GOSIM | 面壁小钢炮 MiniCPM

GOSIM AI Paris 2025

Small but Mighty:

How MiniCPM Made Breakthroughs in the global Open-Source AI Landscape

OpenBMB

Tsinghua University NLP Natural Language Processing

ModelBest

Founding Partner at OpenBMB & ModelBest

Chao JIA

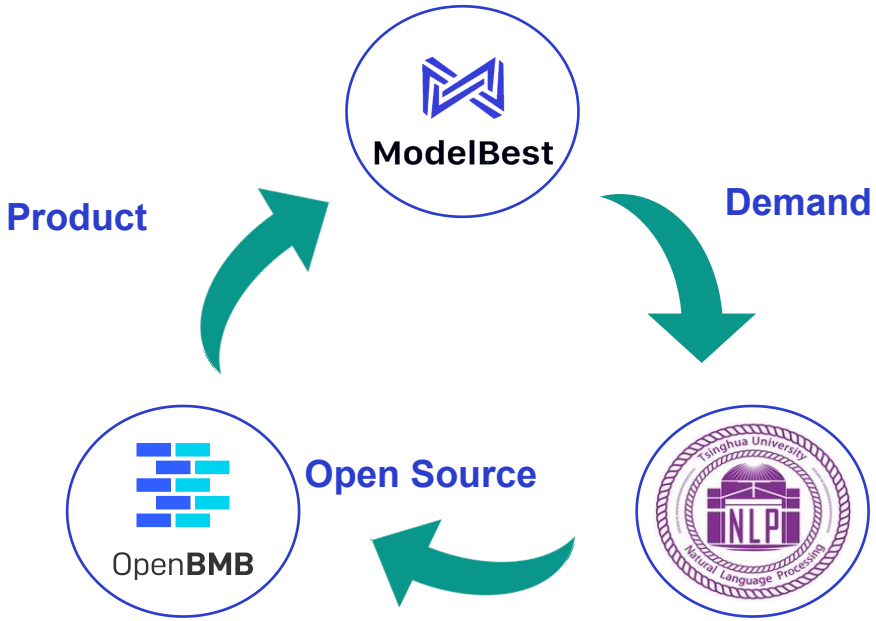
① Who are we?

② What we are doing?

③ What we have achieved?

④ What is the future of AGI?

Innovation Framework



Industry-Academia Partnership Drives Innovation

Over the past two years, ModelBest has grown from a Tsinghua Lab into a Fortune 50 AI innovator.

ModelBest is dedicated to developing models with high-efficiency. MiniCPM on-device model family has repeatedly topped GitHub and Hugging Face trending charts, becoming one of the most popular LLMs on Hugging Face in 2024.

ModelBest aims to bring efficiency models to everyone, every day, and everywhere in the future.

Modelbest: a Team with Insightful Vision and Strong R&D Capability

面壁小钢炮 MiniCPM | GOSIM

First-Generation Large Model

Chinese Pioneer in LLM

2019, The first Knowledge-Augmented Large Models in China, ACL 2019.

2020, Chinese pre-training Model CPM-1

2021, Persistently Learning LLM CPM-2

2022, Controllable Generative LLM CPM-3

Second-Generation Large Model

Continuous Iteration

Launch the CPM-Bee Multilingual Billion-Scale Model, reaching the top of ZeroCLUE

Agent

Global Leader in Agent R&D

Launch the universal agent platform AgentVerse

Release the multi-agent collaborative development framework ChatDev

Introduce the ultra-powerful AI agent application framework XAgent

Third-Generation Large Model

Surpassing GPT-3.5

Introduce the trillion-parameter multimodal large model CPM-Cricket, with scores surpassing GPT-3.5 and logical capabilities on par with GPT-4

Forth-Generation Large Model

Forerunner in On-Device Intelligence

Launch the flagship edge-side model MiniCPM-2B, maintaining a leading position globally

2019 Research Team

2022 Modelbest Company

2023Q1

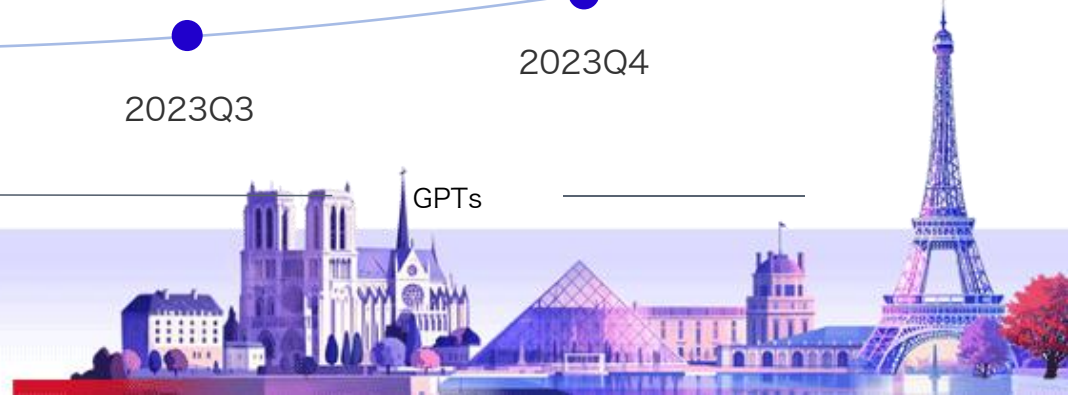
2023Q2

2023Q3

2023Q4

2024H1

GPT 3.5 — GPT 4 — GPTs



GOSIM AI Paris 2025

MiniCPM Family

面壁小钢炮 MiniCPM | GOSIM

Foundation Model

World's first on-device foundation model

2024.2

Knowledge density is more than twice that of Gemma, other **outperforming models such as Mistral-7B, Llama2-13B, and GPT3-175B with only 2.4B parameters**;
CPU inference, lossless quantization and compression to 25% of the original running memory

GitHub Trending 1st
Hugging Face Trending 1st

World's first on-device highly sparse reasoning architecture

2024.7

Dynamic sparse activation, **more suitable for on-device LLM**;
only need 12% neurons,
280% faster decoding

GitHub Trending Top
Hugging Face Trending Top

ChatGPT moment on devices

2024.9

Performance exceeds GPT-3.5, higher knowledge density than Gemma-3; requires only 2GB of RAM after quantization, the **strongest** on-device **Function Calling**

GitHub Trending 1st
Hugging Face Trending 3rd

Multimodal Model

MiniCPM-V2.0

A landmark work in multimodal understanding

2024.4

1.8 megapixel high-definition image decoding technology;
accurately recognizing difficult scenes such as street scenes and long images

GitHub Trending Top
Hugging Face Trending Top

MiniCPM-V2.5

U.S. Shell Copying Out of the Loop

2024.5

Single-image performance **surpasses Gemini Pro, GPT-4V**

GitHub Trending 1st
Hugging Face Trending 1st

MiniCPM-V2.6

Full end-side benchmarking GPT-4V

2024.8

Single Image, Multi-image, Video Understanding **SOTA (up to 20B);**
Device friendly, quantized to only 6G of memory

GitHub Trending 1st
Hugging Face Trending 3rd

MiniCPM-o2.6

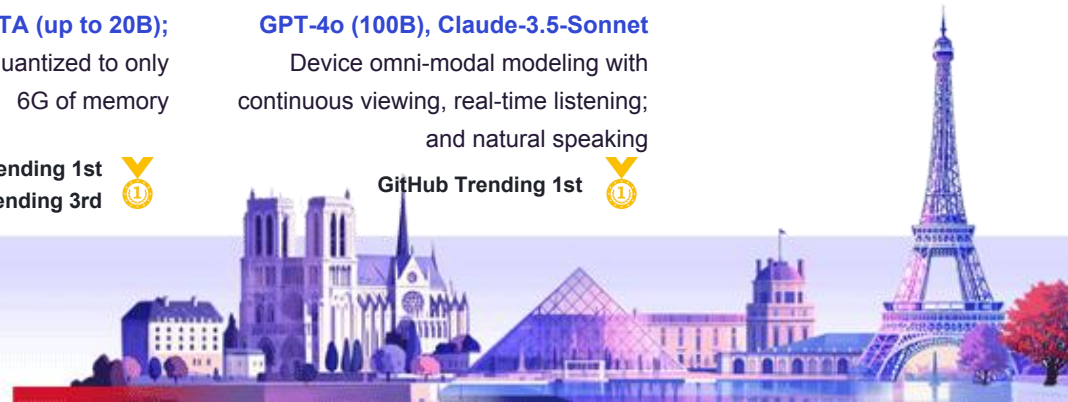
World's first end-side full modal model

2025.1

Multimodal capability **comparable to GPT-4o (100B), Claude-3.5-Sonnet**
Device omni-modal modeling with continuous viewing, real-time listening; and natural speaking

GitHub Trending 1st

2025.3



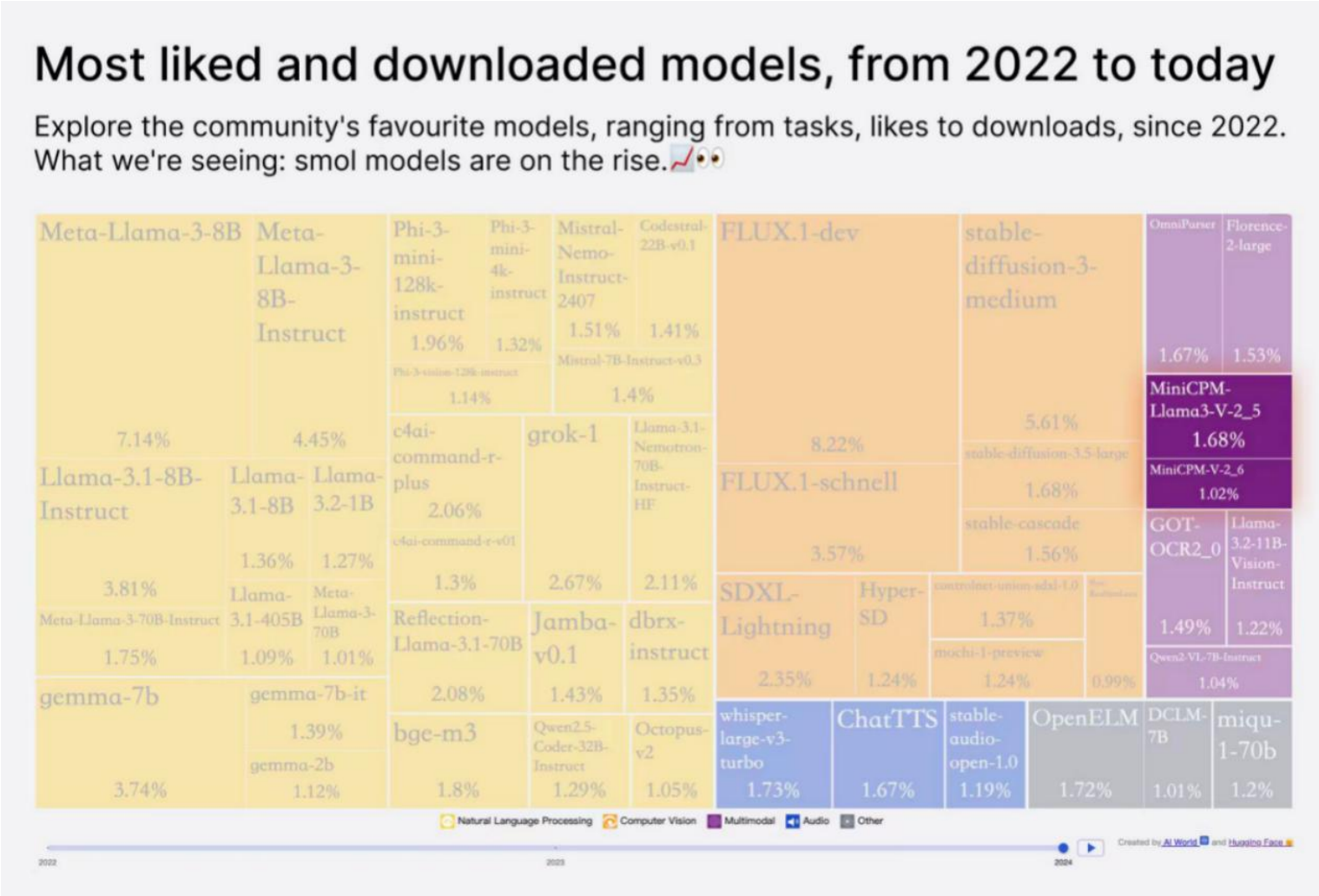
GOSIM AI Paris 2025

Most like and download top 1

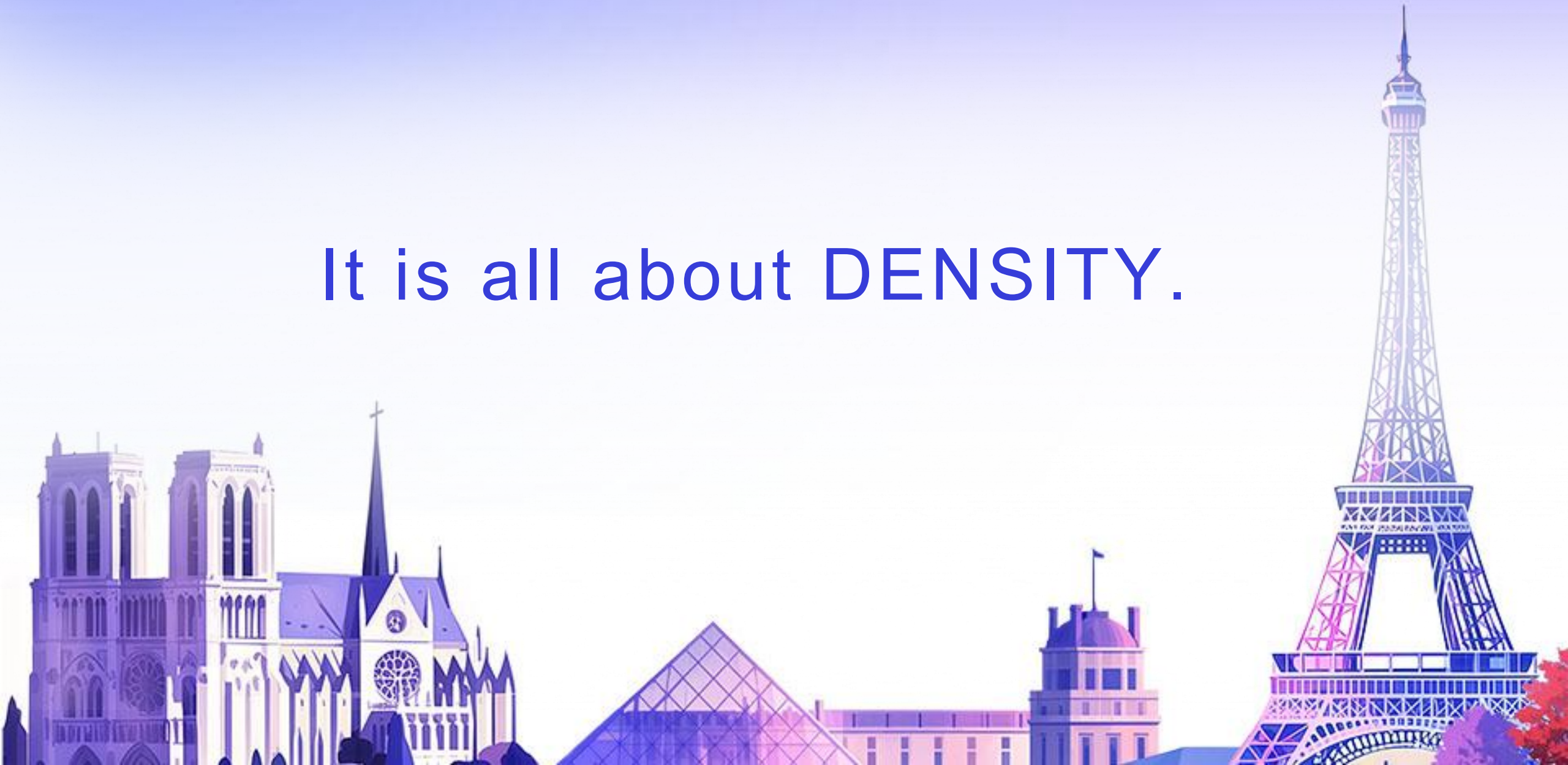
面壁小钢炮 MiniCPM | GOSIM

800 million+
Total downloads

26,700
GitHub Stars



It is all about DENSITY.

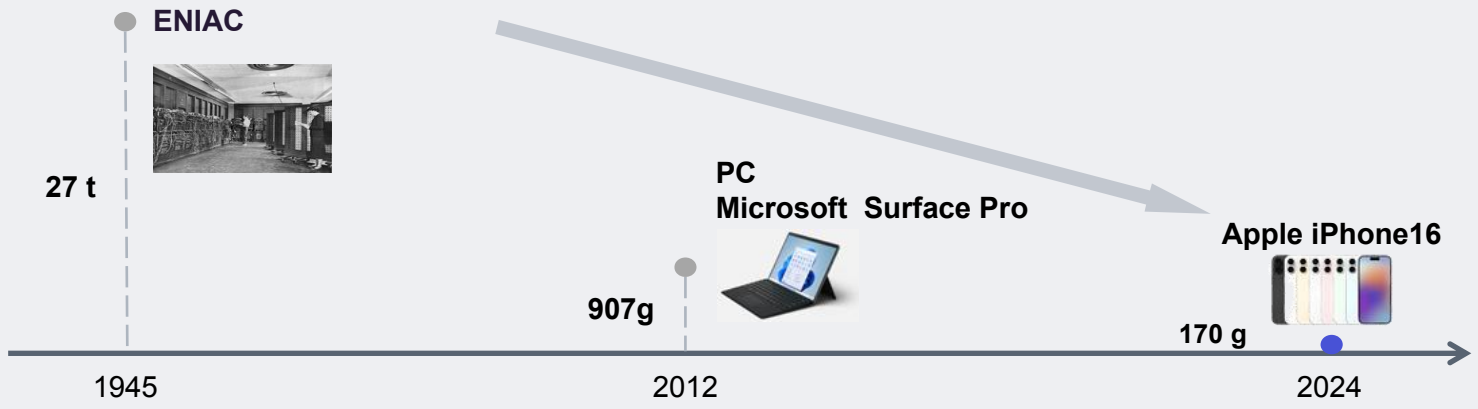


Smaller but More Powerful

面壁小钢炮 MiniCPM | GOSIM

Transistor Density

Computation Efficiency

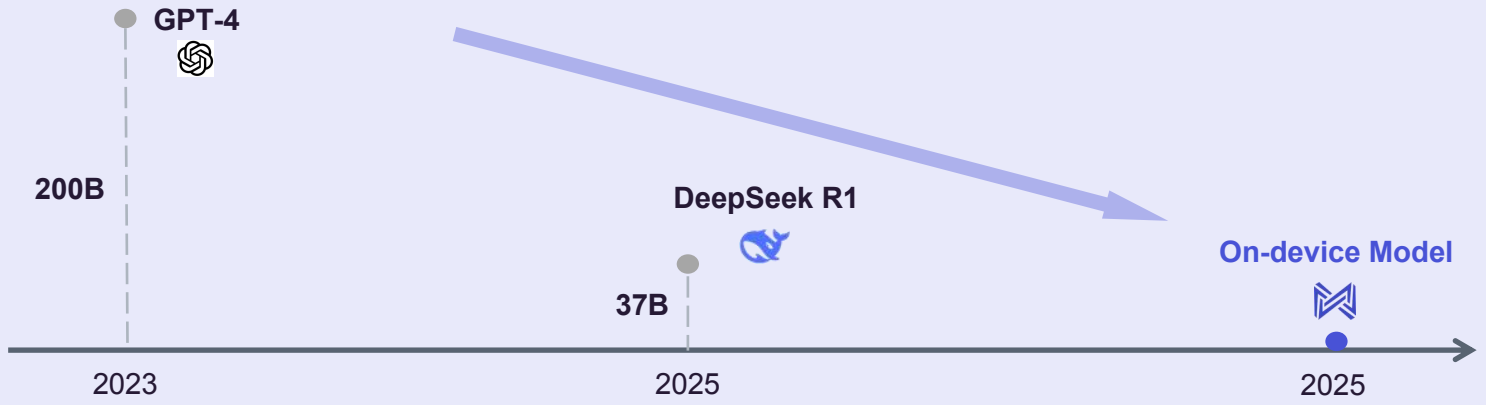


Moore's Law

The number of transistors on chips doubles every 18 months

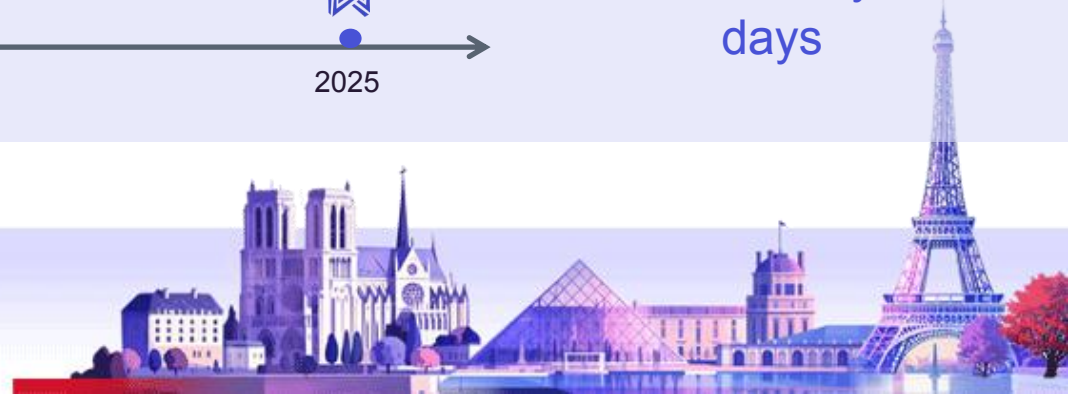
Knowledge Density

Intelligence Efficiency



Densing Law

Knowledge density doubles every 100 days

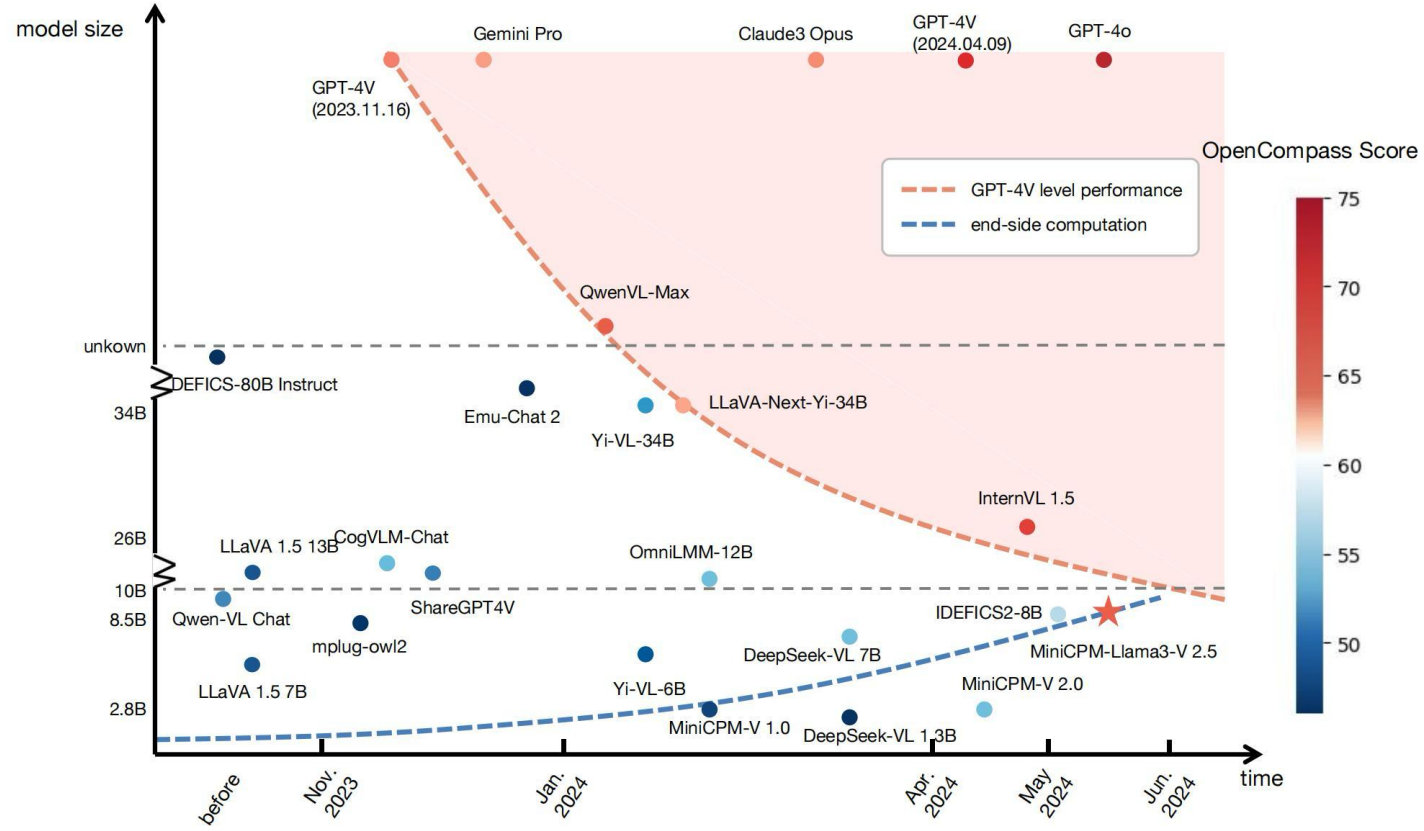


GOSIM AI Paris 2025

Moore's Law for MLLMs

The size of GPT-4V level MLLMs are rapidly decreasing

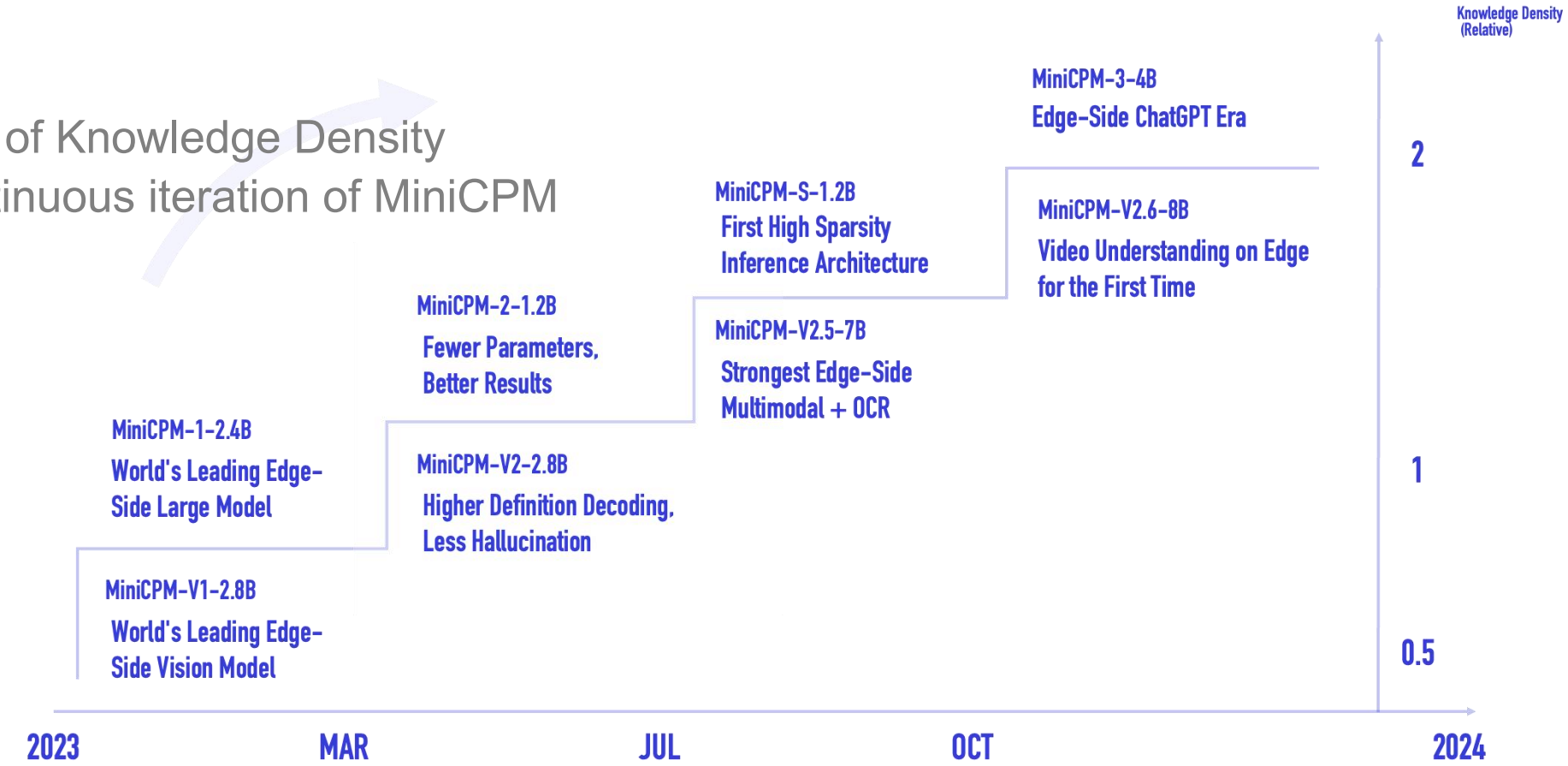
+ Fast growth of end-side device capacity → End-side Intelligence



Yao et al. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. 2024.

The Law of Knowledge Density

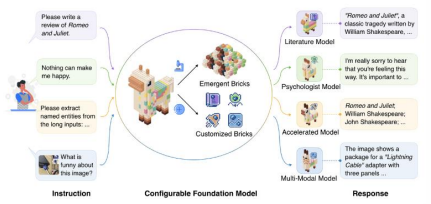
The Law of Knowledge Density
drives the continuous iteration of MiniCPM



Industry-leading LLM R&D System

面壁小钢炮 MiniCPM | GOSIM

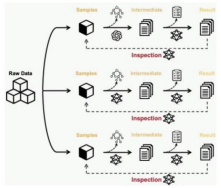
Architecture Efficiency: Modeling Functional Partitions



Training and reasoning computations are dramatically reduced by more than 90%. Model parameters are expected to be reusable

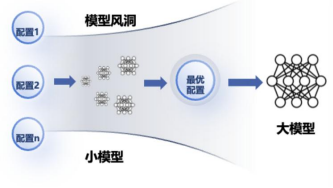
- Model Internal Optimization
  - Module Functional Partitioning
  - ProSparse spontaneous sparse activation enhancement
- Externally Enhanced Breakthroughs
  - Delta Tuning Unified Framework
  - Parameter Incremental Trimming Technology S^3 Delta

Data Efficiency: High Quality of Data Governance



- **UltraChat**
  - **800+** models in use, 5th in the world
- **UltraFeedback**
  - **1,200+** models in use, ranked 3rd in the world
- **UltraInteract**
  - **Reduce** errors formatting **by 90%**
- **RLAIF-V**
  - **Top** Hugging Face Trending

Learning Efficiency: Model Wind Tunnel Technology

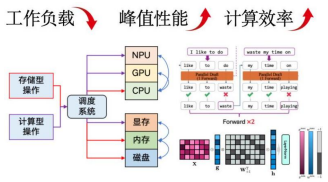


Self-developed "Model Wind Tunnel" platform, quickly formulate the optimal training program for different parameter scales and data conditions.

- Scaling Hyper-parameters Invariant LM
- Optimal Batch Size
- Optimal Learning Rate
- Optimal Learning Rate Scheduler

Multi-stage learning rate scheduler WSD: proposed based on model wind tunnels, became a built-in algorithm for Hugging Face's Transformers, and was widely adopted by LlaMA, DeepSeek open-source models

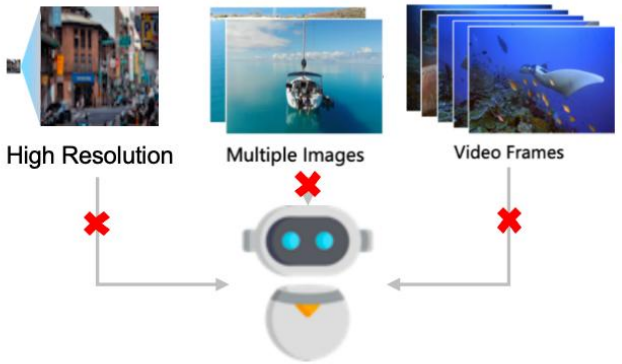
Computation Efficiency: Algorithmic Innovation

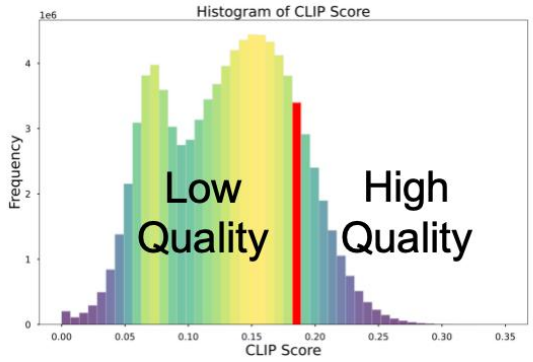


- **Training**
  - Fine-grained pipeline parallelism, fine-grained sequence parallelism
- **Reasoning**
  - Speculative sampling, **1.6-2x** acceleration
  - Articulated Snake Decoding, compared to autoregressive decoding **3.9x**
- **Quantization**
  - Ultra Low Bit-width Parameter Representation
  - Int-1 quantization through QAT, **90%** less storage, deployable on smartphones

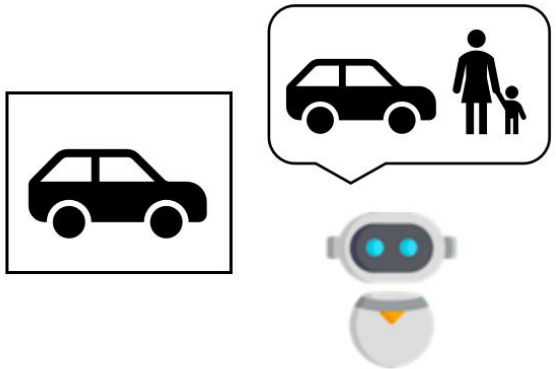
Key Research for MiniCPM-V

Model Architecture	Training	Data
Limited visual resolution and duration	**Weak in non-English languages**	**Severe hallucination, hard to trust**
• Fixed low resolution • High token usage • No multi-image and video encoding capability	• Most open-source MLLMs are English-only • High-quality Chinese multimodal data is limited	• Image-output inconsistency • Over 60% hallucination rate in open-source MLLMs • Unreliable for high-stakes tasks





Chinese multimodal data distribution



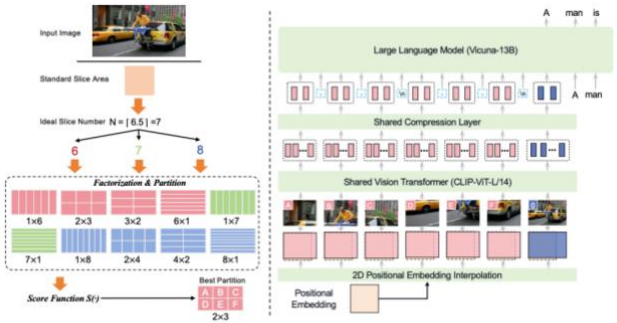
Key Research for MiniCPM-V

Efficient Architecture

Limited visual resolution and duration

Unified high-resolution visual encoding framework

- Support native aspect ratio
- Efficient token compression
- Unified single-image, multi-image and video modeling



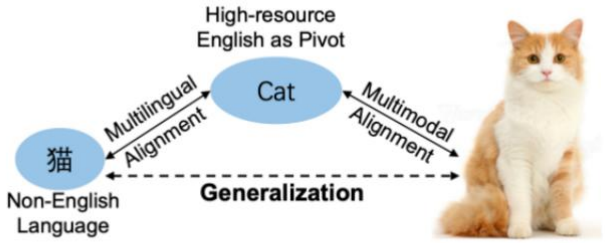
LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. ECCV 2024.

Efficient Training

Weak in non-English languages

Multilingual multimodal generalization

- Pre-training with only English text-image data
- Cross-lingual multimodal generalization in Chinese



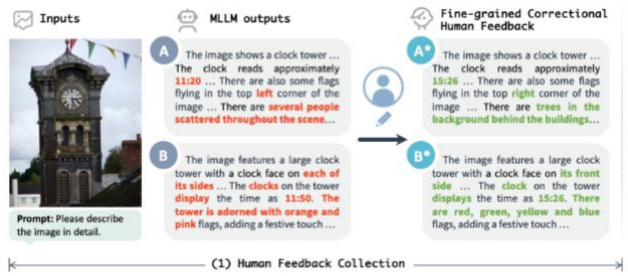
Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. ICLR 2024. Spotlight.

High-quality Data

Severe hallucination, hard to trust

Multimodal feedback data construction

- Aligning model behavior through fine-grained human feedback data
- Significantly reduced multimodal hallucinations

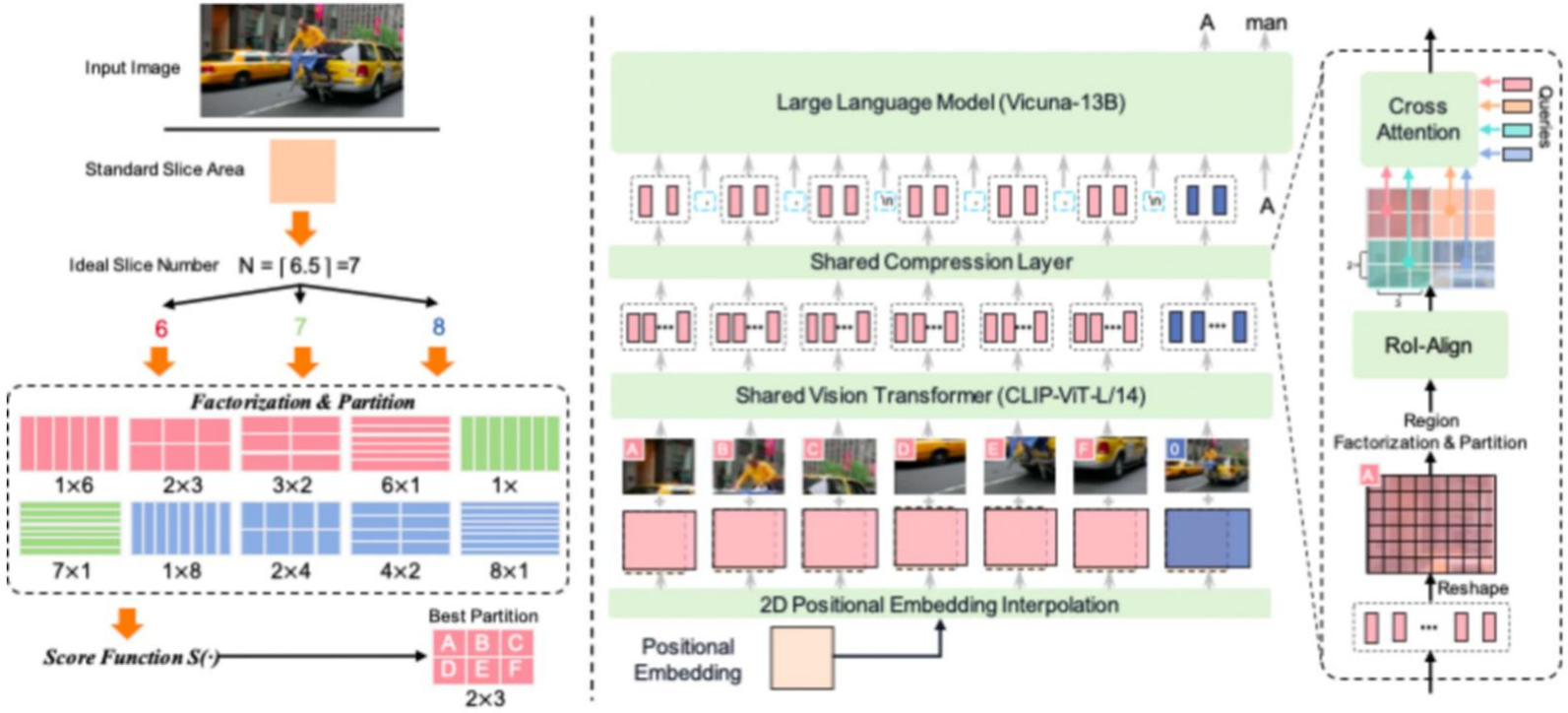


RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. CVPR 2024.

Architecture: High-res Visual Encoding Framework

Adaptivity

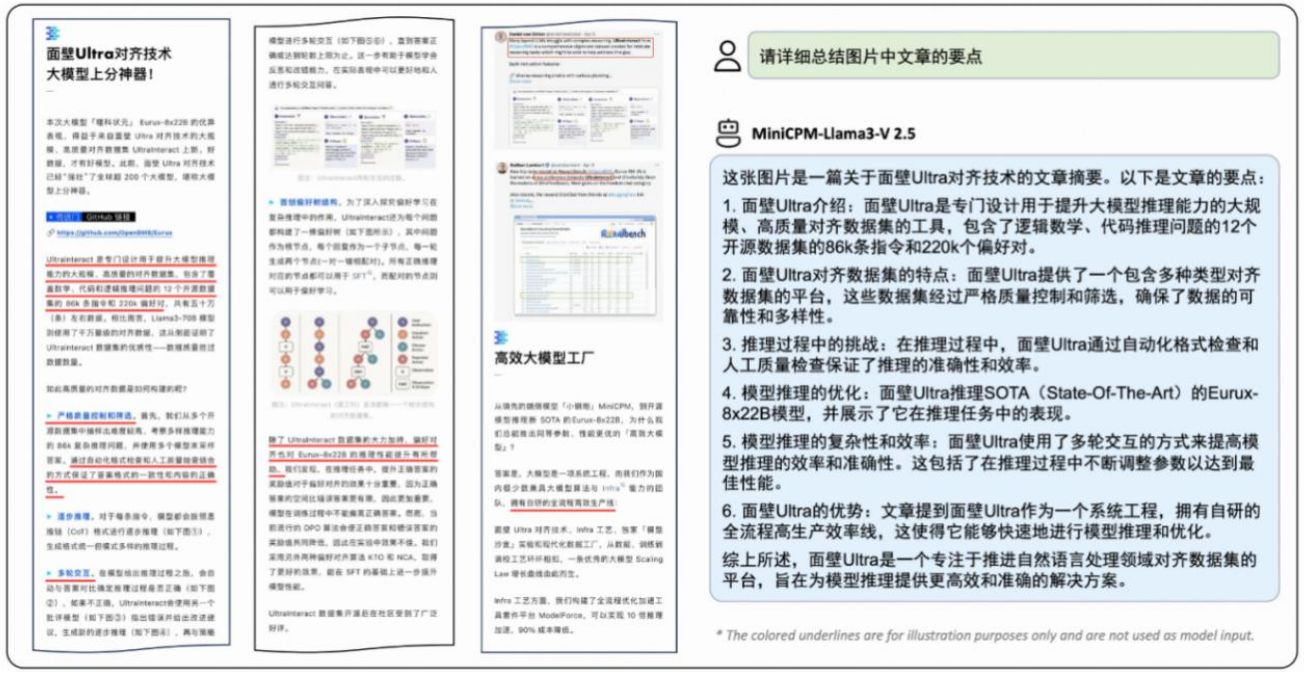
- **Image Modularization:** Divide native-resolution images into smaller **variable-sized slices**
- **Principles:** Slices should be similar to the pretraining setting of ViT



LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. ECCV 2024.

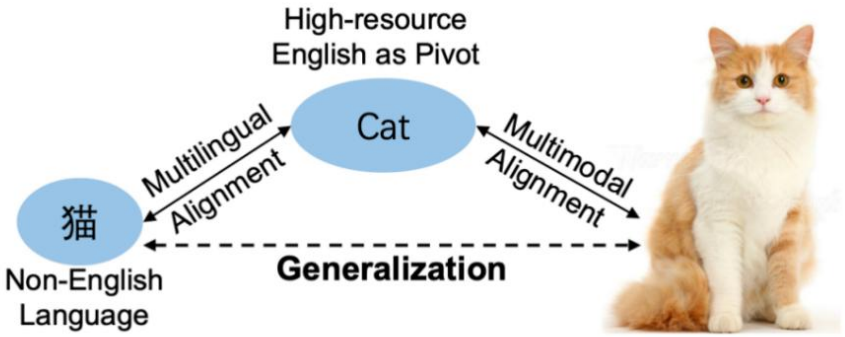
MiniCPM-V: Efficient End-side MLLM

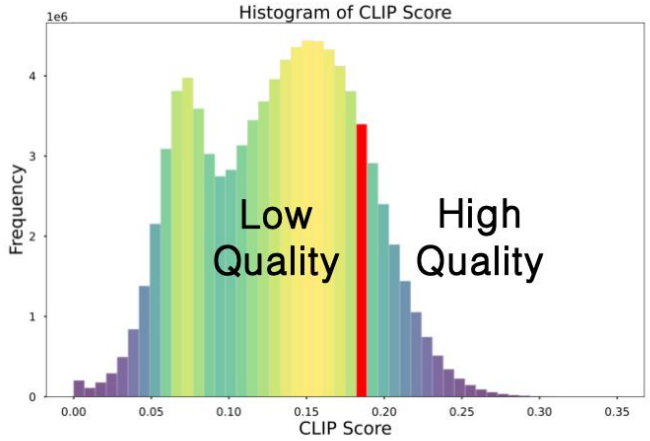
- **High-Resolution Image Encoding**
  - Support images of up to 1.8 million pixels (e.g., 1344x1344) in any aspect ratio
  - SOTA performance on OCRBench
  - Token density = # pixels at maximum resolution / # visual tokens



Training: Multimodal Multilingual Generalization

- **Problem:** MLLMs are typically strong in English, but weak in other languages
- **Challenge:** Limited data scale and quality for non-English languages
- **Common Solutions:** Intensive data collection and cleaning for each language
- **Our Solution:**
  - **Multimodal Generalization:** Transfer multilingual capability from multilingual LLM
  - **High Efficiency:** Lightweight instruction tuning on Chinese data as activation



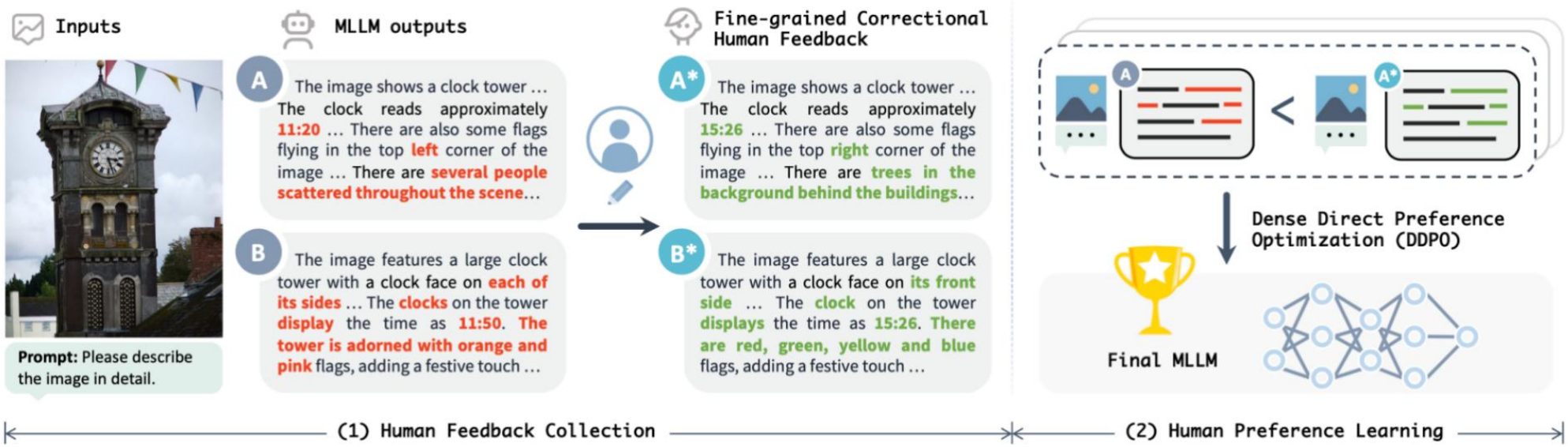


Chinese multimodal data distribution

Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. ICLR 2024. Spotlight

Data: Feedback Data Construction

- **Preference Formulation:** Collect fine-grained human feedback in the form of segment-level corrections on hallucinations
- **Dense-DPO:** Emphasize more on segments of hallucinations vs. corrections



$$Y = \boxed{Y_p} + Y_s + Y_n$$ (Truly preferred behavior; Non-robust bias; Language variance)

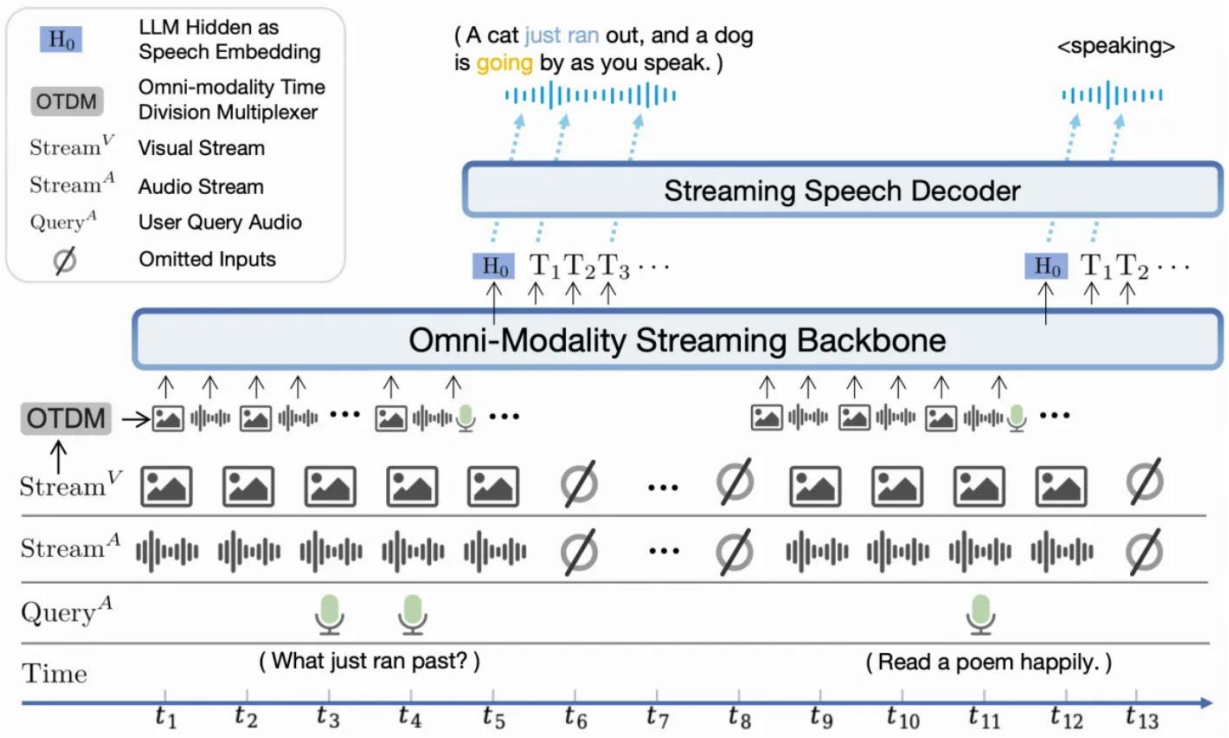
RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. CVPR 2024.

MiniCPM-V: Efficient End-side MLLM

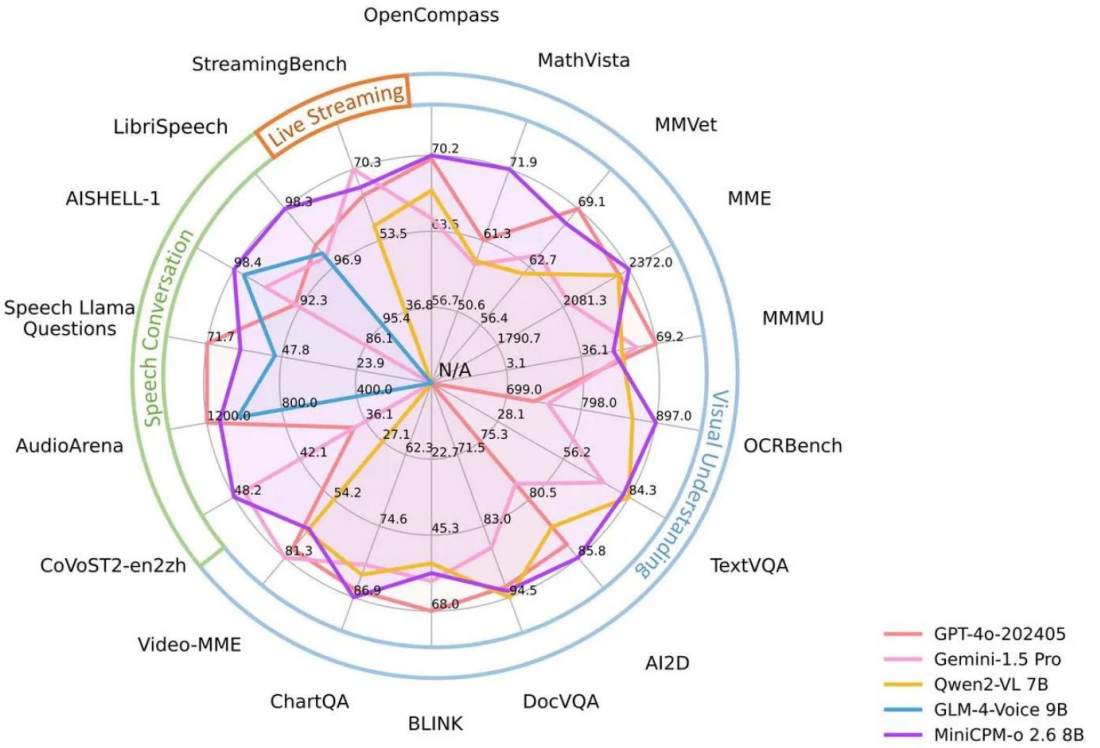
- SOTA performance for video understanding among models under 30B
- Support real-time video analysis on iPad Pro for the first time

MiniCPM-o 2.6 （8B）

- A GPT-4o Level MLLM for Vision, Speech, and Multimodal Live Streaming on Your Phone



Architecture



Evaluation

GOSIM AI Paris 2025

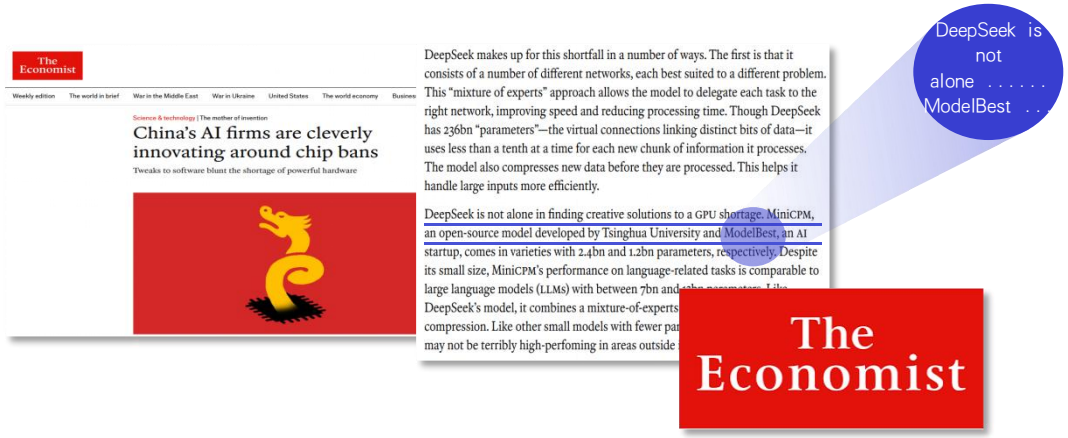
MiniCPM-o 2.6:Efficient End-side MLLM

World-renowned Media Coverage

The Economist: China's AI firms are cleverly innovating around chip bans

DeepSeek is not alone in finding creative solutions to a GPU Shortage. MiniCPM, an open-source model developed by Tsinghua University and ModelBest ..... MiniCPM's tiny size makes it well-suited for personal devices.

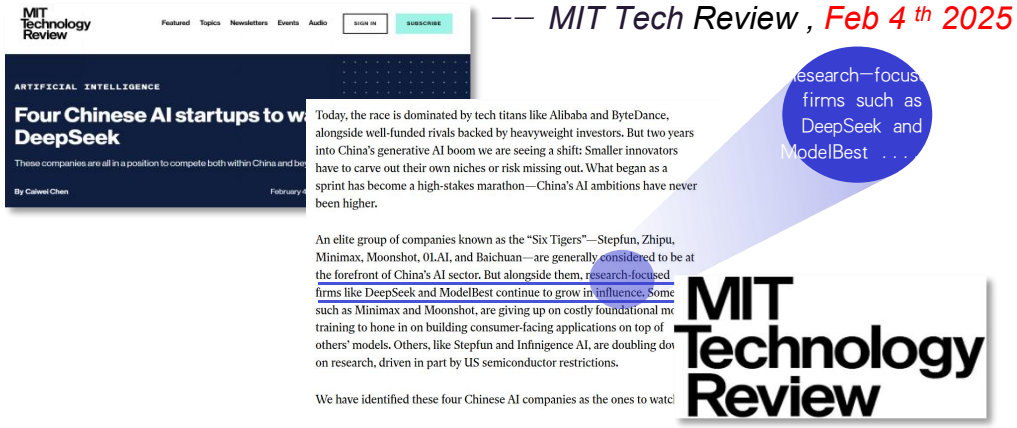
—— *The Economist, Sep 19 th 2024*



MIT Technology Review: Four Chinese AI startups to watch beyond DeepSeek

Research-focused firms like DeepSeek and ModelBest continue to grow in influence ..... the company has distinguished itself by leaning into efficiency and embracing the trend of small language models.

—— *MIT Tech Review , Feb 4 th 2025*



"There's a series of mind-blowing tech reports and open-source models coming from China (DeepSeek, MiniCPM, UltraFeedback...) ..... MiniCPM (amazing super small model – deep dive in the experiments)."

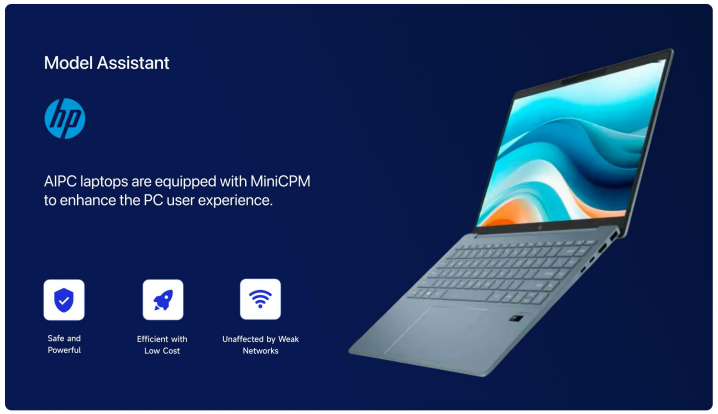
-- Hugging Face Co-Founder | Thomas Wolf

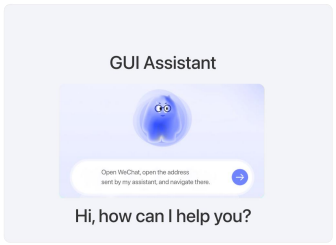
Empowering Each Device for Everyone

面壁小钢炮 MiniCPM | GOSIM

AIPC、AI Phone — Your Smarter Personal AI Assistant

intel. MEDIATEK





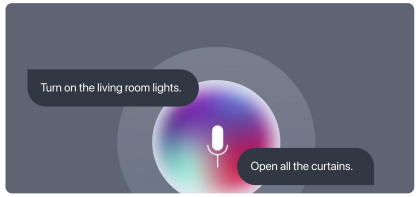
Unaffected Weak Networks

Faster Local Response Speed and Stronger Privacy Protection in a Weak or Disconnected Network Environment.

On-device Efficiency

On-device intelligent agent, giving simple oral instructions to complete complex operations.

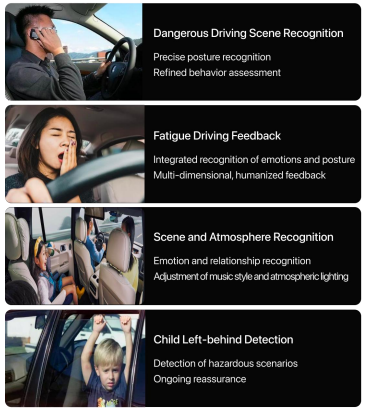
AIoT — Your Smarter Personal AI Household Manager and AI Translator



Becoming the Smart Home Household Manager, a Caring and Thoughtful Home Smart Center.

Intelligent Cabin — Smarter vLLMs in the Car

ThunderSoft 长城汽车



Cabin Brain

Understands the Intentions of the Driver and Passengers Automatically Executes Tasks Based on Vehicle Conditions

In-Cabin Perception

Real-time Detection of the Environment and Behavior Adjusts Vehicle Status or Triggers Interaction Accordingly

External Perception

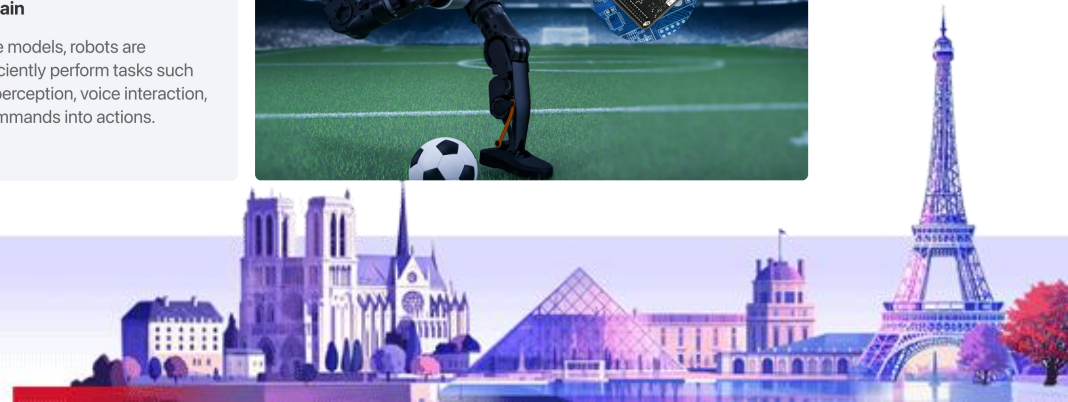
Precisely Identifies and Analyzes Exterior People Instantly Triggers Vehicle Control and Interaction

Embodied Intelligence — Robots with an On-device Brain

BOOSTER ROBOTICS

On-device Brain

By leveraging edge models, robots are empowered to efficiently perform tasks such as environmental perception, voice interaction, and translating commands into actions.



GOSIM AI Paris 2025

Global Partners

面壁小钢炮 MiniCPM | GOSIM

intel

Qualcomm

ARM

MEDIATEK

hp

HUAWEI

Lenovo

NVIDIA

长城汽车

德赛西威

BOOSTER ROBOTICS

SIEMENS

SAIC 上汽集团 SAIC MOTOR

ThunderSoft

Timekettle® 时空壶

百度智能云 cloud.baidu.com

BAAI 智源研究院

义乌小商品城 chinagoods.com

知乎

招商银行 CHINA MERCHANTS BANK

The Next Big Thing

面壁小钢炮 MiniCPM | GOSIM

Capability Leap

Understanding Knowledge → Understanding Intent → ① Deep Thinking → ③ Autonomous Learning → Autonomous Collaboration

Innovation

OpenAI GPT

OpenAI ChatGPT

OpenAI o1

2018 — 2022 — 2025 — 2026 — 2027

Efficiency Enhancement

High knowledge-density On-device Model

MiniCPM-1、2、3

High-density model training method

② Deep-thinking On-device Model

MiniCPM-4、5

Reasoning capability surpassing Deepseek-R1

④ Personalized On-device Model

MiniCPM-X

Omni On-device Large Model

MiniCPM-4o、5o

Omni-modal capability comparable to GPT-4o/5o

GOSIM | 面壁小钢炮 MiniCPM

GOSIM AI Paris 2025

Welcome to follow us

X/Twitter

GitHub

HuggingFace

GOSIM | 面壁小钢炮 MiniCPM

GOSIM AI Paris 2025

AGI FOR LIVES

GOSIM | 面壁小钢炮 MiniCPM

GOSIM AI Paris 2025

THANK YOU