AI large model principle foundation

This section only describes the knowledge theory related to the multi-mode AI large model. Those who are not interested can ignore this section.

This section does not involve the operation and use of the robot dog.

The Generation of AI Big Models

1. Technical architecture evolution

1. The core of the multimodal large model is to integrate multi-source data such as text, images, audio, and video. Its architecture has undergone a transformation from a single modality to cross-modal fusion:

- Early unimodal models: such as AlexNet (image classification) and BERT (text processing), are designed only for a single task and require independent training of different models.
- Breakthroughs in Transformer and Large Language Model (LLM): Cross-modal semantic alignment is achieved through a unified framework (such as the GPT series, CLIP), mapping different data into the same semantic space and reducing information loss.
- End-to-end multimodal modeling: Such as GPT-4o and Google Gemini, which directly process multimodal input and output through a single model, eliminating intermediate conversion steps and improving efficiency.

2. Key Components and Training Methods

- Encoder: Converts data of different modalities (such as image pixels and audio waveforms) into a unified high-dimensional feature vector. For example, a visual encoder extracts image semantics and a text encoder generates word embeddings.
- Cross-modal attention mechanism: dynamically adjust the weights of each modality. For example, Microsoft BEiT-3 uses cross-modal attention to achieve deep association between text and images5.
- Pre-training and fine-tuning: Pre-training on large-scale multimodal data (such as LAION-5B), and then fine-tuning for downstream tasks (such as robot control and medical diagnosis) to improve generalization capabilities.

3. Cross-modal alignment and knowledge fusion

- Alignment techniques: For example, CLIP aligns image and text features through contrastive learning to achieve zero-shot classification of open vocabulary.
- External knowledge enhancement: Models such as KOSMOS-1 introduce medical knowledge bases to improve the accuracy of complex question answering.

2. AI big model application level

1. Robotics and Embodied Intelligence

- General-purpose robots: Multimodal LLM gives robots the ability to reason and learn autonomously. For example, Tesla Optimus can adapt to unstructured environments by integrating multiple sensors such as vision and touch.
- Real-time interaction and control: Google's RT-2 model directly converts multimodal input into action encoding, significantly improving the success rate in unknown tasks.

- Industry case: Boston Dynamics Spot serves as a tour guide in museums, emphasizing interactive entertainment rather than pure functionality.

2. Generative content creation
   Vinyl video and 3D modeling: OpenAI Sora can generate high-fidelity videos, and Stable Diffusion 3 supports 3D content generation, promoting innovation in the film, television, and gaming industries.
   Digital humans and virtual assistants: such as Google Project Astra and Tencent MM-LLMs, which enable natural conversations and real-time video editing.
3. Deep penetration of vertical industries

   Medical diagnosis: Shukun Technology's "Digital Human Body" platform integrates medical images and medical records to improve diagnostic efficiency5.
   Industrial quality inspection: Multimodal models combined with synthetic data can detect complex defects and reduce the error rate by 90%.
   Financial anti-fraud: Cross-modal correlation analysis (such as voice + transaction records) has an accuracy rate of 98%.

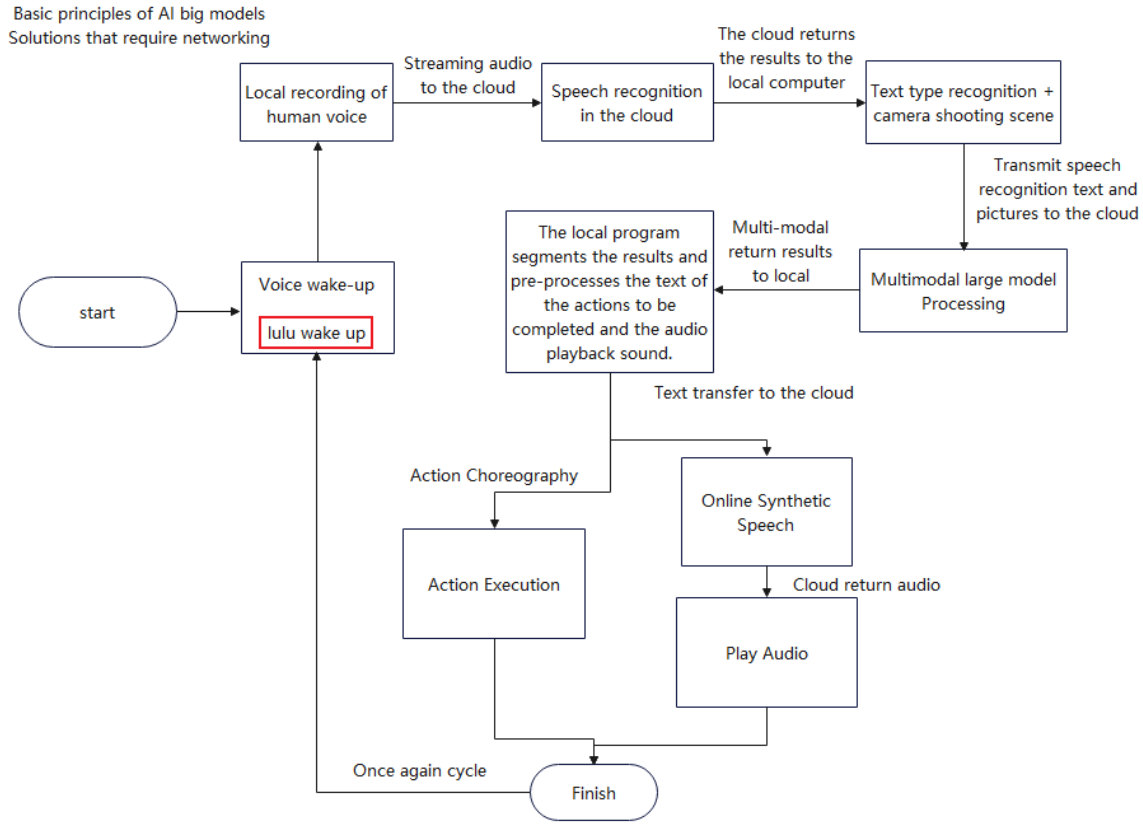
3. Summary

Multimodal large models are reshaping the capabilities of AI through unified architecture and cross-modal fusion, and their applications have shown great potential in fields ranging from robotics to medical care and finance.

In the future, technology needs to continue to make breakthroughs in computing power optimization, ethical governance, and modality expansion to realize the vision of "human-machine symbiosis".

4. Application examples of multimodal robot dogs

The solution of the embodied intelligence multimodal combined online platform of the robot dog is as follows:


Basic principles of AI big models
Solutions that require networking
Local recording of human voice
Streaming audio to the cloud
Speech recognition in the cloud
The cloud returns the results to the local computer
Text type recognition + camera shooting scene
Transmit speech recognition text and pictures to the cloud
Multimodal large model Processing
Multi-modal return results to local
The local program segments the results and pre-processes the text of the actions to be completed and the audio playback sound.
start
Voice wake-up
lulu wake up
Text transfer to the cloud
Action Choreography
Online Synthetic Speech
Action Execution
Cloud return audio
Play Audio
Once again cycle
Finish