Title: Huawei PanGu

URL: https://en.wikipedia.org/wiki/Huawei_PanGu

PageID: 76087358

Categories: Category:2023 in artificial intelligence, Category:2023 software, Category:Generative pre-trained transformers, Category:Huawei products, Category:Large language models, Category:Multimodal interaction

---

Huawei PanGu , PanGu , PanGu-Σ or PanGu-π (Chinese : ■■■■■ ; pinyin : páng■ dà móxíng) is a multimodal large language model developed by Huawei . It was announced on July 7, 2023. [1]

The name of the large learning language model, PanGu , was derived from the Chinese mythology and folklore of Pangu , a primordial character related to the creation of the world. [2]

History

Early development

In April 2023, Huawei released a paper detailing the development of PanGu-Σ, a colossal language model featuring 1.085 trillion parameters. Developed within Huawei's MindSpore 5 framework, PanGu-Σ underwent training for over 100 days on a cluster system equipped with 512 Ascend 910 AI accelerator chips, processing 329 billion tokens in more than 40 natural and programming languages. [3]

PanGu-Σ incorporates Random Routed Experts (RRE) and the Transformer decoder architecture, allowing easy extraction of sub-models for various applications like conversation, translation, code production, and natural language interpretation. The model achieves 6.3 times faster training throughput compared to MoE models with the same hyper-parameters. In the Chinese domain, it outperforms previous state-of-the-art models across 16 tasks in a zero-shot setting. Trained on datasets from 40 domains, including Chinese, English, Bilingual, and code, PanGu-Σ excels in few-shot natural-language understanding , open-domain discussion, question answering, machine translation, and code creation. [4] [5]

Launch

During the Huawei Developer Conference on July 7, 2023, Huawei introduced PanGu 3.0, a large language model (LLM), tailored for sectors like government, finance, manufacturing, mining, and meteorology utilizing Huawei Cloud [zh] solutions. In the subsequent month, Huawei launched the Celia Virtual Assistant with advanced AI features, capable of generating long text replies based on user voice commands and set to release with HarmonyOS 4.0 for eligible devices. [6] [7]

The LLM was designed for enterprises seeking advantages in the AI industry, focusing on task execution over creative work, unlike traditional models used for general purposes like chatbots, poetry, and visual content creation. [8]

Using the same technology as ChatGPT , Huawei's LLM features a hierarchical architecture, allowing customers to adapt the model to various tasks and train it on their own datasets, making it versatile across various industries. [9]

Updates

On August 5, 2023, Huawei partnered with European Centre for Medium-Range Weather Forecasts (ECMWF) to launch a global weather forecasting AI model. This model used Huawei Cloud solutions and the PanGu-Weather Model with MindSpore . It is accessible on the ECMWF website and aims to provide accurate weather data. [10] [11]

On December 19, 2023, Huawei announced its financial services on the PanGu-powered AI Finance platform for the global market. The tech giant introduced this product at the 2023 Huawei Cloud Fintech Summit, aiming to reshape the digital finance industry with efficient features to boost

Fintech firms worldwide. The platform incorporated a variety of advanced technologies, including AI, big data analytics, and blockchain. [12]

On June 21, 2024, at HDC 2024, Huawei announced upgraded PanGu 5.0 alongside HarmonyOS NEXT . This version integrated with Harmony Intelligence , which features a smarter Celia (Xiaoyi) and focuses on generative AI updates to its LLM platform for creating new content, such as text, code, or images. Aiming to make PanGu accessible to a wide range of developers and businesses, it offered scalable options: smaller models requiring less computational power for those with limited resources, and larger models with increased capacities for complex tasks requiring more processing power. [13]

Technical specifications

PanGu Large Model 3.0, designed for industry use, was structured with a 5+N+X three-tier architecture. [14]

First Layer (L0): Comprises PanGu's five basic large models to provide a variety of capabilities for different industry scenarios. These include Natural Language Processing (NLP) models, Visual models, Multimodal models, Prediction models, and Scientific Computing models.

Second Layer (L1): Consists of N large industry-specific models. These models are trained using public data from various industries, such as government, finance, manufacturing, mining, and weather. Additionally, it uses customers' own data from L0 and L1 to train proprietary models tailored for each customer.

Third Layer (L2): Provides customers with detailed scenario-specific models. This layer focuses on specific applications or business needs, offering ready-to-use model services.

The updated Huawei PanGu Model 5.0 by Huawei Cloud business division offered three key features: adaptability for different business scenarios, multi-style modeling, and advanced intelligence. Huawei divided the AI model platform into four series, each with different parameter scales: [15]

PanGu E Series: The Embedded version supports smart apps on phones, tablets, PCs, and other devices, with a parameter scale of 1 billion.

PanGu P Series: The Professional version features a 10-billion parameter scale, ideal for low-latency and low-cost reasoning conditions.

PanGu U Series: The Ultra version comes in two variants, with 135 billion and 230 billion parameters, capable of handling complex tasks and serving as a base for large models.

PanGu S Series: The Super PanGu is the top-tier edition, featuring trillion-level parameters, designed to manage advanced AI technology scenarios such as cross-domain or multi-tasking applications.

Controversy

On July 4, 2025, some researchers alleged on GitHub that there is an extremely high similarity in the attention parameter distribution between the Pangu Pro MoE model and Alibaba 's Qwen model, using "model fingerprinting" technology. The next day, Huawei Noah's Ark Lab, the development team, responded that Pangu is a foundational large model self-developed on Ascend hardware and not incrementally trained on other models. They added that they had made compliant attributions in strict accordance with open-source licenses, a common practice in the community. The original repository with the accusation has since been deleted. [16] [17] [18]

See also

Large language model

Gemini

GPT-4

References

v

t

e

Huawei Ascend

Ascend P P1 P2 P6 P7

P1

P2

P6

P7

Ascend D D quad quad XL D1 1 XL D2

D quad quad XL

quad XL

D1 1 XL

1 XL

D2

Ascend G G6 G7 G300 G312 G330 G600 G750

G6

G7

G300

G312

G330

G600

G750

Ascend Mate Mate 2 Mate 7

Mate 2

Mate 7

W1

W2

Ascend Y300

Huawei P8 P8 Max

P8 Max

P9 9 Plus 9 lite

9 Plus

9 lite

P10

P20

P30

P40 40 lite

40 lite
P50 50 Pocket
50 Pocket
P60
Pura 70 70 Pro 70 Pro+ 70 Ultra
70 Pro
70 Pro+
70 Ultra
Pura X
Huawei Mate S
8
9 9 lite
9 lite
10
SE
20
30
40
50
60
70
X
Xs
X2
Xs 2
X3
X5
X6
XT Ultimate Design
Huawei Nova & Nova Plus lite 2017 lite+ 2 3 3+ Smart Young Youth
lite 2017 lite+ 2 3 3+
2017
lite+
2
3 3+
3+
Smart
Young

Youth

2 2 Plus 2i 2 lite 2s

2 Plus

2i

2 lite

2s

3 3e 3i

3e

3i

4 4e

4e

5 5 Pro 5i 5i Pro 5T 5z

5 Pro

5i

5i Pro

5T

5z

6 6 SE

6 SE

7 7 Pro 7i 7 SE Youth

7 Pro

7i

7 SE Youth

Youth

8 8 Pro 8 SE Youth 8i

8 Pro

8 SE Youth

Youth

8i

9 9 SE

9 SE

10 10 Pro 10 SE 10z

10 Pro

10 SE

10z

11 11 Pro/Ultra 11 SE 11i

11 Pro/Ultra

11 SE

11i

12 12 Pro/Ultra 12 SE 12i 12s/12 Lite

12 Pro/Ultra

12 SE

12i

12s/12 Lite

13 13 Pro 13i

13 Pro

13i

Flip

Y60

Y61/Y62/Y62 Plus

Y70/Y70 Plus/Y71

Y72 Y72S

Y72S

Y90

Y91

Huawei G7 Plus/G8

G9 Plus 9 lite

9 lite

G Play Mini

Mini

GT3

GR3 2017

2017

GR5 2017 5 mini

2017

5 mini

Y3 (Y360) 3 II 2017 2018

3 II

2017

2018

Y5 (Y560) 5c 5 II 2017 2018 Prime lite 2019 5p

5c

5 II

2017

2018 Prime lite

Prime

lite

2019

5p

Y6 6 Pro 6 II Compact 2017 Pro 2018 Prime 2019 Prime Pro 6s 6p

6 Pro

6 II Compact

Compact

2017 Pro

Pro

2018 Prime

Prime

2019 Prime Pro

Prime

Pro

6s

6p

Y7 7 Prime 2018 Prime Pro 2019 Prime Pro 7p 7a

7 Prime

2018 Prime Pro

Prime

Pro

2019 Prime Pro

Prime

Pro

7p

7a

Y8 8s 8p

8s

8p

Y9 2018 2019 Prime 9s 9a

2018

2019 Prime

Prime

9s

9a

Y Max

Y625

Y635

Huawei G6600 Passport

M835

Pocket 2

Pocket S

Premia

Sonic

STREAM X GL07S

T156

T158

T161L

T201

T208

T211

T261L

T300

U120

U121

U1000

U1100

U1250

U1270

U1310

U2801

U3300

U7310

U7510

T-Mobile Tap (U7519)

U8100

U8110

IDEOS U8150

T-Mobile Pulse (U8220)

U8230

U8800

U9130 Compass

U9150

Nexus 6P

Huawei Ideos Tablet S7

Honor Tablet 5

Mediapad M5

Mediapad M6

MatePad Pro

MateBook

MateBook X Pro

Watch GT

X Gentle Monster Eyewear

Watch GT 2 2e 2 Pro

2e

2 Pro

Watch Fit

X Gentle Monster Eyewear II

Kirin

Kunpeng , Ascend

Atlas

Tiangang (5G)

Balong (modem: 5G/5.5G)

EulerOS

openEuler

UniProton

NestOS

HarmonyOS NEXT kernel ; OpenHarmony version history

NEXT

kernel ;

OpenHarmony

version history

LiteOS

EMUI

Huawei VRP

E5

E220

SingleRAN

4G eLTE

Ark Compiler

BiSheng Compiler

DevEco Studio

ArkTS

eTS

Cangjie

Hvigor

EROFS

Huawei PanGu

MindSpore

ArkUI

ArkUI-X

GaussDB

openGauss

ArkData

MetaERP

Celia

Huawei HiCar

NearLink

Softswitches

Next generation home location register

Internet Protocol Multimedia Subsystems

xDSL

Passive optical network

Network switches

Service delivery platforms

Base station subsystems

Mobile Services AppGallery HMS Core Petal Maps

AppGallery

HMS Core

Petal Maps

Network integration

Object storage

Music

Video

Cloud

GameCenter

Themes

Health

Find Device

Ren Zhengfei (CEO)

Liang Hua (Chairman)

Sun Yafang (former chairwoman)

Meng Wanzhou (deputy chair & CFO)
Xu Zhijun (deputy chair)
New IP
Developer Conference
Developer Day
Symantec
HiSilicon
Criticism Li Hongyuan incident
Li Hongyuan incident
Huawei the Beautiful
Nano Memory
FreeBuds
Sound X
Sound Joy
Vision
MateView
MateBook
MateStation
HarmonyOS Sans
Huawei MRP
nweb (ArkWeb) layer
Ox Horn Campus
Intelligent Automotive Solution
Intelligent Mobility Alliance
SiCarrier
Huawei Lianqiuhu
Commons
Category
v
t
e
History timeline
timeline
Companies
Projects
Parameter Hyperparameter
Hyperparameter
Loss functions

Regression Bias–variance tradeoff Double descent Overfitting

Bias–variance tradeoff

Double descent

Overfitting

Clustering

Gradient descent SGD Quasi-Newton method Conjugate gradient method

SGD

Quasi-Newton method

Conjugate gradient method

Backpropagation

Attention

Convolution

Normalization Batchnorm

Batchnorm

Activation Softmax Sigmoid Rectifier

Softmax

Sigmoid

Rectifier

Gating

Weight initialization

Regularization

Datasets Augmentation

Augmentation

Prompt engineering

Reinforcement learning Q-learning SARSA Imitation Policy gradient

Q-learning

SARSA

Imitation

Policy gradient

Diffusion

Latent diffusion model

Autoregression

Adversary

RAG

Uncanny valley

RLHF

Self-supervised learning

Reflection

Recursive self-improvement

Hallucination

Word embedding

Vibe coding

Machine learning In-context learning

In-context learning

Artificial neural network Deep learning

Deep learning

Language model Large language model NMT

Large language model

NMT

Reasoning language model

Model Context Protocol

Intelligent agent

Artificial human companion

Humanity's Last Exam

Artificial general intelligence (AGI)

AlexNet

WaveNet

Human image synthesis

HWR

OCR

Computer vision

Speech synthesis 15.ai ElevenLabs

15.ai

ElevenLabs

Speech recognition Whisper

Whisper

Facial recognition

AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)
Gemma
Grok
LaMDA
BLOOM
DBRX
Project Debater
IBM Watson
IBM Watsonx
Granite
PanGu-Σ
DeepSeek
Qwen
AlphaGo
AlphaZero
OpenAI Five
Self-driving car
MuZero
Action selection AutoGPT
AutoGPT
Robot control
Alan Turing
Warren Sturgis McCulloch
Walter Pitts
John von Neumann
Claude Shannon
Shun'ichi Amari
Kunihiko Fukushima
Takeo Kanade
Marvin Minsky
John McCarthy
Nathaniel Rochester
Allen Newell
Cliff Shaw
Herbert A. Simon
Oliver Selfridge
Frank Rosenblatt
Bernard Widrow

Joseph Weizenbaum
Seymour Papert
Seppo Linnainmaa
Paul Werbos
Geoffrey Hinton
John Hopfield
Jürgen Schmidhuber
Yann LeCun
Yoshua Bengio
Lotfi A. Zadeh
Stephen Grossberg
Alex Graves
James Goodnight
Andrew Ng
Fei-Fei Li
Alex Krizhevsky
Ilya Sutskever
Oriol Vinyals
Quoc V. Le
Ian Goodfellow
Demis Hassabis
David Silver
Andrej Karpathy
Ashish Vaswani
Noam Shazeer
Aidan Gomez
John Schulman
Mustafa Suleyman
Jan Leike
Daniel Kokotajlo
François Chollet
Neural Turing machine
Differentiable neural computer
Transformer Vision transformer (ViT)
Vision transformer (ViT)
Recurrent neural network (RNN)
Long short-term memory (LSTM)
Gated recurrent unit (GRU)

Echo state network

Multilayer perceptron (MLP)

Convolutional neural network (CNN)

Residual neural network (RNN)

Highway network

Mamba

Autoencoder

Variational autoencoder (VAE)

Generative adversarial network (GAN)

Graph neural network (GNN)

Category