Coordination Patterns for Energy-Aware Efficiency in LLM-Based Multi-Agent Systems: A systematic Literature Review

Literature Review Protocol

Literature Research Rationale:

The rapid adoption of Large Language Models (LLMs) has led to the emergence of LLM-based multi-agent systems (MAS), where multiple autonomous agents collaborate to solve complex tasks. While these systems improve modularity, scalability, and reasoning capabilities, they also introduce significant computational and energy overhead due to repeated model invocations, inter-agent communication, and coordination complexity.

Recent research has proposed various architectural and coordination patterns—such as centralized orchestration, hierarchical control, decentralized peer-to-peer interaction, and role-based specialization—to compose multiple agents into larger systems. Although many of these approaches aim to improve efficiency, scalability, or cost-effectiveness, energy consumption is rarely addressed explicitly. Instead, studies often optimize energy-related proxy metrics, including token usage, number of LLM calls, latency, and computational overhead, which are known to correlate with energy consumption in LLM-based systems.

This literature review aims to systematically analyze existing research on architectural and coordination patterns in LLM-based multi-agent systems, with a specific focus on how these patterns impact energy-aware efficiency through proxy metrics. By synthesizing state-of-the-art approaches, this study seeks to identify dominant patterns, commonly used efficiency metrics, methodological limitations, and open research gaps. The findings will contribute to a clearer understanding of how agent composition strategies influence energy-related efficiency and will support the design of more sustainable LLM-based multi-agent systems.

Research Questions :

How have architectural and coordination patterns in LLM-based multi-agent systems been designed and evaluated with respect to energy-aware efficiency or energy-related proxy metrics?

Rationale:

While energy efficiency is an increasingly important concern in LLM-based systems, existing research often evaluates efficiency indirectly through proxy metrics such as token usage, latency, and system cost. This question aims to collect and analyze how different multi-agent architectures and coordination mechanisms influence these metrics and, by extension, energy-aware efficiency..

Sub-RQ1: What architectural and coordination patterns are commonly used to compose LLM-based multi-agent systems?

Rationale:
LLM-based multi-agent systems employ a wide range of architectural and coordination patterns, including centralized orchestrators, hierarchical manager–worker models, decentralized peer-to-peer interactions, and blackboard-based coordination. Identifying and classifying these patterns provides a foundation for understanding how agent composition affects system efficiency.

Sub-RQ2: What energy-related proxy metrics are used to evaluate efficiency in LLM-based multi-agent systems?

Rationale:
Due to the lack of direct energy measurements, most studies rely on proxy metrics such as token consumption, number of LLM invocations, execution latency, parallelism, and context window size. This question examines which metrics are used, how they are measured, and how they are justified as indicators of energy-aware efficiency.

Sub-RQ3: What challenges and limitations are reported in achieving energy-aware efficiency through multi-agent coordination patterns?

Rationale:
Despite proposed optimizations, challenges such as redundant reasoning, coordination overhead, lack of standardized evaluation metrics, and trade-offs with performance or solution quality remain. Understanding these limitations will help identify gaps and inform future research directions.

Keywords:

LLM-Based Multi-Agent Systems, Architectural Patterns, Coordination Patterns, Agent Composition, Energy-Aware Efficiency, Token Efficiency, Latency, Proxy Metrics, Sustainable AI, Large Language Models.

General search query:

("large language model" OR LLM) AND ("multi-agent" OR "multi-agent system" OR "agent-based" OR "agent-based system" OR "autonomous agents") AND ("coordination" OR "orchestration" OR "agent communication" OR "task delegation" OR architecture* OR pattern*) AND ("efficiency" OR "latency" OR "performance" OR "token usage" OR "per token" OR "cost" OR "energy" OR "power consumption" OR "carbon footprint" OR "sustainability")

To ensure comprehensive coverage, synonymous terms for multi-agent systems, coordination mechanisms, and efficiency-related metrics were included to account for variations in terminology across AI, software engineering, and systems research.

Databases to be used and specific search query:

Web of Science	("large language model" OR LLM) AND ("multi-agent" OR "multi-agent system" OR "agent-based" OR "agent-based system" OR "autonomous agents") AND ("coordination" OR "orchestration" OR "agent communication" OR "task delegation" OR architecture* OR pattern*) AND ("efficiency" OR "latency" OR "performance" OR "token usage" OR "per token" OR "cost" OR "energy" OR "power consumption" OR "carbon footprint" OR "sustainability")
IEEE	("large language model" OR LLM) AND ("multi-agent" OR "agent-based" OR "autonomous agents") AND ("coordination" OR "orchestration" OR "agent communication" OR "task delegation" OR architecture* OR pattern*) AND ("efficiency" OR "latency" OR "performance" OR "token usage" OR "per token" OR "cost" OR "energy" OR "power consumption" OR "carbon footprint" OR "sustainability")
ACM Digital Library	("large language model" OR LLM) AND ("multi-agent" OR "multi-agent system" OR "agent-based" OR "agent-based system" OR "autonomous agents") AND ("coordination" OR "orchestration" OR "agent communication" OR "task delegation") AND (architecture* OR pattern*) AND ("efficiency" OR "latency" OR "performance" OR "token usage" OR "per token" OR "cost" OR "energy" OR "power consumption" OR "carbon footprint" OR "sustainability")

Study selection:

Include	**Exclude**
Articles focusing on LLM-based multi-agent systems.	Studies unrelated to LLMs or multi-agent systems.
Studies discussing architectural or coordination patterns.	Papers focusing solely on single-agent LLM applications.
Articles evaluating energy efficiency or using proxy metrics (e.g., token usage, latency, cost)	Articles discussing sustainability without computational or system-level evaluation.
Articles published in peer-reviewed journals or conferences.	Abstract-only papers, posters, editorials, or tutorials

Peer-reviewed journal articles and conference papers	Non-peer-reviewed articles
Studies written in English	
Empirical, experimental, or design-oriented studies	

Data extraction:

General Data:

Author, Title, Year, Abstract, Venue, Database

Specific Data:

- Type of architectural pattern (e.g., centralized, hierarchical, decentralized)
- Coordination mechanism (e.g., orchestration, negotiation, blackboard)
- Number and roles of agents
- Energy-related proxy metrics used (e.g., token usage, latency, number of LLM calls)
- Measurement level (e.g., agent-level, system-level)
- Evaluation setting (e.g., simulation, benchmark, real-world case study)
- Trade-offs reported (e.g., efficiency vs. accuracy or quality)
- Challenges and limitations identified
- Proposed optimization strategies or design guidelines
- Validation method (e.g., empirical evaluation, comparative experiments)

Means of Data Extraction

A Google Sheet will be created to support structured data collection. The sheet will include predefined fields for general bibliographic information and detailed technical attributes related to architecture, coordination patterns, and energy-related proxy metrics. This approach ensures consistency, traceability, and reproducibility of the data extraction process.

Data Analysis and Synthesis

Main Research Question (RQ1)

- Identify and categorize architectural and coordination patterns
- Analyze how each pattern affects energy-related proxy metrics
- Compare efficiency outcomes across different agent compositions
- Synthesize findings into a structured taxonomy of energy-aware patterns

Sub-RQ1

- Classify architectural and coordination patterns
- Identify dominant design strategies in LLM-based MAS
- Analyze trends in agent composition and role specialization

Sub-RQ2

- Identify commonly used proxy metrics
- Compare quantitative evaluation approaches
- Assess how proxy metrics are justified as indicators of energy efficiency

Sub-RQ3

- Identify recurring challenges and limitations
- Analyze trade-offs between efficiency and other system qualities
- Highlight open research gaps and opportunities for future work