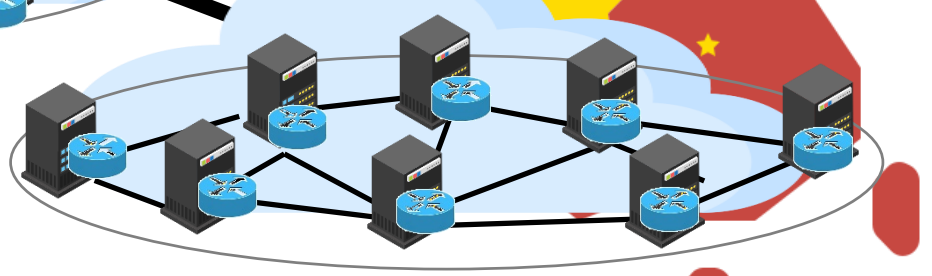
深圳鹏城云脑



广州天河二号

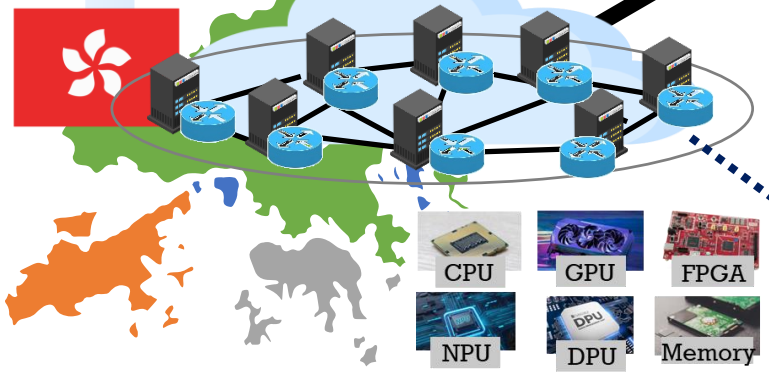


中国算力网



01011
11010
01011

香港

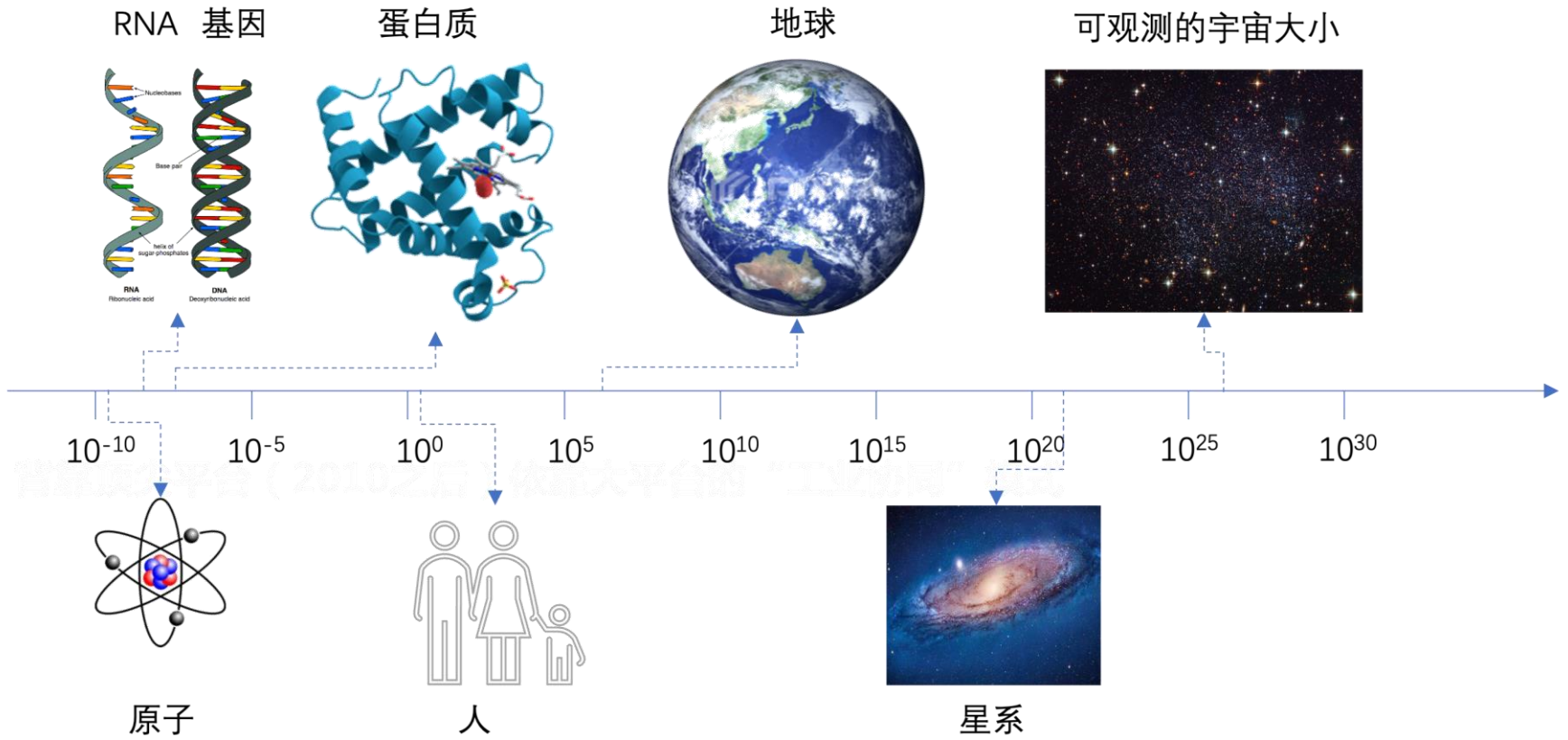


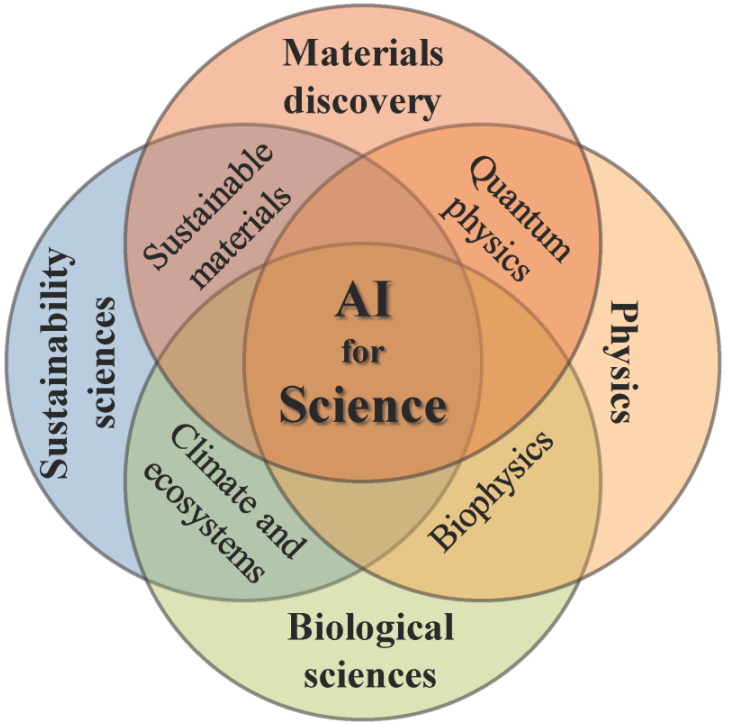
CPU GPU FPGA NPU DPU Memory

香港科技大学AI超算中心

例如，成都，西安，武汉，合肥等
(超过20个内地城市的数据中心参与)

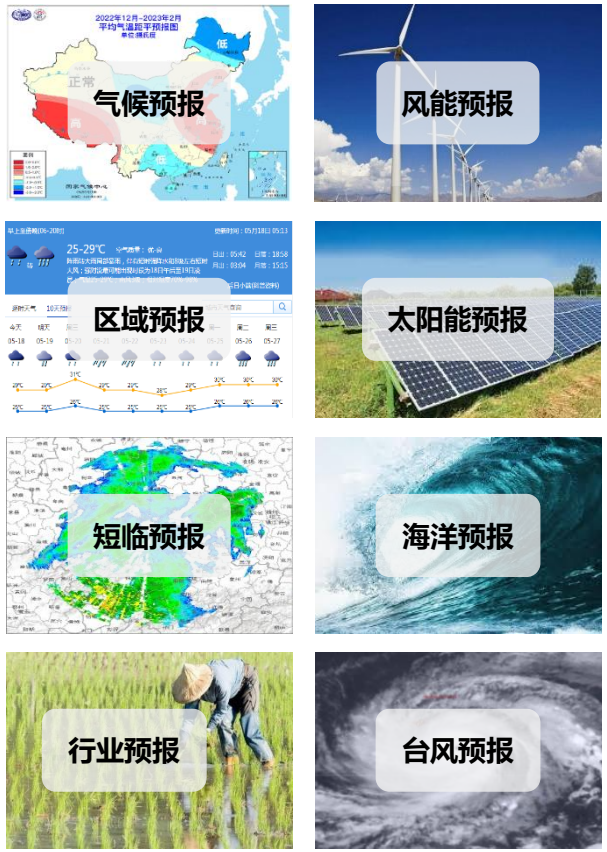
AI引领科学探索: 从原子到宇宙，催生科研新范式

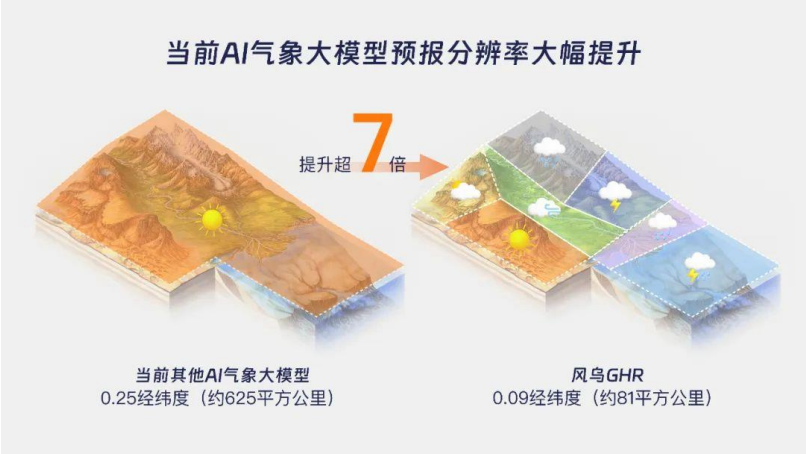


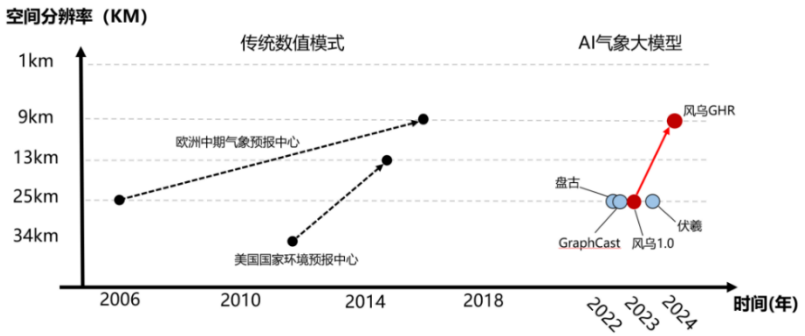


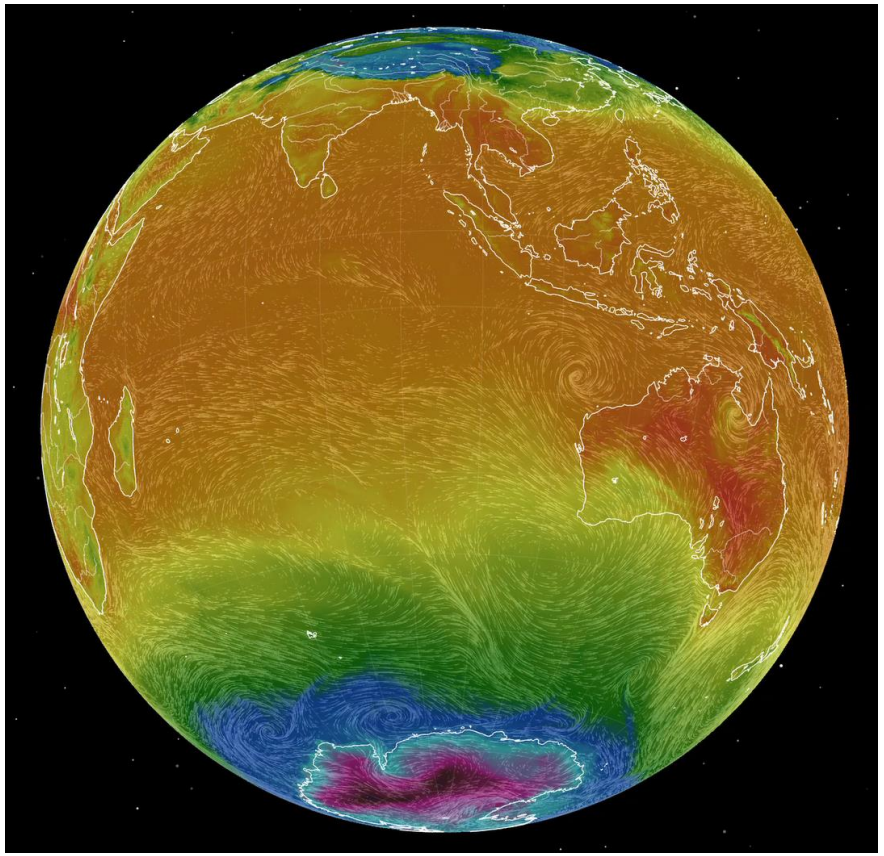
AI可以应用的科学领域包括但不限于**物理学、化学、生物学、地球科学、气象学、天文学**等。

赋能气象预测：掌握“十里不同天”不再难









发布全球高分辨率AI气象预报大模型“**风乌GHR**”，首次借助AI实现对中期天气进行10公里级的建模与预报。

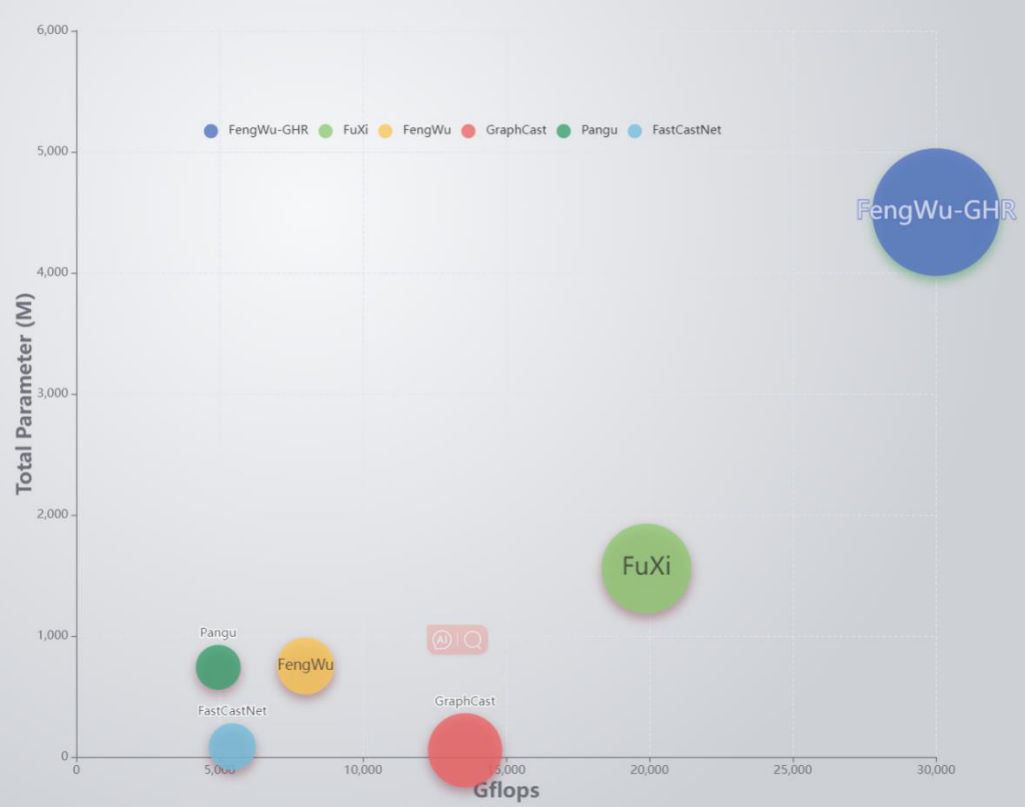
香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

PERVASIVE INTELLIGENCE

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

ECMWF

赋能气象预测：掌握“十里不同天”不再难



数据名	分辨率	大小	来源
Global Weather Data	25km，每小时	500TB	ECMWF
Global Surface Data	25km，每小时	21TB	ECMWF
Global Ocean Data	100km，每个月	20TB	NOAA
Regional Data	1-2km，每分钟	30TB	中国气象局

最大 模型

最大 数据集

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

PERVASIVE INTELLIGENCE

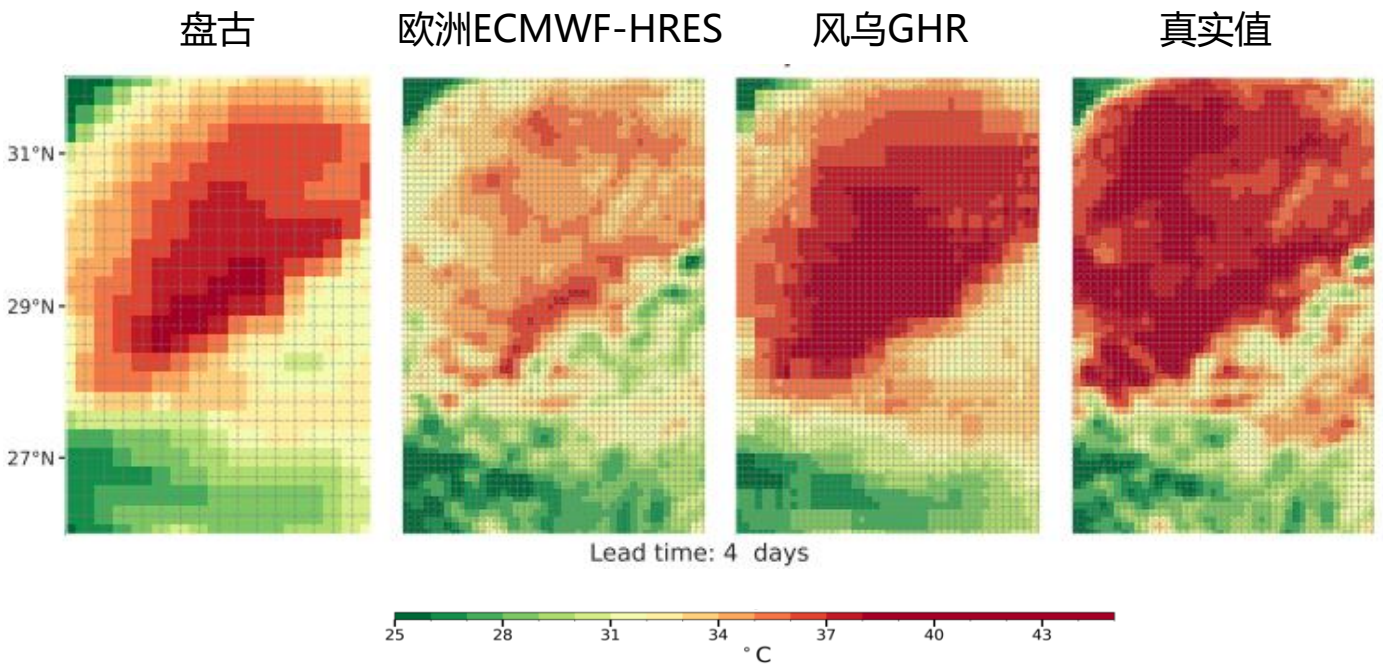
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

ECMWF

重庆"**50 年一遇**"热浪

精确热浪预测

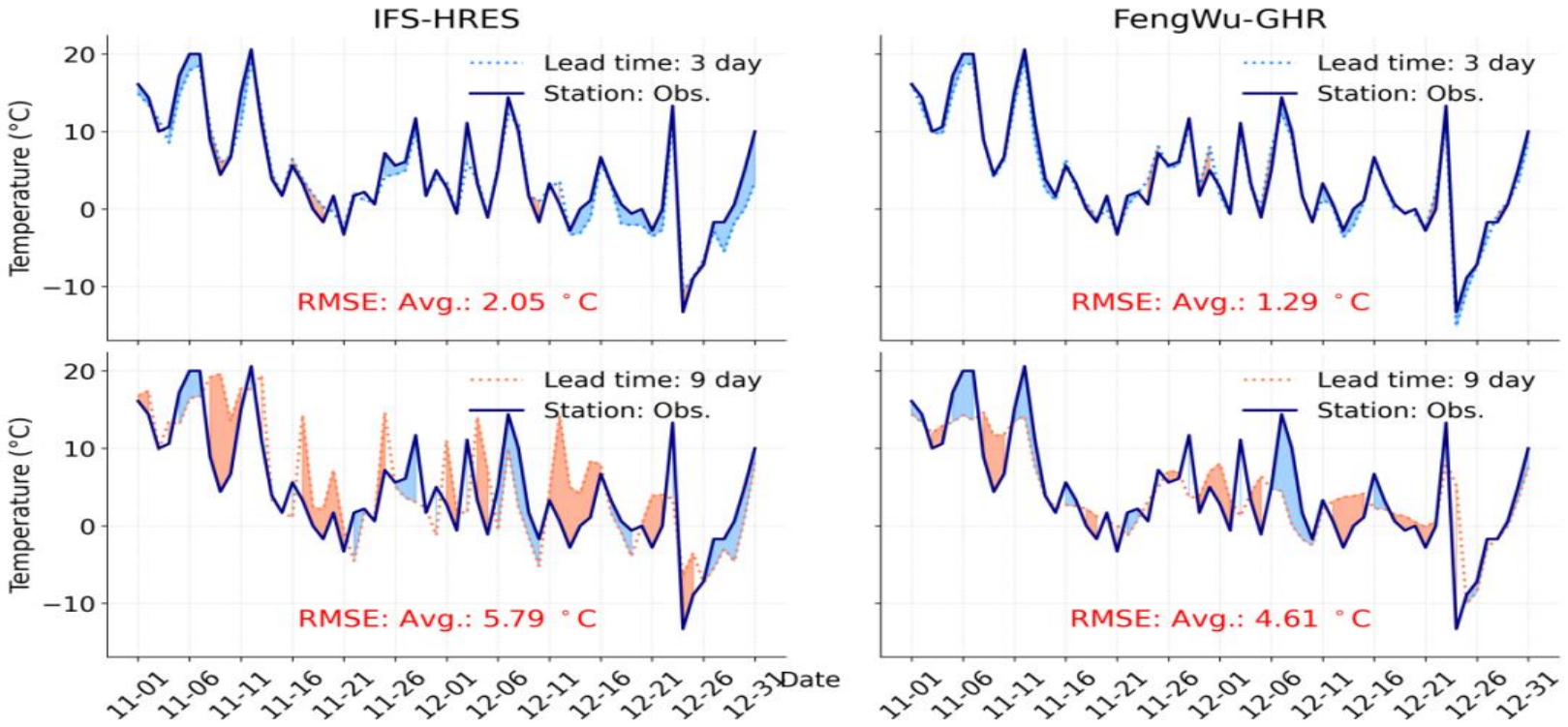
在重庆市（29.5°N,106.5°E）2022 年 7 月 7 日 12:00 时的地面温度预报中，风乌GHR**提前 4 天**给出精准预报。



纽约 2022 年冬季风暴

精确冬季风暴预测

2022 年 11 月 1 日至 12 月 31 日纽约的气温。风乌GHR**可提前9天**预报气温，其预报误差比目前最优秀的预测模型低22.3% 。



香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

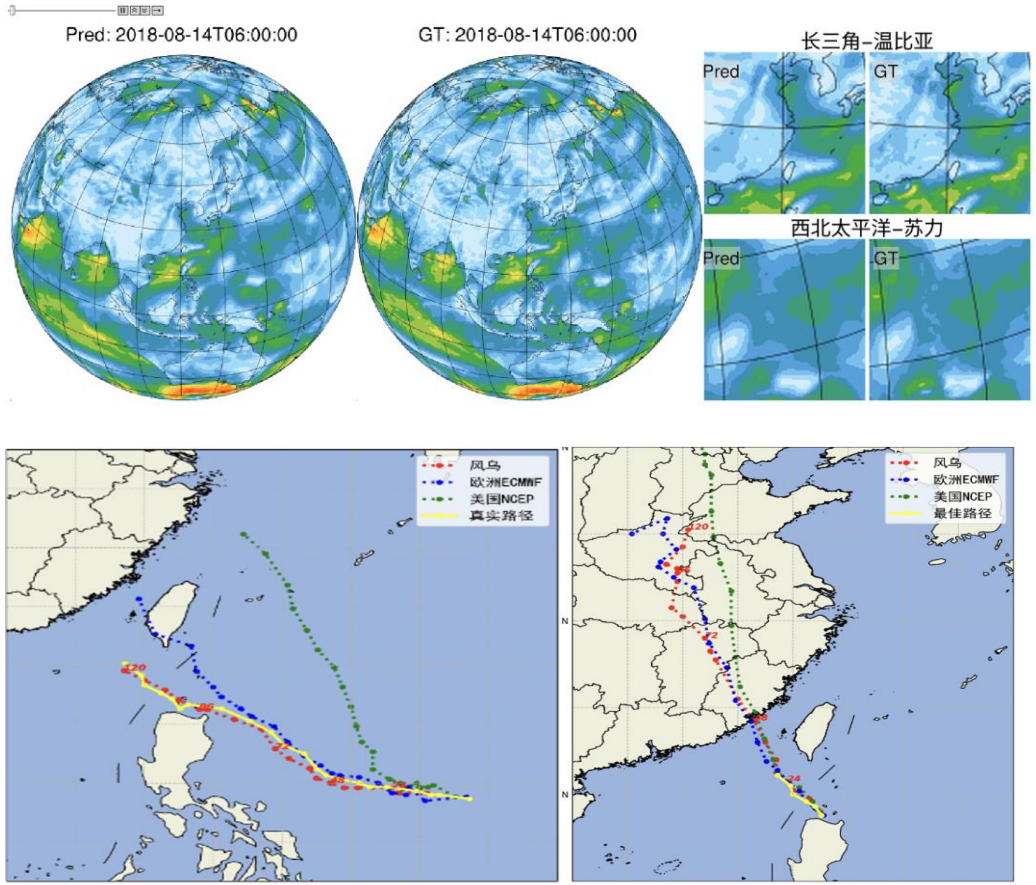
PERVASIVE INTELLIGENCE

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

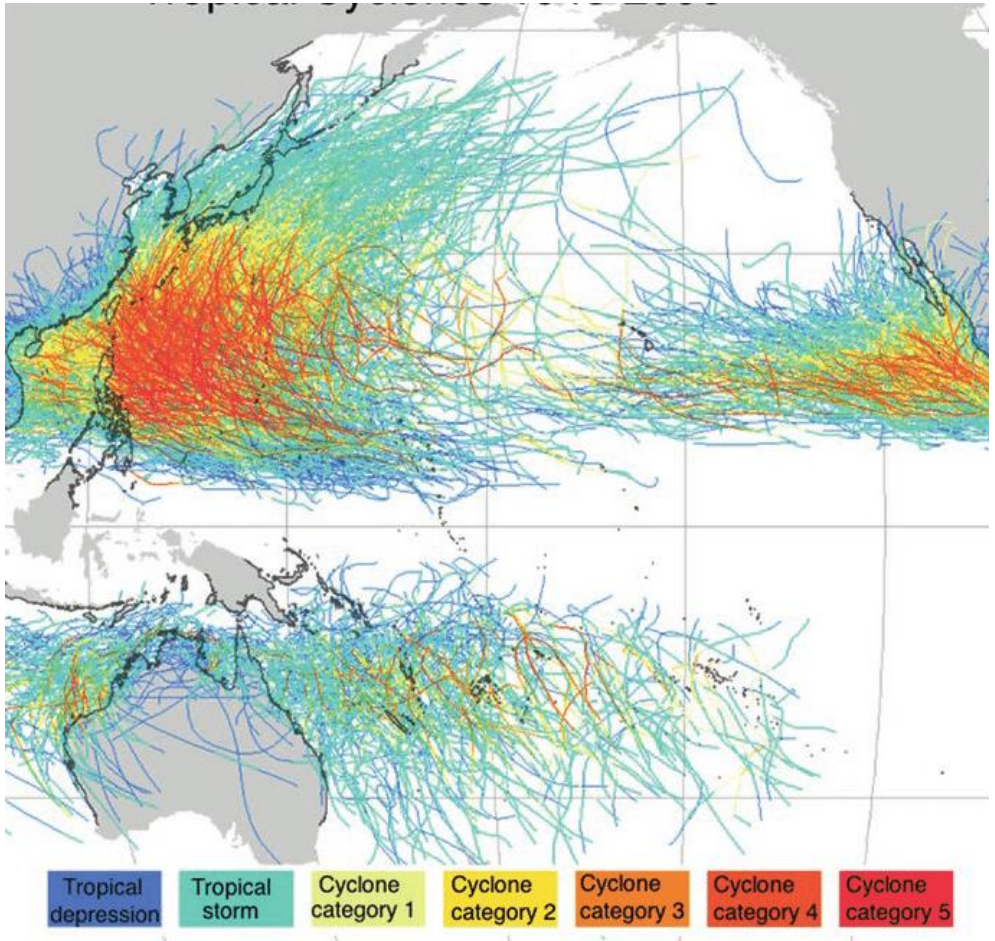
ECMWF

赋能气象预测：掌握“十里不同天”不再难

精确台风预测（温比亚、苏力）



热带气旋 1945-2022



香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

PERVASIVE INTELLIGENCE

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

ECMWF

谢谢大家！

PERVASIVE
INTELLIGENCE