03 Oct 2024

Communication Embeddings and Router to Support LLM-based Multi-agent Systems

Communication Embeddings and Router to Support LLM-based Multi-agent Systems

ABSTRACT

Large language models (LLM) agents are capable of taking actions and performing useful tasks. Complex tasks need collaboration between multiple such agents to accomplish a goal. Existing multi-agent communication frameworks are deficient in supporting such collaboration since these frameworks fail to incorporate critical aspects that would enable effective interactions between agents. This disclosure describes an inter-agent communication framework to improve communication, decision making, and collaboration between LLM agents. In the framework, new embedding paradigms centered around communicative facets can be utilized. The embeddings can enhance communication efficacy and also serve as a foundation for training multi-agent systems to accomplish their objectives effectively. A communication router is utilized that has awareness of inter-agent communication structure and strategies. The router can append a communication embedding to incoming tokens from an LLM and determine outgoing actions for one or more other LLMs that are part of the task workflow.

KEYWORDS

- Large Language Model (LLM)
- LLM agent
- Inter-agent communication
- Collaboration efficiency
- Communication embedding
- Action selector
- Multi-agent system

BACKGROUND

Large language models (LLM) agents are capable of taking actions and performing useful tasks. Complex tasks need collaboration between multiple such agents to accomplish a goal. Existing multi-agent communication frameworks are deficient in supporting such collaboration since these frameworks fail to incorporate critical aspects that would enable effective interactions between agents. Some of the aspects include assessment of task efficiency; identification of potential biases or hallucinations (including by individual LLM agents); formulation of informed decisions; criticize and reflect; monitoring of progress; evaluation of communication effectiveness and productivity; and facilitation of learning from errors to attain shared objectives that may deviate from those that were pursued during individual LLM training for the different agents. Additionally, there are no available embeddings to delineate the aspects during inter-agent communication.

In basic structure, inter-agent communication can be regarded as one of the actions for the agents within the system. However, as the number of agents increases and tasks become more complex, conventional communication techniques can become inadequate to manage inter-agent communication. Communication between agents with an organizational structure during the communication process is essential to preserve efficiency (as the inter-agent communication framework does not interfere with distributed LLM deployment communication such as collectives and parameter DMAs) as well as to ensure success of multi-agent systems/ configurations.

DESCRIPTION

This disclosure describes an inter-agent communication framework to improve communication and collaboration between LLM agents. In the framework, new embedding paradigms centered around communicative facets can be utilized. The embeddings can enhance communication efficacy and also serve as a foundation for training multi-agent systems to accomplish their objectives effectively.

A new communication router is utilized that has awareness of inter-agent communication structure and strategies. The router has prompting capabilities. Additionally, multi-agent systems that employ several LLM agents towards achieving a complex task are designed with the capacity to establish interconnections between communication application programming interfaces (APIs) that facilitate the execution of various actions across multiple modalities. This can include communication with conventional central processing units (CPUs) to carry out tasks for file APIs, storage input/output (I/O) operations, sensor or motor control, streaming, data packet transmission, etc.

In the context of inter-agent communication, the communication router includes substantially distinct components when compared to conventional server routers. Fig. 1 illustrates a schematic representation of the architecture of an agent communication router.

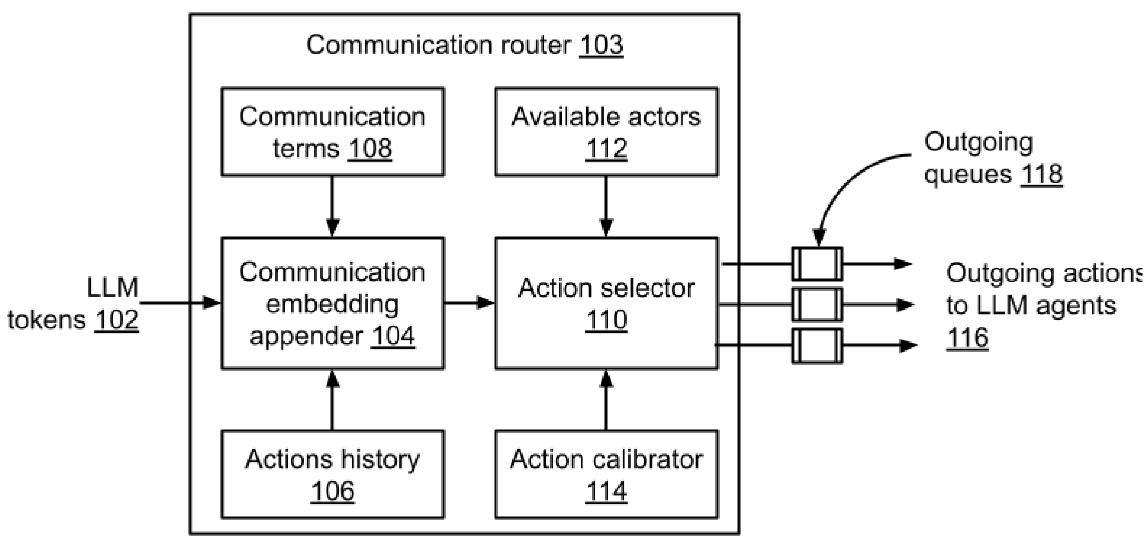

Fig. 1: Inter-agent communication router

In the example of Fig. 1, an inter-agent communication router (103) is illustrated that has capabilities to facilitate efficient communication between agents to achieve a goal. LLM tokens (102) are received at the communication router. A communication embedding appender (104) retrieves action history (106) and communication terms (108) to append communication embeddings to the received tokens. Based on the tokens with the appended communication embeddings, an action selector (110) identifies next agents (for the task workflow) from available actors (112). The action selector also receives calibration input from an action calibrator (114). The outgoing actions (116) to different LLM agents can be stored in respective outgoing queues (118).

Various components of the router can be implemented utilizing deep neural networks (DNNs) (Hopfield network, Boltzmann network, attention networks, etc.) that can enable above-mentioned capabilities. The inter-agent communication has capabilities to interconnect communication APIs to facilitate execution of various actions by LLM agents, across multiple modalities. Such actions can include, for example, defining priorities, delegating tasks, decision making, verification, querying, critical thinking, getting feedback from other agents, reviews, control for scheduling, asking for feedback from human interface, moving task checkpoints and maintaining internal states for execution, etc.

The communication router (inter-agent communication fabric) can facilitate communication between agents. The agents can be deployed on suitable processors, e.g., machine-learning processing units (including CPU, GPU, machine-learning processors, etc.) that are connected via OCI (On-Chip Interconnect) and/or Interchip Interconnect (ICI).

CONCLUSION

This disclosure describes an inter-agent communication framework to improve communication, decision making, and collaboration between LLM agents. In the framework, new embedding paradigms centered around communicative facets can be utilized. The embeddings can enhance communication efficacy and also serve as a foundation for training multi-agent systems to accomplish their objectives effectively. A communication router is utilized that has awareness of inter-agent communication structure and strategies. The router can append a communication embedding to incoming tokens from an LLM and determine outgoing actions for one or more other LLMs that are part of the task workflow.