# Large-Model-Driven Intelligent IoT Decision Systems

Jiaye Liu *

College of Information Science, Yunnan University of Finance and Economics, Yunnan, 650300, China

* Corresponding Author Email: saber0212jjj@gmail.com

**Abstract.** With the development of Internet of things (IoT) technology, the continuous increase of IoT devices and the increasing complexity of application scenarios, the traditional decision system relying on rules has been difficult to meet the intelligent requirements of high dynamic, multi-modal, and global collaboration. In recent years, Large Language Models (LLM) have gradually become an important technical support for building a new generation of IoT decision systems with excellent semantic understanding, cross-modal reasoning, and contextual learning capabilities. This paper reviews the method implementation of large models in IoT decision-making systems, with a focus on elaborating key technical paths such as multimodal input fusion, semantic reasoning and decision generation, reinforcement learning collaboration, edge-cloud collaboration and federated learning, and analyzes the current engineering challenges and research directions. This article emphasizes that the IoT decision making system driven by large models not only promotes the transformation of the traditional IoT from a "data-driven" to a "semantic-driven" approach, providing a feasible path for the practicality of large models in real IoT scenarios, but also offers a technical foundation and research direction for the future development of fields such as intelligent manufacturing and smart cities.

**Keywords:** Internet of Things (IoT); Large Language Models (LLMs); Multimodal Semantic Fusion; Reinforcement Learning; Federated Learning.

## 1. Introduction

With the rapid expansion of the scale of IoT, the number of smart terminals is growing exponentially, and the complexity of data interaction among devices is constantly increasing. Traditional IoT decision systems usually rely on rule-driven or lightweight machine learning models, which are difficult to cope with the real-time processing and global decision-making requirements of multi-source heterogeneous data in dynamic environments. This structural contradiction of "strong perception but weak decision making" has become one of the key bottlenecks restricting the intelligent upgrade of IoT.

In recent years, Large Language Models (LLMs) have shown extensive potential in the decision-making of complex systems due to their powerful semantic understanding and reasoning capabilities. Through the expansion of parameter scale and in-context learning ability, large models can complete multi-task decision making and knowledge transfer without retraining, providing a unified cognitive and decision-making framework for distributed agents in IoT systems [1]. For example, OpenAI's GPT-4 and Google's Gemini family have achieved significant breakthroughs in multimodal perception, complex task decomposition, and policy generation [2]. The introduction of these models offers more possibilities for the decision-making process of the IoT and provides new paths for achieving adaptive and autonomous IoT agents. However, directly applying large models to IoT environments still faces challenges such as the uneven distribution of computational resources, data privacy and security, and real-time constraints. Current research is actively exploring the "large model compression for the edge" and "collaborative inference" strategies to maintain the decision-making performance of the model in the limited resource environment [3].

The advancement of these research directions not only promotes the deep integration of AI and IoT, but also lays a technical foundation for the construction of the next-generation intelligent IoT decision system. This paper will sort out the key methods and paths of large models in the IoT decision system, and construct a unified IoT system framework from multimodal perception, natural

language input and semantic reasoning to generating decisions. Firstly, how to realize the unified semantic encoding input of multimodal data, so that complex IoT device data can be easily processed by large models. Secondly, the decision-making ability of large models in complex scenarios of IoT work was analyzed, and the applicability of methods such as Chain of Thought (CoT) and Low-Rank Adaptation of Large Language Models (LoRA) was discussed. To enhance the model's accuracy and interpretability in complex task decision-making, we designed task-specific prompts, elicited its reasoning processes, and performed small-scale fine-tuning. This work investigates the synergy between large language models and reinforcement learning, aiming to enhance decision reliability and achieve self-consistent, adaptive cognitive-action cycles. Finally, combined with the issues of privacy and computing power, we discuss the feasibility of edge-cloud collaborative architecture and federated learning in practical deployment. Based on the above content, this paper aims to provide a systematic analysis and technical references for the construction of a new generation of highly reliable and highly interpretable intelligent IoT decision-making systems.

## 2. Overview of the System Architecture

### 2.1. Four Layers of Layered Design.

The system adopts a layered architecture to balance the real-time performance and model abilityFirstly, the original data such as sensor time series, images, speech and text logs are collected in the Perception layer. The Edge layer performs data preprocessing, low-latency rules and model lightweight inference. The processed data were transmitted to the Cognitive Decision layer, which took the large model as the core to undertake cross-modal semantic fusion, reasoning and decision generation. Finally, through the Execution & Feedback layer, the decision is converted to the device control command, and the effect is returned for online/offline update.

### 2.2. Principles of Design.

In practice, systems should follow the engineering principle of ‘edge preprocessing + centralized or collaborative inference + local closed-loop execution’ to reduce latency, protect privacy, and preserve the interpretability of decisions. The edge layer undertakes data preprocessing such as fast data filtering and lightweight computation to reduce the upward pressure of the original data. In the centralized cognitive decision layer, the collaborative reasoning module was responsible for the complex semantic understanding and reasoning decision-making of the task, and carried out the global coordination and optimization of each layer to ensure the accuracy and consistency of decision-making. The local closed-loop execution mechanism can maintain the autonomous operation and safe response of the system in the case of network fluctuation or delay. Through hierarchical collaboration, the system can achieve a balance between real-time performance, accuracy and stability in the case of limited local computing power.

## 3. Core Path

### 3.1. Multimodal Input Fusion and Unified Semantic Encoding

In IoT working environment, data forms are diverse, including time series, images, voice, and text data, which need to be mapped into a unified feature space for large model processing. To solve the above problems, multi-modal fusion encoders (e.g., convolutional or Transformer-based encoders for time-series data, convolutional or vision-Transformer encoders for visual inputs, and embedding-based encoders for text) can be used for large models to do cross-modal inference. This idea is inspired by the general sequence modeling capabilities of Transformers [4], and the success of vision-language joint representations such as CLIP on cross-modal pairing tasks [5].

DeRieux et al. proposed a transformer-based system to support various input and output tasks for IoT data in smart cities using a pure encoder architecture with interchangeable input embeddings and

multiple task heads [6]. They combined time-series sensor data, visual plant disease classification, and image tasks into a model for processing [6]. Experimental results show that when dealing with multimodal tasks, the system not only outperforms some baseline models in memory and computational efficiency, but also has the same performance as the single task, which indicates that the method has engineering feasibility and versatility in IoT scenarios [6].

The engineering example also further reflects that the data lightweight preprocessing at the edge can reduce the uplink transmission delay of the original high-dimensional data. Edge nodes first perform denoising, sampling, and lightweight embedding. The aggregated embeddings are then sent to central or near-edge nodes for deeper semantic fusion, so as to ensure both efficiency and accuracy in resource-constrained IoT environments.

### 3.2. Semantic Reasoning and Decision Generation

The decision output not only includes control commands (such as PLC calls, equipment switches), but also includes underlying decision rationale, confidence levels and risk assessments. Therefore, the system needs to be accurate and interpretable to meet the auditable requirements of industrial and security scenarios. Introducing Chain-of-Thought (CoT) reasoning provides an effective solution to this need.

Chain-of-Thought (CoT) is a prompt strategy to enhance the reasoning capability of large language models. By explicitly allowing the model to output a step-by-step reasoning process before generating the final answer, the accuracy and interpretability of complex logic tasks are improved [7].The core idea is to decompose the problem into a series of coherent intermediate reasoning steps, so that the model forms a structured path of "step-wise thinking" during the generation process.In the IoT decision system, CoT can be used to generate decision recommendations with causal logic chains, from state perception to risk assessment, and finally to realize control actions, thus significantly improving the trustworthiness and auditability of the system in multi-constraint decision scenarios.

Chain-of-Thought (CoT) reasoning significantly enhances the interpretability and accuracy of complex decision-making tasks [7]. In practice, models can be enabled to generate structured decision drafts through carefully crafted prompts or fine-tuning.

For specific IoT tasks, to avoid the high cost of full fine-tuning, Parameter-Efficient Fine-Tuning (PEFT) techniques such as Low-Rank Adaptation (LoRA) can be employed. These methods allow for small-scale adjustments, enabling the model to rapidly adapt to new task environments or sensor modalities at significantly reduced costs, while improving task-specific accuracy [8].

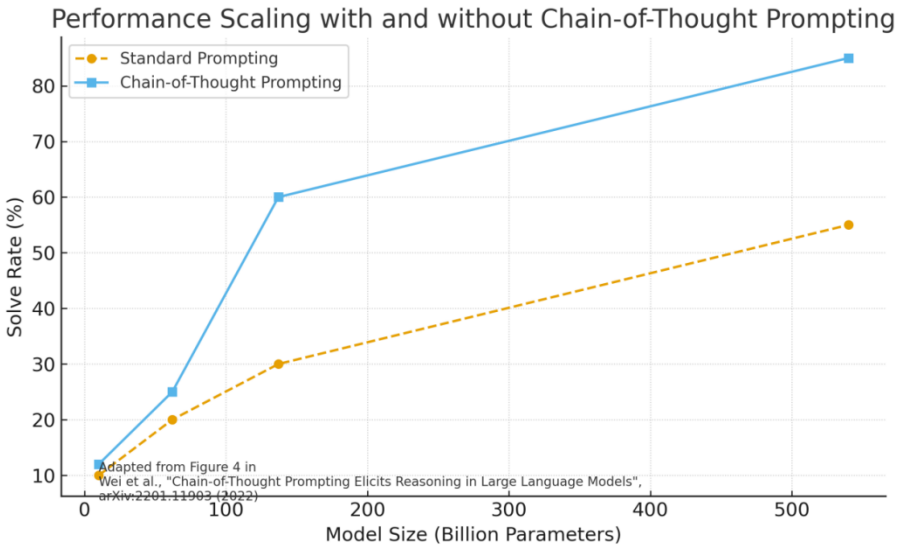


**Fig. 1** Performance scaling of few-shot prompting with and without chain-of-thought reasoning [7].

Figure 1 shows the redrawn performance comparison plots based on the CoT research results proposed by Wei et al. This figure compares the performance of large models in few-shot inference

tasks under different strategies, indicating that CoT can significantly improve the accuracy of the model in complex tasks, and the model can still maintain stable performance improvement when the task size and reasoning difficulty continue to increase.

### 3.3. Synergy between Reinforcement Learning and Large Models: The Cognition-Action Loop

Reinforcement learning (RL) is a sequential decision-making method based on the "state-action-reward" feedback mechanism, which aims to maximize the cumulative reward through interaction with the environment.RL uses Policy to guide the agent to take the optimal action in different states, and continuously updates the policy function according to the feedback signal.

In IoT control tasks, RL can make the device learn the optimal control policy in the dynamic environment through long-term interaction.For example, in intelligent energy management or traffic flow scheduling, RL can adaptively optimize global system performance based on historical behavior and real-time feedback.When combined with the large model, RL can improve the efficiency and robustness of policy generation by using the semantic reasoning ability of LLM, so as to realize the "Cognitive-Action Loop" intelligent decision system.

Many IoT control problems are sequential decision-making problems in nature and lend themselves to reinforcement learning (RL). Incorporating Large Models as the 'semantic reasoning module' within RL policies allows the system to balance long-term rewards with contextual reasoning capabilities. Decision Transformer introduces the idea of sequence modeling into RL, and directly generates action sequences through Transformers or autoregressive models, which provides a feasible path for IoT control [9]. Chen et al. demonstrated on the D4RL dataset, such as robot grasping, navigation, industrial control, and other tasks, that the sequence modeling method can generate stable action sequences in IoT control scenarios with multi-sensor inputs, and can achieve comparable or even better performance than traditional RL methods in multi-sensor IoT control tasks. Thus, the engineering feasibility of DT in IoT scenarios is demonstrated [9]. In engineering practice, the candidate actions or policy fragments generated by the large model can be used as the supplementary input of DT, so that the system can improve the global consistency and interpretability of the policy while ensuring the reasoning efficiency.

### 3.4. Edge-cloud Collaboration and Federated Learning

Federated learning is a distributed machine learning framework that allows multiple edge devices to co-train a global model without sharing raw data [10]. Assuming that there are k clients (devices) and each client k has a local dataset $P_k$ with a data size of $n_k$, the authors formalize the federated optimization problem as Eq. (1).$F_k(\omega)$ is the average loss over client k.

$$\mathrm{f}(\omega) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\omega). \tag{1}$$

In traditional methods, all samples are regarded as coming from the same data set, and are uniformly calculated and processed on the server. Conversely, Federated Learning utilizes a distributed structure weighted by client data size $n_k$,preserving the statistical features of both local data $P_k$ and loss information $F_k(\omega)$.Since the objective function is decomposed into the local term $F_k(\omega)$ of each client, the client can perform local multi-step gradient descent, which allows each terminal device to complete the local model update without uploading the original data, improves the calculation amount of each round, reduces the number of communications with the server, and enhances the privacy protection [10].In IoT scenarios, FL is especially suitable for data-sensitive application environments such as medical, industrial monitoring and smart home.By combining model distillation and encrypted communication mechanism, FL can realize cross-device knowledge sharing and global optimization under the premise of ensuring privacy and security, and effectively deal with the problem of data silos in edge computing scenarios.

The deployment mode can adopt the "lightweight local model + central aggregation" mode, where the local model is responsible for real-time response and embedding extraction, and the cloud or near

end server is responsible for global model aggregation and centralized processing of LLM inference requests.Model distillation techniques, such as DistilBERT, can be used to generate compressed model versions suitable for deployment on edge devices [11].This study is not a simple truncated model, but a "three-in-one" loss function strategy based on the lightweight model.

This strategy combines a traditional Masked Language Modeling (MLM) loss with a Distillation Loss over the core label probability distribution, and a Cosine Embedding Loss.This enables DistilBERT to significantly reduce GPU memory footprint and improve inference speed, while maintaining generalization capabilities comparable to the original model.

## 4. Engineering challenges and countermeasures

In the actual deployment, it is first faced with the problem of lightweight computing resources and models. The large pre-trained model has a massive parameter counts, which is not conducive to edge deployment. Countermeasures include efficient parameter optimization methods such as knowledge distillation, quantization, LoRA, and heterogeneous computing (GPU/TPU/AI accelerator) collaboration. Privacy and security challenges, federated learning, differential privacy, and trusted execution environment should be combined to reduce data leakage and model inversion risk [8].In high-stakes industrial or medical environments that demand rigorous interpretability and compliance, the system produces structured decisions containing reasoning chains and evidence citations—primarily via Chain-of-Thought (CoT)—to facilitate auditing and human intervention. Finally, the scene transfer ability is required under specific tasks: cross-scene transfer is realized through In-Context Learning and Parameter Efficient Fine-Tuning (e.g., LoRA) with a small number of examples to reduce the need for large-scale retraining [5].

## 5. Conclusion

This paper presents a systematic review of Large Model-driven IoT decision-making systems, delving into critical areas such as unified multi-modal semantic encoding, the integration of Chain-of-Thought (CoT), Reinforcement Learning, and Federated Learning, as well as resource optimization strategies for Edge-Cloud collaboration. The results show that the large model can significantly improve the semantic understanding ability, information fusion ability, and decision-making ability of IoT systems in complex environments. At the same time, by integrating the multi-modal coding, RL interaction strategies, and cross-device collaboration and privacy protection mechanisms of federated learning mentioned in the article, it can effectively alleviate the common problems in traditional IoT environments, such as limited computing resources, complex data types, privacy sensitivity, and difficulties in cross-scenario transfer, thereby enhancing the flexibility of decision-making in IoT systems in real-world scenarios.

The contributions of this research are mainly reflected in three aspects. Firstly, this paper constructs a system framework including multimodal perception, semantic reasoning, policy generation, and collaborative deployment, providing structured theoretical support for the role mechanism of large models in the IoT. After that, the main technical paths of big model landing in real IoT scenarios were clarified, which provided a reusable technical reference for industry practice. In conclusion, this work emphasizes interpretability, lightweighting, and collaborative computing as pillars of IoT decision-making, establishing a methodological basis for next-generation trustworthy, autonomous AIoT and smart city infrastructures.

## References

[1] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models. 2023. https://arxiv.org/abs/2302.13971.

[2] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. 2023. https://arxiv.org/abs/2303.08774.

[3] Singh R, Gill S S. Edge AI: a survey. Internet of Things and Cyber-Physical Systems, 2023, 3: 71-92.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 6000-6010.

[5] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021, 139: 8748-8763.

[6] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin: Harbin Institute of Technology, 2011.

[7] DeRieux A C, Saad W, Zuo W, et al. A transformer framework for data fusion and multi-task learning in smart cities. 2022. https://arxiv.org/abs/2211.10506.

[8] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[9] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022.

[10] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, FL, USA: PMLR, 2017: 1273-1282.

[11] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019. https://arxiv.org/abs/1910.01108.