ITUPublications

International Telecommunication Union
Standardization Sector

ITU-T Technical Report

(10/2025)

FSTR.HAI

Holistic AI: A survey about a technical framework for artificial intelligence (AI) services and AI capabilities

Technical Report ITU-T FSTR.HAI
Holistic AI: A survey about a technical framework for artificial intelligence (AI) services and AI capabilities

Summary

Technical Report ITU-T FSTR.HAI establishes a holistic framework for Al foundation models and provides guidance and reference for the usage or application of Al foundation models, in conditions including computing resource sharing, data sharing and infrastructures sharing. It mainly focuses on holistic AI framework for multiple AI applications and models. This Technical Report includes the introduction of holistic AI and ideas of a framework for AI service and AI capabilities, a holistic framework for multiple AI applications and models, and mechanisms of holistic AI, enabling technologies of holistic AI.

Keywords

AI, AI capabilities, AI service, foundation model, framework, holistic AI, large language model, multimodal.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

© ITU 2026

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

Page

1 Scope ... 1

2 References ... 1

3 Definitions ... 1

3.1 Terms defined elsewhere ... 1

3.2 Terms defined in this Technical Report ... 1

4 Abbreviations and acronyms ... 1

5 Conventions ... 2

6 Introduction ... 2

6.1 Background ... 2

6.2 Necessity of the survey of holistic AI and related technologies ... 3

6.3 Concepts for holistic AI ... 3

7 Holistic framework for multiple AI applications and models ... 3

8 Mechanisms of "Holistic AI" ... 6

9 Enabling technologies of "Holistic AI" ... 7

9.1 AI-exchange system ... 7

9.2 AI capabilities atomization ... 8

9.3 End-to-end close loop optimization ... 8

9.4 Network native AI ... 9

9.5 Trustworthiness ... 10

Bibliography ... 11

Technical Report ITU-T FSTR.HAI

Holistic AI: A survey about a technical framework for artificial intelligence (AI) services and AI capabilities

1 Scope

This Technical Report mainly focuses on the holistic AI framework for multiple AI applications and models.

The scope of this Technical Report includes:

- Introduction of holistic AI and ideas of a framework for AI services and AI capabilities;
- A holistic framework for multiple AI applications and models;
- Mechanisms of holistic AI and related progress;
- Enabling technologies of holistic AI.

2 References

None.

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Technical Report

None.

4 Abbreviations and acronyms

This Technical Report uses the following abbreviations and acronyms:

AI	Artificial Intelligence
DeRy	Deep Model Reassembly
E2E	End-to-End
5G	Fifth Generation
GPT	Generative Pre-Trained
GPU	Graphics Processing Unit
HAI	Holistic Artificial Intelligence
HNN	Holistic Neural Network
LLM	Large Language Model
MAC	Memory Attention and Composition
OS	Operating System

5 Conventions

None.

6 Introduction

6.1 Background

The generative pre-trained (GPT) transformer language model has elevated the overall level of machine intelligence to an unprecedented level. Compared with small models for specific tasks, foundation models have several significant advantages, including but not limited to the following:

- **Generic understanding capability**: Large language models (LLMs) represented by GPT can be used for the problems, instructions, dialogue context, long texts from users and even various logical reasoning questions. The universality of LLMs in various types of natural language tasks and the improvement effect of LLMs on various downstream tasks are significant and indisputable natural language understanding, which is not only crucial but also difficult.
- **Information integration and compression capability**: LLM takes most of the accumulation of human civilization, including but not limited to books, papers, codes and the Internet as input contents, and carries out unsupervised learning with the goal of prediction or filling, forming a large model with a parameter scale of billions or even hundreds of billions. The model has acquired extensive knowledge and achieved extremely strong reasoning ability.
- **Diverse generative capability**: Current large models, including language models, visual models, speech models and multimodal models, have significantly improved quality and diversity compared with previous technologies, changing the previous monotonous and tedious state of machine generated content.
- **Emergence capability**: Emergence is defined as an ability that cannot be observed in small models but that suddenly emerges only in large models, e.g., the ability to suddenly see when model parameters reach a certain scale. Emergence ability has also attracted attention to AI security, and as the model parameters continue to increase, large models may produce unpredictable and harmful abilities.

The success of large models has sparked a new technological trend and revolution; it is widely predicted that large models will accelerate the empowerment of various industries and disciplines. The potential changes it can bring are enormous and even revolutionary, which brings great hope and concerns. However, there are still many challenges when large models move towards large-scale applications:

- **Illusions**: These are also described as "large but unstable", as there is a certain probability that a model's understanding of a problem is not robust enough, the generated contents cannot match the facts and the reasoning is not rigorous enough.
- **High cost of training and deploying**: The model parameters of a typical LLM are generally in the range of billions or even trillions. Every time such a large model is trained, it requires the investment of trillions of tokens of data and thousands of high-performance graphics processing units (GPUs) and costs several months along with millions of dollars. Only a few enterprises and research institutions can invest continuously for such large model training. Most large models are currently deployed in the cloud because only cloud-based machines can run them efficiently.
- **Supply and demand gap**: There is a huge demand from both industry users and individual users of intelligent services, but this is restricted to the limited types of models.

6.2 Necessity of the survey of holistic AI and related technologies

An AI foundation model is any AI model that is trained on broad data that can be adapted to a wide range of downstream tasks. The sheer scale and scope of AI foundation models from the last few years have stretched our imagination of what is possible. At the same time, existing AI foundation models have the potential to accentuate harms, and their characteristics are in general poorly understood. Given their impending widespread deployment, they have become a topic of intense scrutiny. AI foundation models acquire various capabilities that can power applications. There are five main potential capabilities, and the potential limits will consider the philosophy of understanding of language, vision, robotics, reasoning and search, interaction. The capabilities of AI foundation models indicate that they have the potential to transform various sectors and industries, extending the roles AI plays in society. This Technical Report establishes a holistic framework for AI foundation models and provides guidance and reference for the usage or application of AI foundation models, in conditions including computing resource sharing, data sharing and infrastructures sharing.

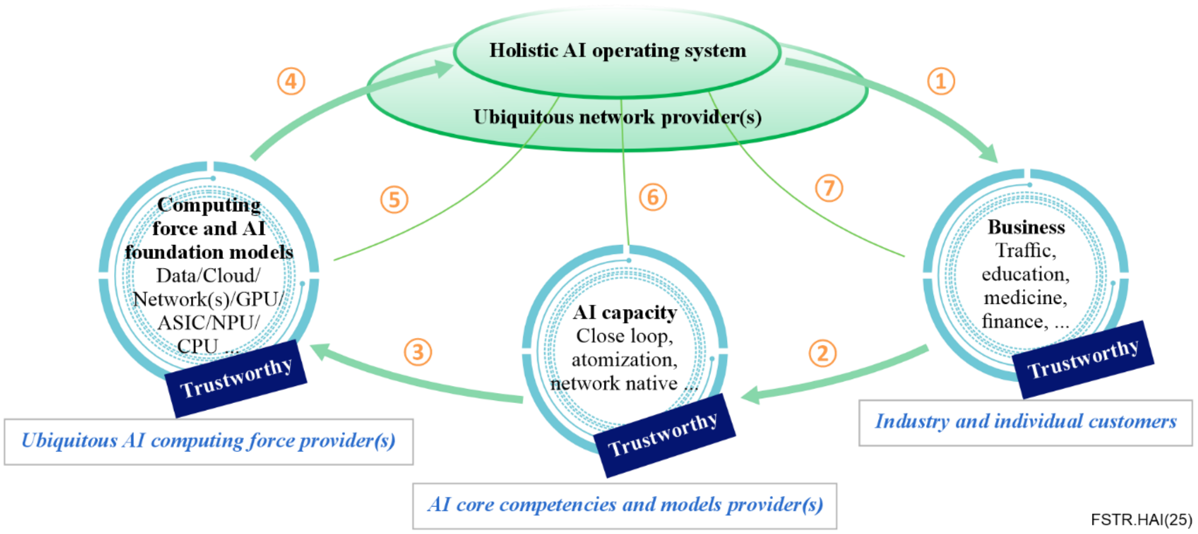
6.3 Concepts for holistic AI

The followings are new concepts based on holistic AI:

- **AI-exchange**: A scheduling strategy that can achieve AI capability matching and can independently allocate relevant capabilities and resources among intelligent business demanders, AI models and capability providers, ubiquitous AI computing resource providers and ubiquitous network resource providers.
- **Holistic AI**: A technical framework in which different AI models, computing resources and networks can be flexibly scheduled with the processing of both numerical and non-numerical data, so as to meet any requirements of AI stakeholders.
- **Holistic neural network (HNN)**: A neural network whose models and capabilities have been atomized.

7 Holistic framework for multiple AI applications and models

In order to make the holistic framework for AI foundation models intuitive and understandable in terms of means of standardization, descriptions and requirements, a holistic framework for AI foundation models is illustrated in Figure 1.


Holistic AI operating system
Ubiquitous network provider(s)
④
①
⑤
⑥
⑦
Computing force and AI foundation models
Data/Cloud/ Network(s)/GPU/ ASIC/NPU/ CPU ...
Trustworthy
AI capacity
Close loop, atomization, network native ...
Trustworthy
Business
Traffic, education, medicine, finance, ...
Trustworthy
③
②
Ubiquitous AI computing force provider(s)
AI core competencies and models provider(s)
Industry and individual customers
FSTR.HAI(25)

① The HAI OS provides relevant service(s) to specific business:

- **Services:** Intelligent customer service, recommender system, information searching, code generation, content generation, intelligent assistant, knowledge insight, data analysis, data processing, IT system operation and maintenance, software development, safety and security,
- **Business:** Traffic, education, medicine, finance, smart city, agriculture, industry, travel, multimedia, entertainment, biotechnology, robot, etc. The business can be for industry customers and it can be also for individual customers.

② The requirement(s) or demand(s) from/of business is suggested to customize the specific "AI capacity".

③ The AI computing force can be dispatched or managed depending on specific AI capacity.

④ The trained AI model(s) are provided to the HAI OS to do the relevant operation(s).

⑤ The HAI OS can operate the AI computing force directly, including but not limited to the hardwares, softwares, data centres, etc.

⑥ The HAI OS can operate the AI capacities too, including but not limited to the capacity atomization, network native, and other related AI capacities.

⑦ The business can feedback to HAI OS.

Figure 1 – Holistic AI framework for multiple applications and models

Holistic AI (HAI) relies on ubiquitous networks and AI computing resources to achieve flexible and efficient configuration, scheduling, training and deployment of AI models and capabilities in an open environment, to meet increasingly diverse digital business needs, while ensuring the trustworthiness and controllability of large-scale intelligent businesses within the framework. Based on HAI, user intelligence needs can be flexibly expressed in natural language, graphics, images, component orchestration and other ways. HAI uses large models as the main technical foundation to understand user needs and form execution solutions. The execution solution includes models, capabilities, data and calculations needed to match business needs. Network resources; HAI deploys models and capabilities to corresponding computing network resources, flexibly schedules and jointly optimizes to meet the requirements of business in production environments This technical framework involves multiple unique core technologies, including big loop AI for AI services, atomized AI capabilities, network native AI and secure and trusty AI services.

In Figure 1, the demand side of intelligent business, AI foundation models and capability providers, AI computing resource providers, and network service providers are independent of each other. The systematic intelligent core operation system (HAI-OS) analyses the rich intelligent business needs of various industries and individual customers, and schedules corresponding AI models, computing resources and network resources to jointly meet the intelligent computing needs of customers. HAI-OS can be centralized or decentralized. The following are the general descriptions of the main blocks in Figure 1:

– **Industry and individual customers (intelligent service users)**: Intelligent service users can use natural language, use cases, tools and models provided by HAI system to describe requirements. HAI-OS analyses the requirements and generates an execution plan. Any action nodes in the execution plan correspond to the AI model or tool registered in HAI-

OS. The intelligent service user can adjust and configure the attributes of the workflow and node, such as input and output methods, accuracy requirements, service range scope, speed requirements, security and other requirements. If the capability required by the user is beyond the scope of the current HAI-OS registration capability, the service party can release this task, which will be developed or registered by the AI capability party. The current large-scale model has a strong ability to integrate information, understanding and reasoning, which makes HAI-OS possible.

- **AI core competencies and models provider(s)**: AI models and capabilities can be provided by general-purpose LLMs or special capabilities, e.g., voice recognition engine for customer service, motion posture recognition capability. the providers of multiple technologies can be multiple or one. The generality, adaptation mode, compliance and deployment computing power requirements of these models and capabilities will be verified and tested at the time of registration and each subsequent update, and their accuracy and security can be verified on the agreed data set. When these capabilities are matched by business requirements, they will be further verified under specific business scenarios on the test data set provided by the business party. If the business requirements cannot be met, the business party can use the AI model and the tools provided by the capability provider for tuning, retraining or iterative training. Therefore, the AI model and capability providers provide not only a reasoning service but also self-test tools, tuning and even iteration tools.
- **Ubiquitous AI computing force providers**: Providers include cloud computing resources, various edge computing resources and end-to-end computing resources of the data centre. These computing resources can be assets of governments, companies or individuals. Two-way communication takes place between computing resources and HAI-OS. Computing resources report their information dynamically and the HAI-OS release task is completed by the response of computing resources, so that HAI-OS can more quickly deploy the AI capabilities required by the business on the most efficient computing nodes that can meet business requirements. Each node can deploy one or more AI capabilities or AI computing resources. The matching between businesses depends on both technical indicators and prices. A universal computing resource is not only a multiclass computing power at the cloud edge, but also a heterogeneous computing resource, including computer processors, GPU servers, embedded neural network processors and even personal computers with idle time. By integrating different computing resources and using them efficiently, the pan AI computing resource provider alleviates the high cost of large model training and deployment to a certain extent.
- **Ubiquitous network operators(s)**: When serving smart computing, the network needs to transfer the data sensed by various sensing devices and various size models to the universal computing resources including cloud edge computing resource for processing or computing. Therefore, in order to provide intelligent services for most of society, not only the deep integration of network and computing resource but also the deep integration of network and AI computing resources such as data and models is required. Data is transferred and calculated during network transmission.
- **Holistic intelligent system operator(s)**: HAI-OS is the core system of HAI. It needs to accurately understand the user's intelligent needs and output the implementation scheme. The general large model provides the theoretical and technical basis for the completion of this task. Users can describe their needs in natural language, graphics, pictures or component arrangement. After the task decomposition and design, HAI-OS can further confirm, verify and adjust new requirements with users. HAI-OS uses the large model and user's multiple interactions to finally complete the implementation scheme, including the matching AI capabilities, models, data, computing network resources and test methods. HAI-OS efficiently schedules and manages the connected computing resources and AI capabilities in an open, real and dynamic environment, serving a wide range of intelligent

needs. Its core functions include the standardization of computing resources, AI capabilities, data, models, intelligent business description, the operation management of AI services, and the mechanism of quantification, evaluation, verification, security and controllability.

Through the cooperation of the above blocks, HAI integrates the advantages of large and small models and balances the model of large model training and service with the needs of intelligence.

8 Mechanisms of holistic AI

The relationship between HAI and the large model is shown in Figure 2. Both large models and small and medium-sized models can be used as components of HAI. The HAI system emphasizes the collaboration and mutual learning mechanism between large and small models. Small models can be derived from larger base models. The related technologies derived from models include teacher-student models, quantification, compression and pruning. Small and medium-sized models can also be completely independent. However, how to transfer the intelligence of small and medium-sized models to large models and collaborate with them has not been fully studied at present.

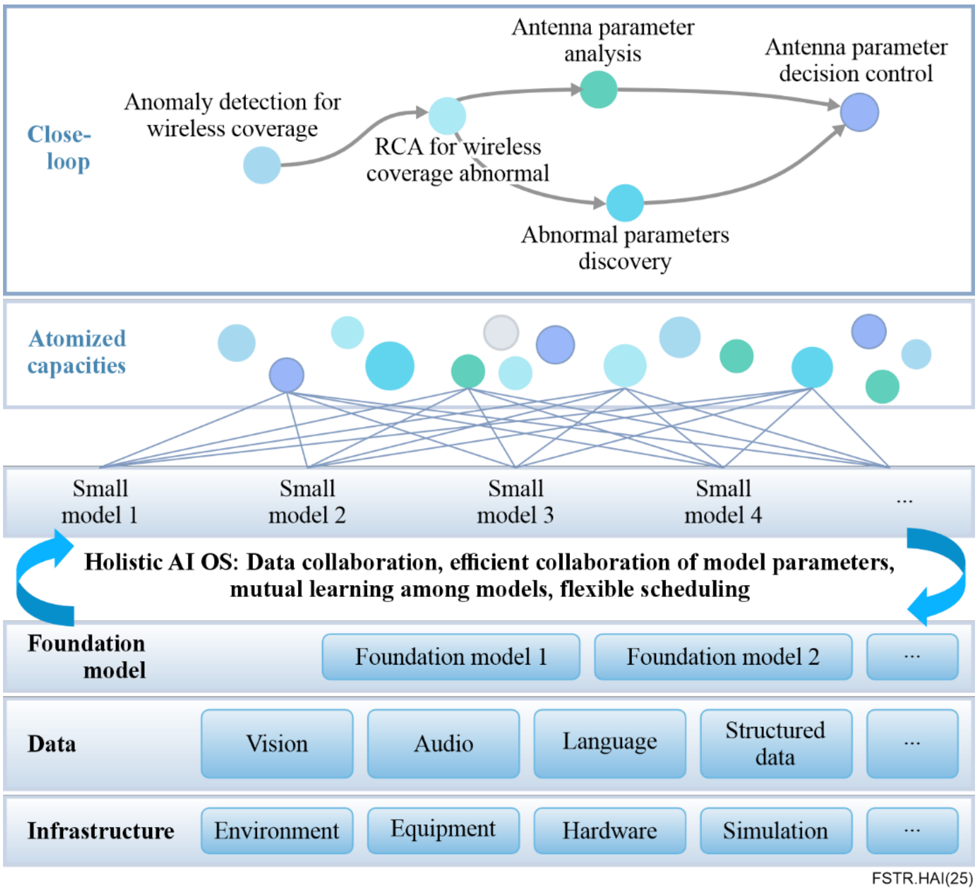

Close-loop
Anomaly detection for wireless coverage
RCA for wireless coverage abnormal
Antenna parameter analysis
Antenna parameter decision control
Abnormal parameters discovery
Atomized capacities
Small model 1
Small model 2
Small model 3
Small model 4
...
Holistic AI OS: Data collaboration, efficient collaboration of model parameters, mutual learning among models, flexible scheduling
Foundation model
Foundation model 1
Foundation model 2
...
Data
Vision
Audio
Language
Structured data
...
Infrastructure
Environment
Equipment
Hardware
Simulation
...
FSTR.HAI(25)

Figure 2 – AI foundation models to HAI

HAI-OS needs to understand business language and match scheduling and computing network resources to meet business needs. Therefore, HAI-OS will rely on large models to understand complex business requirements, perceive and understand computing network resources and models, and thus be able to build a bridge in between. HAI is a technical system and framework that supports the implementation of large-scale model applications.

9 Enabling technologies of holistic AI

9.1 AI-exchange system

AI-exchange is the scheduling strategy and framework for HAI, i.e., the scheduling for AI models, computing and network resources based on demands and requirements of business. AI-exchange is also an operation and bidding system between the "demand side" and "supply side" for AI resources and capabilities, and the AI resources and capabilities can be subject to real-time scheduling to meet business requirements and demands. The main processes are as follows:

- The operator/provider reports the demand/requirement information of AI foundation models to AI-exchange; in the meantime, the demand/requirement information can also be polled by AI-exchange.
- Understanding the diverse and complex intelligent demands/requirements of users, and developing the relevant solution of capability testing; broadcast the testing solution and the foundation model capabilities, which will be provided by the model provider through online bidding.
- Run and complete the planned test, judge whether the testing solution can meet/satisfy the user's demands/requirements, determine the final executive solution and return to the first or second step if adjustments are needed to the requirements or solution.
- The final broadcasting executive solutions, related model information and security requirements are suggested to be provided by the computing resource providers/operators through bidding as needed, to offer a suitable solution.
- Test the solutions; if passed, the computing resource provider can initiate the operating solution and provide relevant services to stakeholders.

The AI-exchange makes operators/providers, "demand side" and "supply side" independent from each other. The stakeholders can define a clear understanding of their preference and find the most suitable solution. The foundation model providers can focus on the AI models, other than binding with the providers of computing resources. The providers/operators can focus on improving computational and transmission efficiency, in the meantime, optimizing the execution of models. See also Figure 3.

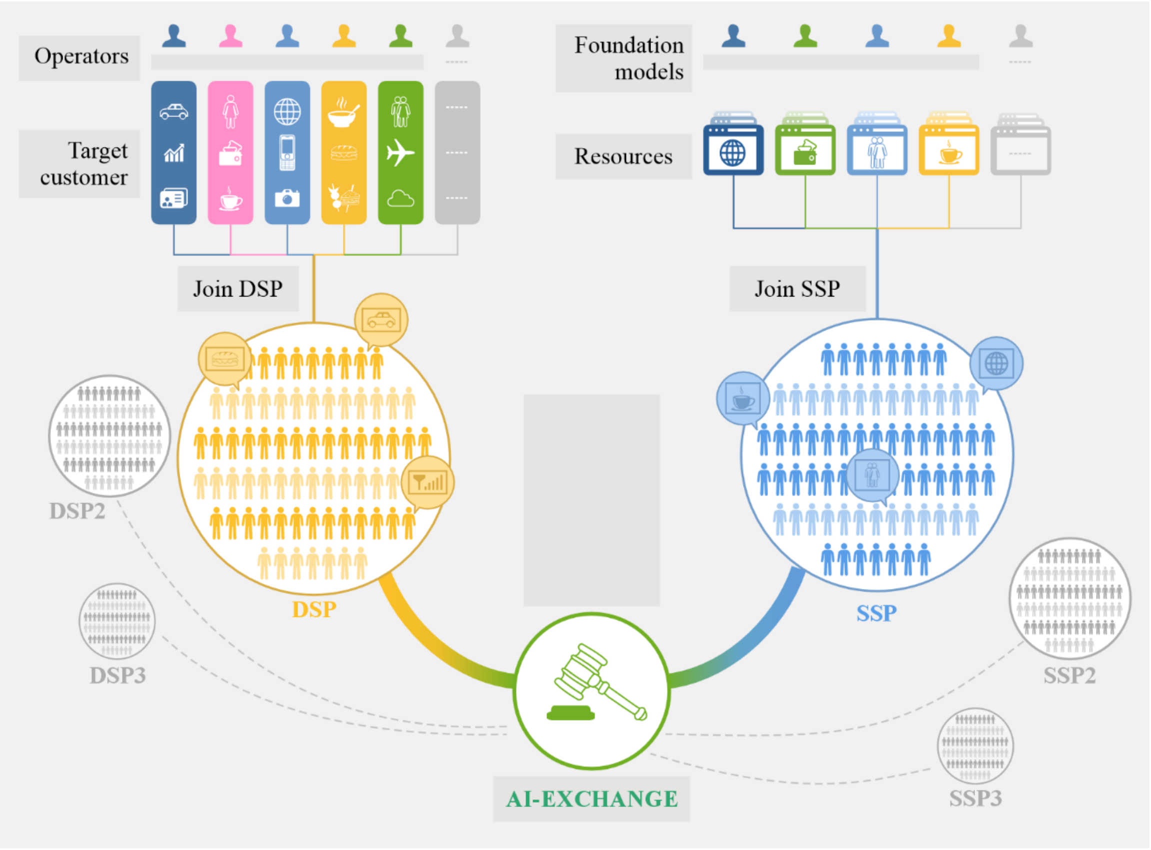


Figure 3 – AI-exchange system of HAI

9.2 AI capability atomization

The generalization of intelligent services requires not only a universal large model in the cloud but also flexible specialized small and medium-sized models that focus on completing typical tasks with high repetition rates. These small and medium-sized models can be trained independently or derived from larger models. From the dimensions of perception, cognition and prediction from the data dimensions of speech, vision, natural language processing, structured data and from the application scenario dimensions of indoor or outdoor, a series of common specialized abilities can be abstracted, and the required models can be constructed. The atomized definition of AI capability is the foundation for using AI-exchange to match AI models and decompose complex tasks. Although the overall intelligence of general models is rapidly improving and can complete almost infinite cognitive tasks, in the future, general models and specialized models will coexist for a long time, just as work and life require not only a large number of educated ordinary people but also a large number of professional talents to focus on specific tasks. The definition of atomic models and atomic capabilities needs to meet several standards: high reusability, clear input–output and functionality, not too small to result in model collaboration costs exceeding computational costs, suitable for independent research and development, and complementary to basic model capabilities.

9.3 End-to-end close loop optimization

The existing general frameworks can be divided into three categories: designing new model structures to adapt to cascading multitasking, end-to-end optimization mechanisms for white box models and end-to-end optimization mechanisms for black box models. The HAI optimization mechanism belongs to the third category, i.e., the E2E optimization mechanisms for black box models.

The first type is to design a new model structure to adapt to the cascading of multiple tasks. An E2E differentiable inference architecture to implement multistep inference tasks was proposed, where each step of the model's inference process is implemented by a memory attention and composition (MAC) module, and multiple MAC modules can be connected in series to achieve multistep inference. Each MAC module consists of a memory unit, a read-in unit and a write-in unit, and multiple MAC units connected can still perform end-to-end differentiable optimization.

It is believed that reasoning is the ability to process previously acquired knowledge and derive new inferences or answer new questions. Reasoning is a fundamental component of an intelligent agent, and typical reasoning processes include transfer relationships, logical relationships, mathematical calculations and comparisons. However, neural network models, including basic foundation models, perform a large-scale correlation calculation rather than such logical combinations. On the one hand, MAC imitates the design pattern of contemporary computer architecture by separating control and memory units, thereby breaking down complex tasks into multistep inference processes. At the same time, MAC uses a serialization cascade method to ensure that multiple MAC modules can still be optimized end-to-end based on back propagation algorithm after connection. MAC achieved the optimal results at that time in multiple inference tasks based on visual information proposed a cyclical memory converter model that uses modules similar to cyclical memory units to expand the length of inputs that the converter model can accept, objectively achieving the separation of memory. The lower layer of converters is dense and shared, while the forward part of the upper layer of converters is sparse and selectively activated by random routing experts. Input from different domains is represented by different words. Different fields correspond to different experts at the upper level. Therefore, a domain model can be derived from the base model for different domains.

The second type is the joint optimization mechanism of white box models. The white box model refers to a model whose structure and parameters are completely open. The idea of deep model reassembly (DeRy) was proposed. Given a series of heterogeneous foundation models trained from different data sources, DeRy first decomposes the foundation model into multiple building blocks, where each model building block corresponds to multiple deep network layers. These building blocks are recombined to form a new model that meets functional needs and resource constraints. DeRy uses a coverage set optimization method to decompose various pre-trained models and search for similar functional blocks in the building blocks to form a set. The building blocks in a similar set can be selected as needed to meet various constraint conditions. The model reorganized through the above method can achieve excellent accuracy on image net classification tasks without training, and the fine-tuned classification accuracy can be further improved by 4.6%. However, the above-mentioned joint restructuring method relies heavily on the partitioning strategy of the model proposed.

A more concise approach for joint optimization of open-source white box models is to use a 1D convolutional neural network to stitch the layers of different models together and perform end-to-end optimization. In the optimization process, only the parameters of the stitching layer are updated, and the initialization methods for the stitching layer parameters include minimum variance and Kaiming initialization. The parameter optimization method continues to use stochastic gradient descent. Using a one-dimensional convolutional neural network as a suture layer can suture a variety of visual models, such as a standard diffusion model, hierarchical visual converter and convolutional neural network. Experiments show that suture models can achieve accuracy similar to a single model in image net classification tasks. They also conducted extensive experiments to explore how to select the models to be stitched, which model layers to stitch between and the effectiveness of different optimization methods. The advantage of this method is that it can reuse the intelligence learned by many foundation models, providing the possibility of mixing multiple models, but the search space of the models is large and time-consuming. The function of the stitching layer is to map the activation features of one model layer to the feature space of another model layer. However, there are significant differences in the model structure and activation features learned by different domains, and it is insufficient to fit these differences solely through feature space mapping. Therefore, this method is currently only applicable to models in the stitching domain.

9.4 Network native AI

The purpose of HAI is to make intelligence ubiquitous, and the convenience and ubiquity of intelligence are equally important as the generation of intelligence. The supply of intelligence relies on ubiquitous communication networks to securely, accurately and cost effectively deliver AI services to every user. The evolution of current AI algorithms and AI computing chips are mutually reinforcing and complementary. AI algorithms are designed based on the characteristics of current AI computing power. Current AI models and algorithms can efficiently perform parallel computing, accelerate inference and share model parameters, data and tensors on AI computing power represented by GPUs. It can be said that AI algorithms understand computing power, but not network transmission AI algorithms. The starting point for proposing research on network native AI is to fill this gap, allowing AI algorithms to fully consider the characteristics of communication network transmission from the design stage, so that AI models can efficiently collaborate between computing cloud edge ends and become ubiquitous and secure ubiquitous services like the fifth generation of mobile communication systems (5G).

In response to the demand for the collaborative evolution of computing power network cloud edge models, it is proposed to study the scheduling, orchestration and optimization techniques of computing power network cloud edge models from the application layer dimension. An adaptive and highly elastic model scheduling and orchestration optimization method is designed to optimize service orchestration and training task scheduling strategies based on the results of computing power network resource situational awareness, supporting low latency edge cloud model migration. The scheduling, orchestration and optimization techniques for edge cloud models mainly include training task scheduling and resource allocation. Firstly, the nodes of the computing power network need to be integrated to model the computing power resources in the network. Through the integration and fusion of computing power network resources, the free migration of micro services and optimal scheduling of task requests between different computing power nodes and between computing power nodes and network nodes can be achieved, thereby achieving the sub task goal of optimal utilization of computing power network resources and minimizing the training delay of the size model. Secondly, it is necessary to enhance the inference mechanism of AI models to maximize their utilization of network capabilities, in order to achieve low latency edge cloud model migration and provide efficient service to users.

9.5 Trustworthiness

The security guarantee of AI data, AI models and AI capabilities in services is an important enabler for HAI to serve users on a large scale. Secure and trustworthy AI adheres to the principle of "AI for good" and focuses on tackling the systematic framework that ensures the full process of AI services can be supervised, including researching the basic theories and methods of traceability, trustworthiness, auditability, accountability, value consistency and attack prevention. The security and trustworthiness of AI models, especially the security and controllability of large models, has recently become a hot topic in social discussions and technological breakthroughs with the rapid popularity of large models. In HAI, more emphasis is placed on the protection of data security, model security and business security in large-scale operations.

Bibliography

[b-Liang] Liang, L. et al. Kag. (2025), *Boosting llms in professional domains via knowledge augmented generation* In Proceedings of the Companion Proceedings of the ACM on Web Conference, pp. 334–343.

[b-Liu] Liu, J.; Wang, Q.; Wang, J.; Cai, X. (2024), *Speculative Decoding via Early-exiting for Faster LLM Inference with Thompson Sampling Control Mechanism*, In Proceedings of the Findings of the Association for Computational Linguistics ACL 2024. pp. 3027–3043.

[b-Lu] Lu, J.; Pang, Z.; Xiao, M.; Zhu, Y.; Xia, R.; Zhang, J. (2024), *Merge, ensemble, and cooperate! A survey on collaborative strategies in the era of large language models,* arXiv preprint arXiv:2407.06089.

[b-Qu] Qu, C.; Dai, S.; Wei, X.; Cai, H.; Wang, S.; Yin, D.; Xu, J.; Wen, J.R. (2025), *Tool learning with large language models: A survey*. Frontiers of Computer Science, Vol. 19, 198343.

[b-Tak] Tak, D., et al. (2024), *A foundation model for generalized brain MRI analysis*, medRxiv.

[b-Wang] Wang, F., et al. (2024), *A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness*, arXiv:2411.03350.

[b-Wu] Wu, J.; Zhu, J.; Liu, Y.; Xu, M.; Jin, Y. (2025), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, *Agentic Reasoning: A Streamlined Framework for Enhancing LLM Reasoning with Agentic Tools*.; Volume 1: Long Papers, pp. 28489-28503.

[b-Xi] Xi, Z., et al. (2025), *The rise and potential of large language model based agents: A survey*, Science China Information Sciences, Vol. 68, 121101.

[b-Zhang] Zhang, Y. et al., (2025), *Qwen3 Embedding*: *Advancing Text Embedding and Reranking Through Foundation Models,* arXiv preprint arXiv:2506.05176.