

Composable Modular Intelligence: A Memory-Centric Modular AI Architecture for Adaptive Cognitive Systems

Suganthi J, Dr.B. Jaishanthi

suganthijothi1991@gmail.com, bjaishanthi@tagore-engg.ac.in

Department of Computer Science and Engineering
Tagore Engineering College
Rathinamangalam, Chennai.

Abstract

Recent advances in Artificial Intelligence (AI) have shown impressive capabilities using large neural architectures like Mixture of Experts (MoE) and efficient fine-tuning methods such as LoRA. However, these systems mainly assume that intelligence exists within the weights of a single neural network. This paper introduces a new framework called Composable Modular Intelligence (CMI). It redefines intelligence as the coordinated interaction among separate cognitive modules. The architecture includes specialized micro-models, an orchestration engine for smart routing, and a multi-layer external memory system that can adapt in real-time without needing weight retraining. This architecture promotes scalable, flexible, and personalized AI systems through modular composition, contextual reasoning, and continuous memory evolution. This framework aims to overcome the limitations of existing MoE and LoRA systems in terms of adaptability, runtime flexibility, and memory-driven thinking.

Keywords— Modular AI, Mixture of Experts, LoRA, Cognitive AI, Memory Systems, AI Orchestration, Composable Intelligence.

INTRODUCTION

Artificial Intelligence systems are moving from single large architectures to distributed and composable systems. Current leading systems like Mixture of Experts (MoE) and efficient fine-tuning methods offer scalability and computational efficiency, but they still depend heavily on centralized neural representations. These architectures often lack the ability to adapt

dynamically, extend cognitive functions in real time, and maintain persistent memory-driven reasoning.

Human cognition works through the coordination of various specialized cognitive processes including memory, reasoning, language understanding, and decision-making. Inspired by this concept, this work presents Composable Modular Intelligence (CMI), a modular AI framework in which intelligence arises from the interaction of independently trained micro-models managed by a lightweight orchestrator.

The proposed system merges modular neural reasoning, an external memory infrastructure, and adaptive orchestration methods to support scalable and lifelong learning systems. Unlike traditional monolithic models, this architecture allows for interchangeable cognitive modules, real-time memory evolution, and selective task-oriented reasoning.

II. LITERATURE REVIEW

A. Modular and Composable AI Architectures

S. Jain et al. proposed the "Composition of Experts" architecture that dynamically activates specialized large language models through routing methods. Their work enhanced scalability and efficiency in modular AI systems. However, the approach concentrated mainly on expert routing and did not consider persistent memory-driven cognition or real-time personalization.

B. AlKhamissi et al. introduced modular reasoning systems inspired by biology through specialized cognitive modules for logical reasoning, social reasoning, and language reasoning. While this system improved understanding, the modules remained closely

linked within transformer architectures, limiting their ability to extend in real time.

C. AI Orchestration and Compound Intelligence Systems

Berkeley AI Research (BAIR) highlighted the shift from isolated models to compound AI systems that incorporate orchestration layers, retrieval systems, and tool usage pipelines. Although the framework stressed system-level intelligence, it did not present a formal modular cognitive architecture.

Rudra and Agrawal proposed AI pipelines that integrate orchestration, memory systems, contextual routing, and tool coordination. Their focus was primarily on software orchestration rather than flexible modular cognition.

D. Memory-Centric Cognitive AI Systems

K. Wedel introduced Contextual Memory Intelligence (CMI), positioning memory as an adaptive reasoning framework for collaboration between humans and AI. Similarly, S. Jia et al. suggested a unified memory system influenced by human episodic, semantic, and procedural memory.

Despite progress in memory-driven AI systems, current strategies do not support dynamic orchestration among independently deployable cognitive modules. Additionally, real-time composition of modules and adaptive memory evolution have not been widely explored.

III. RESEARCH GAP

The literature review reveals several shortcomings in current AI architectures:

- Existing modular systems tend to focus on expert routing rather than adaptive cognition.
- The ability to extend cognitive modules in real time is limited.
- Current memory systems are not fully modularized.
- Most methods rely on retraining or fine-tuning for personalization.
- There is a notable absence of dynamic task-specific composition of independent AI modules.

These limitations highlight the need for a flexible, memory-focused, and composable AI architecture.

IV. PROPOSED WORK

The Composable Modular Intelligence framework introduces four main components:

A. Micro-Models

Micro-models are lightweight, task-specific models trained independently for purposes such as language understanding, legal reasoning, sports analytics, and programming help. Unlike MoE experts confined to a single neural network, these models can be deployed and replaced independently.

B. Orchestrator

The orchestrator is a lightweight coordination model responsible for:

- Task breakdown
- Module selection
- Context-aware routing
- Coordinating multi-step reasoning

The orchestrator learns decision-making strategies instead of storing domain knowledge.

C. NeuralVault Memory System

NeuralVault offers a three-layer external memory infrastructure:

- Episodic Memory – Stores interaction histories using vector databases.
- Semantic Memory – Maintains structured knowledge graphs.
- Procedural Memory – Keeps reinforcement learning feedback and execution strategies.

This memory system evolves continuously during runtime without needing updates to neural weights.

D. Feedback Loop

A continuous feedback loop assesses task outcomes, updates memory structures, and improves orchestration strategies for better performance in the future.

V. SYSTEM ARCHITECTURE

The proposed architecture replaces centralized intelligence with coordinated modular intelligence.

Component

Function

Replaces

Micro-Models

Task-specific modular AI systems

MoE Experts

Orchestrator

Smart routing and coordination

Gating Network

NeuralVault

Adaptive runtime memory system

LoRA/Fine-Tuning

Feedback Loop

Continuous learning and optimization

Static Deployment

The architecture supports selective module activation, efficient resource use, and scalable cognitive growth.

VI. TRAINING PIPELINE

A. Micro-Model Training

Each micro-model is trained on domain-specific datasets using knowledge distillation from larger teacher models. The goal is to achieve low latency and reliable task specialization.

B. Orchestrator Training

The orchestrator trains on synthetic multi-hop tasks that require coordination among multiple modules. Reinforcement learning is used, where the reward is based on successful task completion instead of predicting the next token.

C. Runtime Learning

Memory systems evolve continuously based on user interactions and feedback from the environment. Interaction logs are used periodically for offline retraining of orchestration policies.

VII. COMPARISON WITH EXISTING SYSTEMS

Problem

MoE/LoRA Approach

Proposed CMI Approach

Scalability

Larger Models

More Modular Components

Adaptation

Weight Fine-Tuning

Memory Evolution

Efficiency

Sparse Activation

Selective Module Loading

Personalization

Retraining Required

Runtime Memory Adaptation

Extensibility

Limited

Plug-and-Play Modules

VIII. ADVANTAGES OF THE PROPOSED SYSTEM

The proposed system provides several benefits:

- Modular and scalable architecture
- Real-time cognitive adaptability
- Better personalization without retraining
- Reduced computational load
- Improved understanding through specialized modules

- Support for lifelong learning and contextual reasoning

IX. CONCLUSION

This paper introduced a novel Composable Modular Intelligence framework that redefines intelligence as the coordinated interaction among modular cognitive systems. The architecture combines independently deployable micro-models, orchestration-based reasoning, and a multi-layer adaptive memory system. This framework addresses the limitations of traditional AI architectures by allowing for real-time flexibility, memory-driven adaptation, and personalized cognitive growth. Future work will focus on real-world deployment, optimizing distributed orchestration, and integrating multimodal reasoning systems.

REFERENCES

- [1] W. Cai et al., "A Survey on Mixture of Experts in Large Language Models," IEEE Transactions on Knowledge and Data Engineering, 2025.
- [2] "Parameter-Efficient Fine-Tuning in Large Language Models," Artificial Intelligence Review, Springer, 2025.
- [3] "Parameter-Efficient Fine-Tuning of Large Language Models Using Semantic Knowledge Tuning," Scientific Reports, vol. 14, Art. no. 27890, 2024.
- [4] J. Wu et al., "Omni-Smola: Boosting Generalist Multimodal Models with Soft Mixture of Low-Rank Experts," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14205–14215.
- [5] S. Mu and S. Lin, "A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications," arXiv preprint arXiv:2503.07137, 2025.
- [6] R. Hwang et al., "Pre-Gated MoE: An Algorithm-System Co-Design for Fast and Scalable Mixture-of-Expert Inference," in Proceedings of the ACM/IEEE International Symposium on Computer Architecture (ISCA), 2024.
- [7] D. Li et al., "MixLoRA: Enhancing Large Language Models Fine-Tuning with LoRA-Based Mixture of Experts," arXiv preprint arXiv:2404.15159, 2024.
- [8] W. Fedus, B. Zoph, and N. Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models

with Simple and Efficient Sparsity," Journal of Machine Learning Research, vol. 23, no. 120, pp. 1–39, 2022.

- [9] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," in Proceedings of the International Conference on Learning Representations (ICLR), 2022.
- [10] N. Shazeer et al., "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," in Proceedings of the International Conference on Learning Representations (ICLR), 2017.