US 20250259008A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2025/0259008 A1**

Wen et al. (43) **Pub. Date: Aug. 14, 2025**

(54) **SERVERLESS FUNCTIONAL ROUTING FOR LARGE LANGUAGE MODEL INFERENCE SERVICE**

(71) Applicant: **International Business Machines Corporation**, Armonk, NY (US)

(72) Inventors: **Bo Wen**, Chappaqua, NY (US); **Chen Wang**, Chappaqua, NY (US); **Huamin Chen**, Newton, MA (US)

(21) Appl. No.: **18/436,105**

(22) Filed: **Feb. 8, 2024**

Publication Classification

(51) **Int. Cl.**

G06F 40/40 (2020.01)

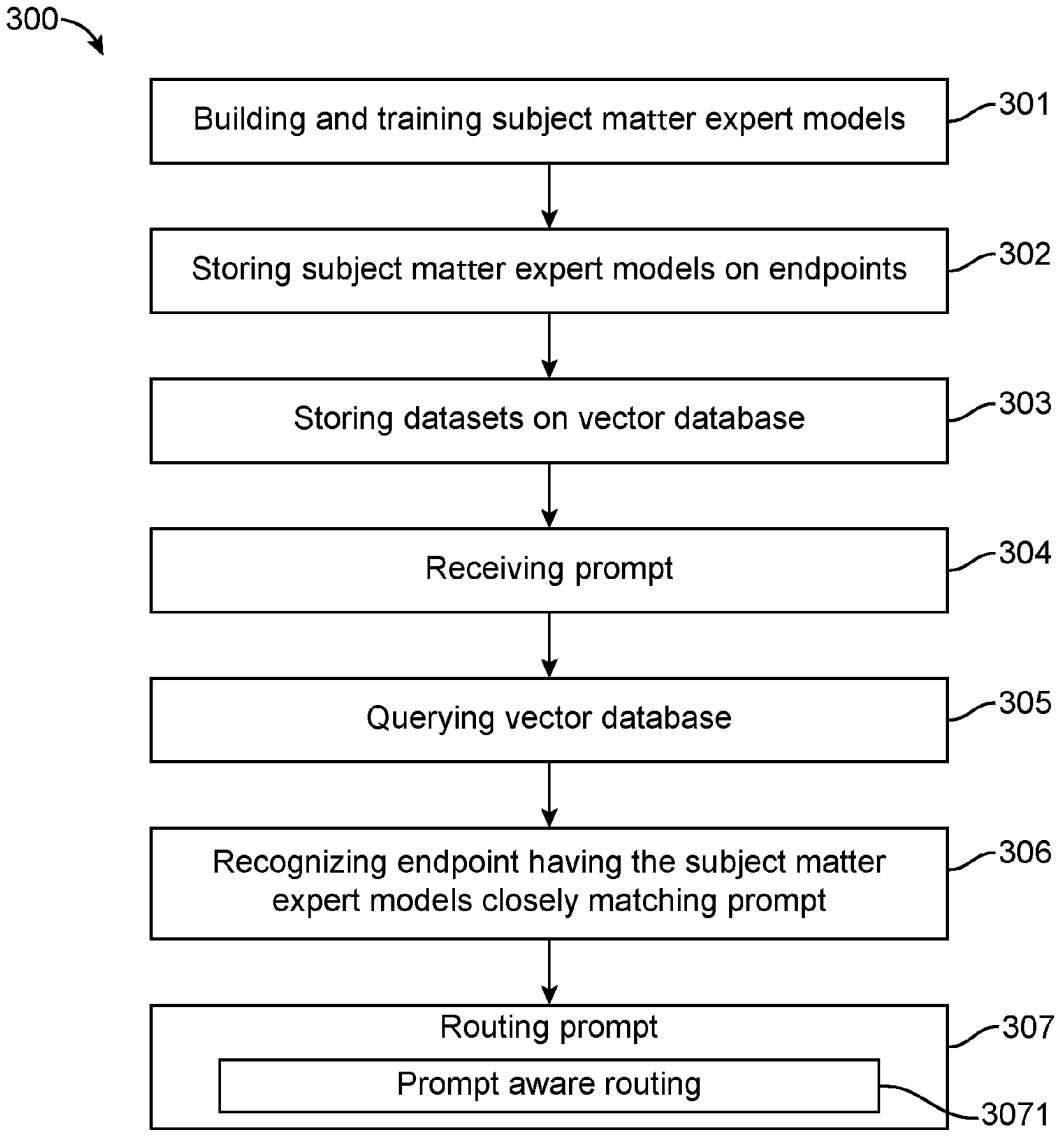
G06F 16/33 (2025.01)

(52) **U.S. Cl.**

CPC ***G06F 40/40*** (2020.01); ***G06F 16/3347*** (2019.01)

(57) **ABSTRACT**

A computer-implemented method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon is provided. The computer-implemented method includes receiving a prompt, querying a database comprising multiple datasets for an indication as to which one of the multiple datasets has a highest level of similarity with the prompt, recognizing one of the multiple endpoints as having the set of the expert models stored thereon which have a closest match with the one of the multiple datasets and routing the prompt to the one of the multiple endpoints having the set of the expert models stored thereon which have the closest match with the one of the multiple datasets.



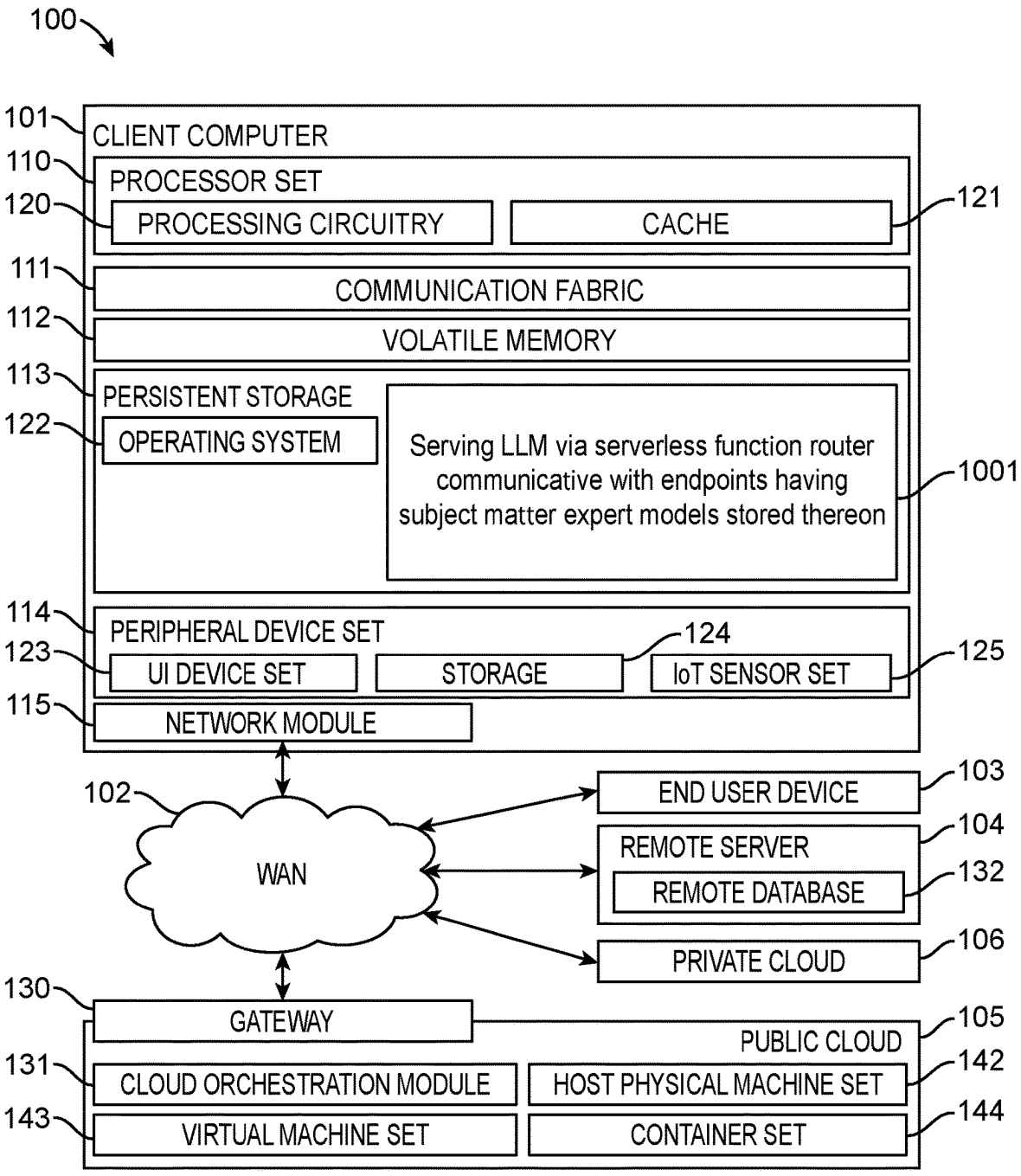

100
101
CLIENT COMPUTER
110
PROCESSOR SET
120
PROCESSING CIRCUITRY
CACHE
121
111
COMMUNICATION FABRIC
112
VOLATILE MEMORY
113
PERSISTENT STORAGE
122
OPERATING SYSTEM
Serving LLM via serverless function router communicative with endpoints having subject matter expert models stored thereon
1001
114
PERIPHERAL DEVICE SET
123
UI DEVICE SET
STORAGE
124
IoT SENSOR SET
125
115
NETWORK MODULE
102
WAN
END USER DEVICE
103
REMOTE SERVER
104
REMOTE DATABASE
132
PRIVATE CLOUD
106
130
GATEWAY
PUBLIC CLOUD
105
131
CLOUD ORCHESTRATION MODULE
HOST PHYSICAL MACHINE SET
142
143
VIRTUAL MACHINE SET
CONTAINER SET
144

FIG. 1

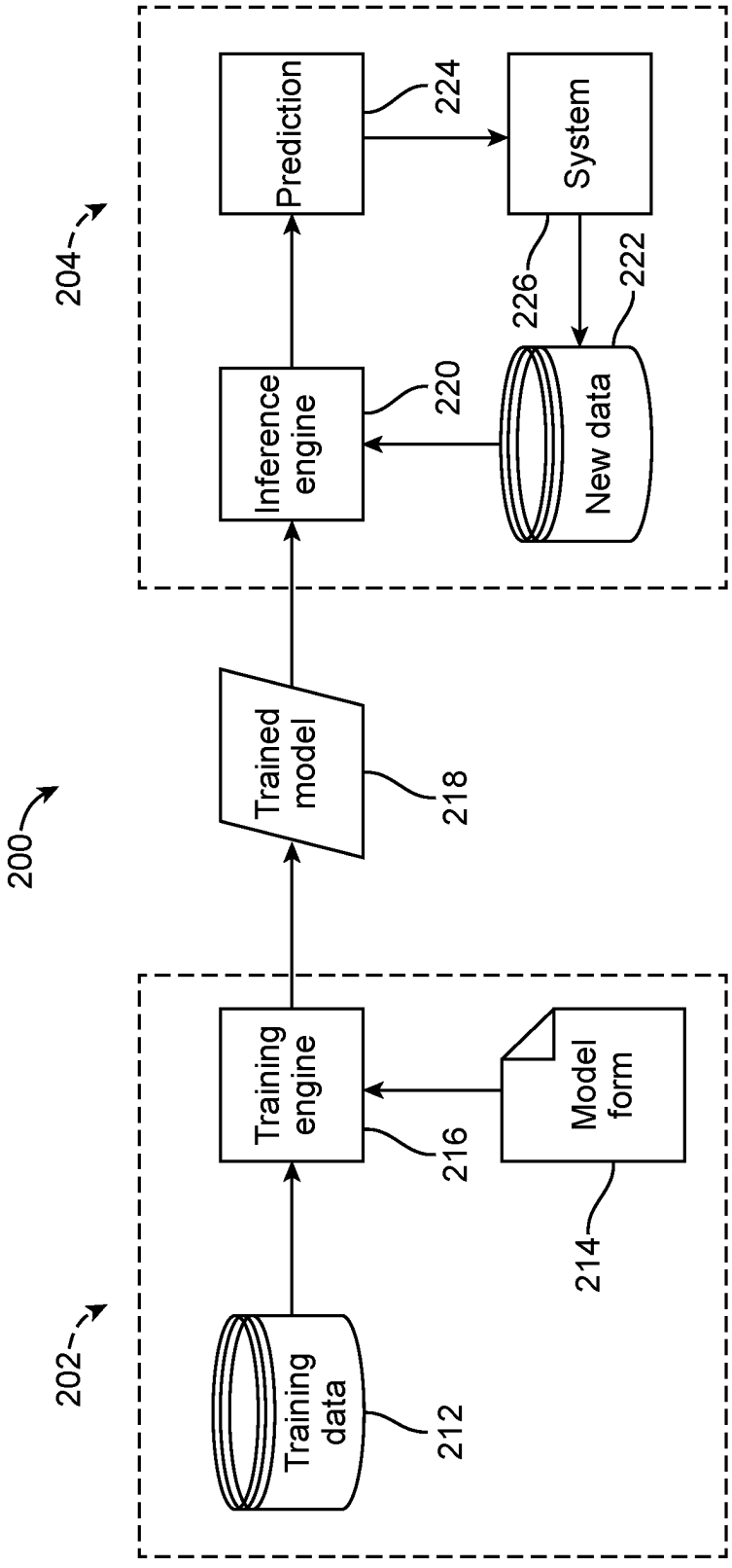

200
202
Training data
212
Training engine
216
Model form
214
Trained model
218
204
Inference engine
220
New data
222
226
Prediction
224
System

FIG. 2

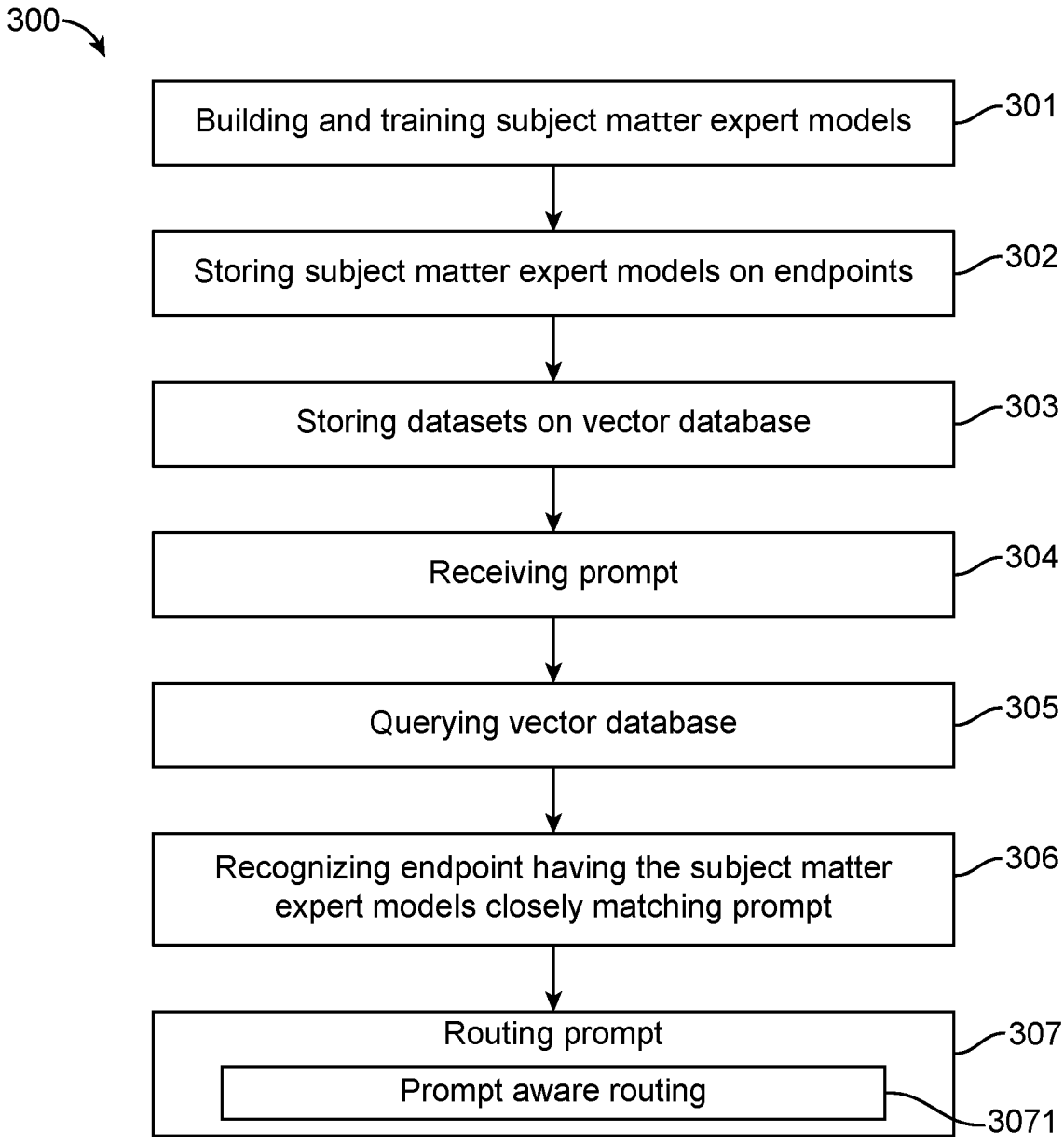

300
Building and training subject matter expert models
301
Storing subject matter expert models on endpoints
302
Storing datasets on vector database
303
Receiving prompt
304
Querying vector database
305
Recognizing endpoint having the subject matter expert models closely matching prompt
306
Routing prompt
307
Prompt aware routing
3071

FIG. 3

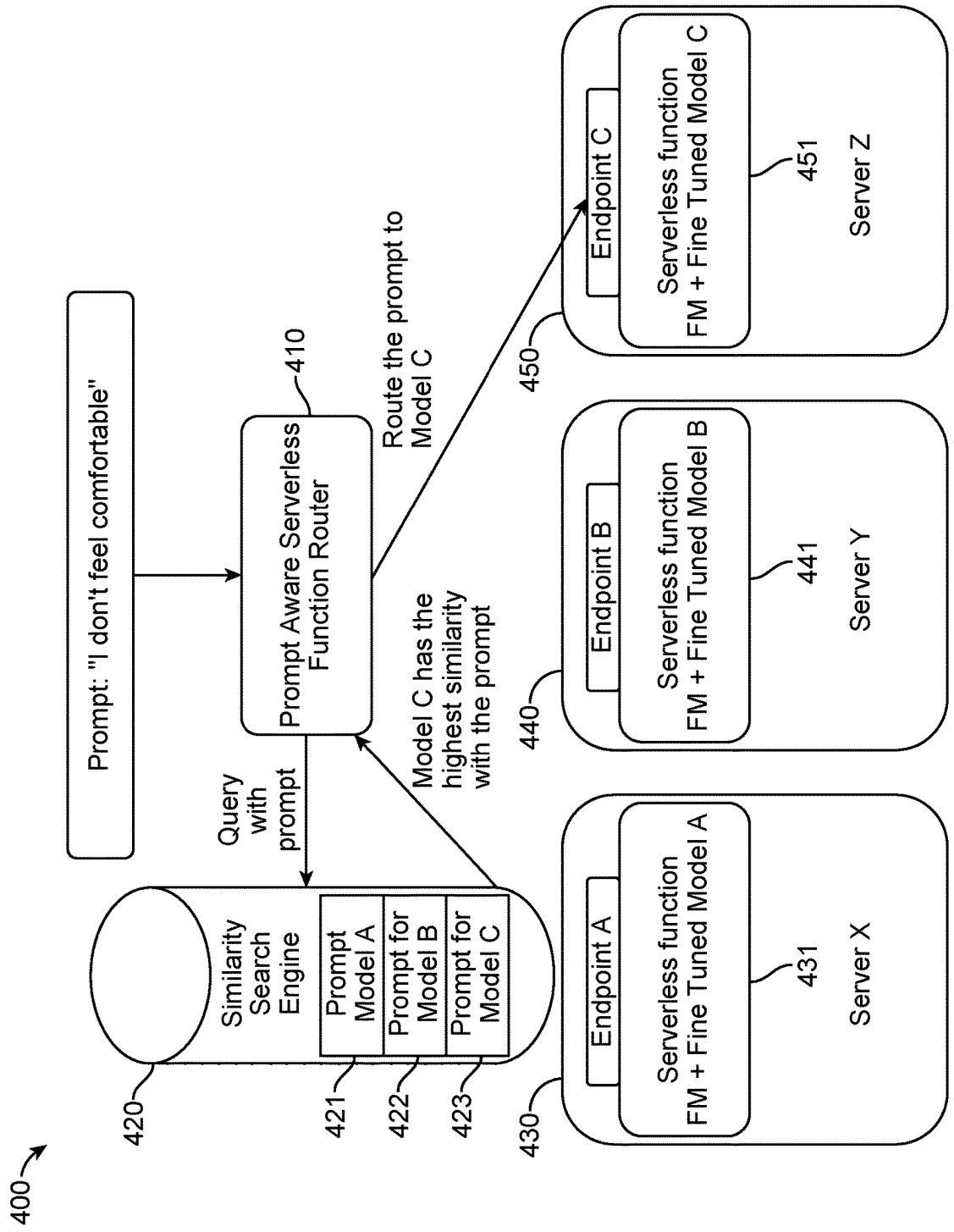

400
Prompt: "I don't feel comfortable"
Prompt Aware Serverless Function Router
410
Query with prompt
Route the prompt to Model C
Model C has the highest similarity with the prompt
Similarity Search Engine
Prompt Model A
Prompt for Model B
Prompt for Model C
420
421
422
423
Endpoint A
Serverless function FM + Fine Tuned Model A
431
Server X
430
Endpoint B
Serverless function FM + Fine Tuned Model B
441
Server Y
440
Endpoint C
Serverless function FM + Fine Tuned Model C
451
Server Z
450

FIG. 4

Model	Prompt	Embedding
A	Is fever a flu symptom?	[0.7, 0.2, 0.1, 0.4, 0.5]
B	How to define a function in Python?	[0.1, 0.6, 0.8, 0.3, 0.2]
C	What is P/E ratio?	[0.3, 0.4, 0.2, 0.7, 0.8]

FIG. 5

$$D = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

Where

- A_i is the i^{th} element of vector A.
- B_i is the i^{th} element of vector B.

FIG. 6

	Is fever a flu symptom?	How to define a function in Python?	What is P/E ratio?
I don't feel comfortable	0.387	0.678	0.583

FIG. 7

SERVERLESS FUNCTIONAL ROUTING FOR LARGE LANGUAGE MODEL INFERENCE SERVICE

BACKGROUND

[0001] The present invention generally relates to large language models in computing systems. More specifically, the present invention relates to cost-effective and quality-assured serverless functional routing for a large language model (LLM) inference service.

[0002] An LLM is a language model that is notable for its ability to achieve general-purpose language generation. LLMs acquire these abilities by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training processes. While LLMs are generally artificial neural networks that are can be built with transformer-based architectures, recent implementations have been based on alternative architectures, such as recurrent neural network variants.

[0003] As an example, LLMs can be used for text generation and other forms of generative artificial intelligence (AI). In these or other cases, LLMs take input text and repeatedly predict next tokens or words. Until recently, fine tuning was the only way a given LLM could be adapted to be able to accomplish specific tasks. It has been found, however, that modern large LLMs can be prompt-engineered to achieve positive results by acquiring knowledge about syntax, semantics and ontology inherent in human language.

SUMMARY

[0004] According to an aspect of the disclosure, a computer-implemented method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon is provided. The computer-implemented method includes receiving a prompt, querying a database comprising multiple datasets for an indication as to which one of the multiple datasets has a highest level of similarity with the prompt, recognizing one of the multiple endpoints as having the set of the expert models stored thereon which have a closest match with the one of the multiple datasets and routing the prompt to the one of the multiple endpoints having the set of the expert models stored thereon which have the closest match with the one of the multiple datasets. In additional or alternative embodiments, the computer-implemented method provides for a fast response to a prompt that does not waste valuable computing resources.

[0005] According to an aspect of the disclosure, a computer program product for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon is provided. The computer program product includes one or more computer readable storage media having computer readable program code collectively stored on the one or more computer readable storage media. The computer readable program code is executed by a processor of a computer system to cause the computer system to perform a method. The method includes receiving a prompt, querying a database comprising multiple datasets for an indication as to which one of the multiple datasets has a highest level of similarity with the prompt, recognizing one of the multiple endpoints as having the set of the expert models stored thereon which have a closest match with the one of the multiple datasets and routing the prompt to the one of the multiple endpoints having the set of the expert models stored thereon which have the closest match with the one of the multiple datasets. In additional or alternative embodiments, the method provides for a fast response to a prompt that does not waste valuable computing resources.

[0006] According to an aspect of the disclosure, a computing system is provided and includes a processor, a memory coupled to the processor and one or more computer readable storage media coupled to the processor. The one or more computer readable storage media collectively contain instructions that are executed by the processor via the memory to implement a method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon. In additional or alternative embodiments, the method for serving the LLM application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon provides for a fast response to a prompt that does not waste valuable computing resources.

[0007] Additional technical features and benefits are realized through the techniques of the present invention. Embodiments and aspects of the invention are described in detail herein and are considered a part of the claimed subject matter. For a better understanding, refer to the detailed description and to the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] The specifics of the exclusive rights described herein are particularly pointed out and distinctly claimed in the claims at the conclusion of the specification. The foregoing and other features and advantages of the embodiments of the invention are apparent from the following detailed description taken in conjunction with the accompanying drawings in which:

[0009] FIG. **1** is a schematic diagram of a computing environment for executing a computer-implemented method for operating a chip handling assembly in accordance with one or more embodiments;

[0010] FIG. **2** is a block diagram of components of a machine learning training and inference system in accordance with one or more embodiments;

[0011] FIG. **3** is a flow diagram illustrating a computer-implemented method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon in accordance with one or more embodiments;

[0012] FIG. **4** is a schematic diagram of an architecture for an execution of the computer-implemented method of FIG. **3** in accordance with one or more embodiments;

[0013] FIG. **5** is a table illustrating models, hypothetical prompts for the models and embedding data in accordance with one or more embodiments;

[0014] FIG. **6** is an equation illustrating Euclidean distance between prompts and models in accordance with one or more embodiments; and

[0015] FIG. **7** is a table illustrating models and distance data between a prompt and each of the models in accordance with one or more embodiments.

[0016] The diagrams depicted herein are illustrative. There can be many variations to the diagram or the operations described therein without departing from the spirit of the invention. For instance, the actions can be performed in a differing order or actions can be added, deleted or modified. Also, the term "coupled" and variations thereof describes having a communications path between two elements and does not imply a direct connection between the elements with no intervening elements/connections between them. All of these variations are considered a part of the specification.

[0017] In the accompanying figures and following detailed description of the described embodiments, the various elements illustrated in the figures are provided with two- or three-digit reference numbers. With minor exceptions, the leftmost digit(s) of each reference number correspond to the figure in which its element is first illustrated.

DETAILED DESCRIPTION

[0018] According to an aspect of the disclosure, a computer-implemented method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon is provided. The computer-implemented method includes receiving a prompt, querying a database comprising multiple datasets for an indication as to which one of the multiple datasets has a highest level of similarity with the prompt, recognizing one of the multiple endpoints as having the set of the expert models stored thereon which have a closest match with the one of the multiple datasets and routing the prompt to the one of the multiple endpoints having the set of the expert models stored thereon which have the closest match with the one of the multiple datasets. In additional or alternative embodiments, the computer-implemented method provides for a fast response to a prompt that does not waste valuable computing resources.

[0019] The receiving of the prompt, the querying of the database, the recognizing of the one of the multiple endpoints and the routing of the prompt are executed by the serverless function router so that all of the operations are contained within a single operational and computational entity.

[0020] The database includes a vector database, which is configured to facilitate or enable the direction of queries to certain closest-matching endpoints.

[0021] Each of the multiple datasets has a closest match with the subject matter expert models stored on one of the multiple endpoints. This way, the computer-implemented method avoids sending a query to a less closely matching endpoint.

[0022] The subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to medical subject matter, the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to financial subject matter and the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to technical subject matter. These particular subject matter experts are just examples, but they cover some of the most valuable subject matter.

[0023] The subject matter expert models include one or more foundation models and one or more fine-tuned models so that each subject matter expert can approach a query in a different manner.

[0024] The serverless function router is a prompt aware router and the routing includes prompt aware routing that is responsive to user input.

[0025] According to an aspect of the disclosure, a computer program product for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon is provided. The computer program product includes one or more computer readable storage media having computer readable program code collectively stored on the one or more computer readable storage media. The computer readable program code is executed by a processor of a computer system to cause the computer system to perform a method. The method includes receiving a prompt, querying a database comprising multiple datasets for an indication as to which one of the multiple datasets has a highest level of similarity with the prompt, recognizing one of the multiple endpoints as having the set of the expert models stored thereon which have a closest match with the one of the multiple datasets and routing the prompt to the one of the multiple endpoints having the set of the expert models stored thereon which have the closest match with the one of the multiple datasets. In additional or alternative embodiments, the method provides for a fast response to a prompt that does not waste valuable computing resources.

[0026] The receiving of the prompt, the querying of the database, the recognizing of the one of the multiple endpoints and the routing of the prompt are executed by the serverless function router so that all of the operations are contained within a single operational and computational entity.

[0027] The database includes a vector database, which is configured to facilitate or enable the direction of queries to certain closest-matching endpoints.

[0028] Each of the multiple datasets has a closest match with the subject matter expert models stored on one of the multiple endpoints. This way, the method avoids sending a query to a less closely matching endpoint.

[0029] The subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to medical subject matter, the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to financial subject matter and the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to technical subject matter. These particular subject matter experts are just examples, but they cover some of the most valuable subject matter.

[0030] The subject matter expert models include one or more foundation models and one or more fine-tuned models so that each subject matter expert can approach a query in a different manner.

[0031] The serverless function router is a prompt aware router and the routing includes prompt aware routing that is responsive to user input.

[0032] According to an aspect of the disclosure, a computing system is provided and includes a processor, a memory coupled to the processor and one or more computer readable storage media coupled to the processor. The one or

more computer readable storage media collectively contain instructions that are executed by the processor via the memory to implement a method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon. In additional or alternative embodiments, the method for serving the LLM application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon provides for a fast response to a prompt that does not waste valuable computing resources.

[0033] The receiving of the prompt, the querying of the database, the recognizing of the one of the multiple endpoints and the routing of the prompt are executed by the serverless function router so that all of the operations are contained within a single operational and computational entity.

[0034] The database includes a vector database, which is configured to facilitate or enable the direction of queries to certain closest-matching endpoints.

[0035] Each of the multiple datasets has a closest match with the subject matter expert models stored on one of the multiple endpoints. This way, the method for serving the LLM application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon avoids sending a query to a less closely matching endpoint.

[0036] The subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to medical subject matter, the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to financial subject matter and the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to technical subject matter. These particular subject matter experts are just examples, but they cover some of the most valuable subject matter.

[0037] The subject matter expert models include one or more foundation models and one or more fine-tuned models so that each subject matter expert can approach a query in a different manner.

[0038] The serverless function router is a prompt aware router and the routing includes prompt aware routing that is responsive to user input.

[0039] Various aspects of the present disclosure are described by narrative text, flowcharts, block diagrams of computer systems and/or block diagrams of the machine logic included in computer program product (CPP) embodiments. With respect to any flowcharts, depending upon the technology involved, the operations can be performed in a different order than what is shown in a given flowchart. For example, again depending upon the technology involved, two operations shown in successive flowchart blocks may be performed in reverse order, as a single integrated step, concurrently, or in a manner at least partially overlapping in time.

[0040] A computer program product embodiment ("CPP embodiment" or "CPP") is a term used in the present disclosure to describe any set of one, or more, storage media (also called "mediums") collectively included in a set of one, or more, storage devices that collectively include machine readable code corresponding to instructions and/or data for performing computer operations specified in a given CPP claim. A "storage device" is any tangible device that can retain and store instructions for use by a computer processor. Without limitation, the computer readable storage medium may be an electronic storage medium, a magnetic storage medium, an optical storage medium, an electromagnetic storage medium, a semiconductor storage medium, a mechanical storage medium, or any suitable combination of the foregoing. Some known types of storage devices that include these mediums include: diskette, hard disk, random access memory (RAM), read-only memory (ROM), erasable programmable read-only memory (EPROM or Flash memory), static random access memory (SRAM), compact disc read-only memory (CD-ROM), digital versatile disk (DVD), memory stick, floppy disk, mechanically encoded device (such as punch cards or pits/lands formed in a major surface of a disc) or any suitable combination of the foregoing. A computer readable storage medium, as that term is used in the present disclosure, is not to be construed as storage in the form of transitory signals per se, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide, light pulses passing through a fiber optic cable, electrical signals communicated through a wire, and/or other transmission media. As will be understood by those of skill in the art, data is typically moved at some occasional points in time during normal operations of a storage device, such as during access, de-fragmentation or garbage collection, but this does not render the storage device as transitory because the data is not transitory while it is stored.

[0041] With reference to FIG. **1**, a computer or computing device **100** that implements a computer-implemented method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon. The computer or computing device **100** of FIG. **1** contains an example of an environment for the execution of at least some of the computer code involved in performing the inventive methods, such as the block **1001** of the computer-implemented method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon.

[0042] In addition to the computer-implemented method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon of block **1001**, the computer or computing device **100** includes, for example, computer **101**, wide area network (WAN) **102**, end user device (EUD) **103**, remote server **104**, public cloud **105**, and private cloud **106**. In this embodiment, computer **101** includes processor set **110** (including processing circuitry **120** and cache **121**), communication fabric **111**, volatile memory **112**, persistent storage **113** (including operating system **122** and the computer-implemented method of block **1001**, as identified above), peripheral device set **114** (including user interface (UI) device set **123**, storage **124**, and Internet of Things (IoT) sensor set **125**), and network module **115**. Remote server **104** includes remote database **130**. Public cloud **105** includes gateway **140**, cloud orchestration module **141**, host physical machine set **142**, virtual machine set **143**, and container set **144**.

[0043] The computer **101** may take the form of a desktop computer, laptop computer, tablet computer, smart phone,

smart watch or other wearable computer, mainframe computer, quantum computer or any other form of computer or mobile device now known or to be developed in the future that is capable of running a program, accessing a network or querying a database, such as remote database **130**. As is well understood in the art of computer technology, and depending upon the technology, performance of a computer-implemented method may be distributed among multiple computers and/or between multiple locations. On the other hand, in this presentation of the computer-implemented method, detailed discussion is focused on a single computer, specifically computer **101**, to keep the presentation as simple as possible. Computer **101** may be located in a cloud, even though it is not shown in a cloud in FIG. **1**. On the other hand, computer **101** is not required to be in a cloud except to any extent as may be affirmatively indicated.

[0044] The processor set **110** includes one, or more, computer processors of any type now known or to be developed in the future. Processing circuitry **120** may be distributed over multiple packages, for example, multiple, coordinated integrated circuit chips. Processing circuitry **120** may implement multiple processor threads and/or multiple processor cores. Cache **121** is memory that is located in the processor chip package(s) and is typically used for data or code that should be available for rapid access by the threads or cores running on processor set **110**. Cache memories are typically organized into multiple levels depending upon relative proximity to the processing circuitry. Alternatively, some, or all, of the cache for the processor set may be located "off chip." In some computing environments, processor set **110** may be designed for working with qubits and performing quantum computing.

[0045] Computer readable program instructions are typically loaded onto computer **101** to cause a series of operational steps to be performed by processor set **110** of computer **101** and thereby effect a computer-implemented method, such that the instructions thus executed will instantiate the methods specified in flowcharts and/or narrative descriptions of computer-implemented methods included in this document (collectively referred to as "the inventive methods"). These computer readable program instructions are stored in various types of computer readable storage media, such as cache **121** and the other storage media discussed below. The program instructions, and associated data, are accessed by processor set **110** to control and direct performance of the inventive methods. In the computer-implemented method, at least some of the instructions for performing the inventive methods may be stored in the block **1001** of the computer-implemented method in persistent storage **113**.

[0046] Communication fabric **111** is the signal conduction path that allows the various components of computer **101** to communicate with each other. Typically, this fabric is made of switches and electrically conductive paths, such as the switches and electrically conductive paths that make up busses, bridges, physical input/output ports and the like. Other types of signal communication paths may be used, such as fiber optic communication paths and/or wireless communication paths.

[0047] Volatile memory **112** is any type of volatile memory now known or to be developed in the future. Examples include dynamic type random access memory (RAM) or static type RAM. Typically, volatile memory **112** is characterized by random access, but this is not required unless affirmatively indicated. In computer **101**, the volatile memory **112** is located in a single package and is internal to computer **101**, but, alternatively or additionally, the volatile memory may be distributed over multiple packages and/or located externally with respect to computer **101**.

[0048] Persistent storage **113** is any form of non-volatile storage for computers that is now known or to be developed in the future. The non-volatility of this storage means that the stored data is maintained regardless of whether power is being supplied to computer **101** and/or directly to persistent storage **113**. Persistent storage **113** may be a read only memory (ROM), but typically at least a portion of the persistent storage allows writing of data, deletion of data and re-writing of data. Some familiar forms of persistent storage include magnetic disks and solid state storage devices. Operating system **122** may take several forms, such as various known proprietary operating systems or open source Portable Operating System Interface-type operating systems that employ a kernel. The code included in the block **1001** of the computer-implemented method typically includes at least some of the computer code involved in performing the inventive methods.

[0049] Peripheral device set **114** includes the set of peripheral devices of computer **101**. Data communication connections between the peripheral devices and the other components of computer **101** may be implemented in various ways, such as Bluetooth connections, Near-Field Communication (NFC) connections, connections made by cables (such as universal serial bus (USB) type cables), insertion-type connections (for example, secure digital (SD) card), connections made through local area communication networks and even connections made through wide area networks such as the internet. In various embodiments, UI device set **123** may include components such as a display screen, speaker, microphone, wearable devices (such as goggles and smart watches), keyboard, mouse, printer, touchpad, game controllers, and haptic devices. Storage **124** is external storage, such as an external hard drive, or insertable storage, such as an SD card. Storage **124** may be persistent and/or volatile. In some embodiments, storage **124** may take the form of a quantum computing storage device for storing data in the form of qubits. In embodiments where computer **101** is required to have a large amount of storage (for example, where computer **101** locally stores and manages a large database) then this storage may be provided by peripheral storage devices designed for storing very large amounts of data, such as a storage area network (SAN) that is shared by multiple, geographically distributed computers. IoT sensor set **125** is made up of sensors that can be used in Internet of Things applications. For example, one sensor may be a thermometer and another sensor may be a motion detector.

[0050] Network module **115** is the collection of computer software, hardware, and firmware that allows computer **701** to communicate with other computers through WAN **102**. Network module **115** may include hardware, such as modems or Wi-Fi signal transceivers, software for packetizing and/or de-packetizing data for communication network transmission, and/or web browser software for communicating data over the internet. In some embodiments, network control functions and network forwarding functions of network module **115** are performed on the same physical hardware device. In other embodiments (for example, embodiments that utilize software-defined networking (SDN)), the control functions and the forwarding functions

of network module **115** are performed on physically separate devices, such that the control functions manage several different network hardware devices. Computer readable program instructions for performing the inventive methods can typically be downloaded to computer **101** from an external computer or external storage device through a network adapter card or network interface included in network module **115**.

[0051] WAN **102** is any wide area network (for example, the internet) capable of communicating computer data over non-local distances by any technology for communicating computer data, now known or to be developed in the future. In some embodiments, the WAN **102** may be replaced and/or supplemented by local area networks (LANs) designed to communicate data between devices located in a local area, such as a Wi-Fi network. The WAN and/or LANs typically include computer hardware such as copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and edge servers.

[0052] End user device (EUD) **103** is any computer system that is used and controlled by an end user (for example, a customer of an enterprise that operates computer **101**), and may take any of the forms discussed above in connection with computer **101**. EUD **103** typically receives helpful and useful data from the operations of computer **101**. For example, in a hypothetical case where computer **101** is designed to provide a recommendation to an end user, this recommendation would typically be communicated from network module **115** of computer **101** through WAN **102** to EUD **103**. In this way, EUD **103** can display, or otherwise present, the recommendation to an end user. In some embodiments, EUD **103** may be a client device, such as thin client, heavy client, mainframe computer, desktop computer and so on.

[0053] Remote server **104** is any computer system that serves at least some data and/or functionality to computer **101**. Remote server **104** may be controlled and used by the same entity that operates computer **101**. Remote server **104** represents the machine(s) that collect and store helpful and useful data for use by other computers, such as computer **101**. For example, in a hypothetical case where computer **101** is designed and programmed to provide a recommendation based on historical data, then this historical data may be provided to computer **101** from remote database **130** of remote server **104**.

[0054] Public cloud **105** is any computer system available for use by multiple entities that provides on-demand availability of computer system resources and/or other computer capabilities, especially data storage (cloud storage) and computing power, without direct active management by the user. Cloud computing typically leverages sharing of resources to achieve coherence and economies of scale. The direct and active management of the computing resources of public cloud **105** is performed by the computer hardware and/or software of cloud orchestration module **141**. The computing resources provided by public cloud **105** are typically implemented by virtual computing environments that run on various computers making up the computers of host physical machine set **142**, which is the universe of physical computers in and/or available to public cloud **105**. The virtual computing environments (VCEs) typically take the form of virtual machines from virtual machine set **143** and/or containers from container set **144**. It is understood that these VCEs may be stored as images and may be transferred among and between the various physical machine hosts, either as images or after instantiation of the VCE. Cloud orchestration module **141** manages the transfer and storage of images, deploys new instantiations of VCEs and manages active instantiations of VCE deployments. Gateway **140** is the collection of computer software, hardware, and firmware that allows public cloud **105** to communicate through WAN **102**.

[0055] Some further explanation of virtualized computing environments (VCEs) will now be provided. VCEs can be stored as "images." A new active instance of the VCE can be instantiated from the image. Two familiar types of VCEs are virtual machines and containers. A container is a VCE that uses operating-system-level virtualization. This refers to an operating system feature in which the kernel allows the existence of multiple isolated user-space instances, called containers. These isolated user-space instances typically behave as real computers from the point of view of programs running in them. A computer program running on an ordinary operating system can utilize all resources of that computer, such as connected devices, files and folders, network shares, CPU power, and quantifiable hardware capabilities. However, programs running inside a container can only use the contents of the container and devices assigned to the container, a feature which is known as containerization.

[0056] Private cloud **106** is similar to public cloud **105**, except that the computing resources are only available for use by a single enterprise. While private cloud **106** is depicted as being in communication with WAN **102**, in other embodiments a private cloud may be disconnected from the internet entirely and only accessible through a local/private network. A hybrid cloud is a composition of multiple clouds of different types (for example, private, community or public cloud types), often respectively implemented by different vendors. Each of the multiple clouds remains a separate and discrete entity, but the larger hybrid cloud architecture is bound together by standardized or proprietary technology that enables orchestration, management, and/or data/application portability between the multiple constituent clouds. In this embodiment, public cloud **105** and private cloud **106** are both part of a larger hybrid cloud.

[0057] Turning now to an overview of technologies that are more specifically relevant to aspects of the invention, LLMs are becoming increasingly common in AI applications. Many issues remain, however, with proper LLM operation and functionality. These issues include the fact that LLM training can be very expensive and model sizes can be very large (i.e., in the magnitude of several GBs) such that loading, reloading and starting an LLM in a serverless functions architecture can be very time consuming. The issues further include problems with routing. While, an LLM inference service may have prompts with different domain contexts, such as medical, financial, technical, etc., to improve routing effectiveness whereby a given query is routed to an appropriate serverless function that has the right model, routing end user prompts to appropriate LLM serverless functions has been shown to be different from traditional L3/L5/L7 routing.

[0058] Turning now to an overview of the aspects of the invention, one or more embodiments of the invention address the above-described shortcomings of the prior art by providing for a mixture of models (MoM) system in which

queries are routed to "expert" models. LLM serverless functions are constructed with foundation models (FMs) and fine tuned models, such as those trained via low rank (LoRA) or other training processes. The fine tuned models are complementary to the FM, yet very lightweight to load and start in serverless functions. Each serverless function service endpoint is annotated with the dataset or the fine tuned model information. Prompt aware routing instead of L3/L5/L7 routing is used to route prompts to the right serverless functions that have the best FM and fine tuned models.

[0059] Notably, a prompt-to-model routing method is provided in which a router queries a vector where a subset of fine tuning training datasets are located. The vector returns a similarity result telling how close the query is to a certain dataset. The router uses the similarity result to route the prompt to the right serverless function service endpoint. The computer-implemented method includes receiving a prompt, querying a database including multiple datasets for an indication as to which one of the multiple datasets has a highest level of similarity with the prompt, recognizing one of the multiple endpoints as having the set of the expert models stored thereon which have a closest match with the one of the multiple datasets and routing the prompt to the one of the multiple endpoints having the set of the expert models stored thereon which have the closest match with the one of the multiple datasets.

[0060] The above-described aspects of the invention address the shortcomings of the prior art by providing a computer-implemented method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon.

[0061] Turning now to a more detailed description of aspects of the present invention, FIG. **2** depicts a block diagram of components of a machine learning training and inference system **200**. The machine learning training and inference system **200**, in accordance with one or more embodiments of the invention, can utilize machine learning techniques to perform tasks, such as a computer-implemented method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon. Embodiments of the invention utilize AI, which includes a variety of so-called machine learning technologies. The phrase "machine learning" broadly describes a function of electronic systems that learn from data. A machine learning system, engine, or module can include a trainable machine learning algorithm that can be trained, such as in an external cloud environment, to learn functional relationships between inputs and outputs, and the resulting model (sometimes referred to as a "trained neural network," "trained model," and/or "trained machine learning model") can be used for managing information during a web conference, for example. In one or more embodiments of the invention, machine learning functionality can be implemented using an artificial neural network (ANN) having the capability to be trained to perform a function. In machine learning and cognitive science, ANNs are a family of statistical learning models inspired by the biological neural networks of animals, and in particular the brain. ANNs can be used to estimate or approximate systems and functions that depend on a large number of inputs. Convolutional neural networks (CNN) are a class of deep, feed-forward ANNs that are particularly useful at tasks such as, but not limited to analyzing visual imagery and natural language processing (NLP). Recurrent neural networks (RNN) are another class of deep, feed-forward ANNs and are particularly useful at tasks such as, but not limited to, unsegmented connected handwriting recognition and speech recognition. Other types of neural networks are also known and can be used in accordance with one or more embodiments of the invention described herein.

[0062] ANNs can be embodied as so-called "neuromorphic" systems of interconnected processor elements that act as simulated "neurons" and exchange "messages" between each other in the form of electronic signals. Similar to the so-called "plasticity" of synaptic neurotransmitter connections that carry messages between biological neurons, the connections in ANNs that carry electronic messages between simulated neurons are provided with numeric weights that correspond to the strength or weakness of a given connection. The weights can be adjusted and tuned based on experience, making ANNs adaptive to inputs and capable of learning. For example, an ANN for handwriting recognition is defined by a set of input neurons that can be activated by the pixels of an input image. After being weighted and transformed by a function determined by the network's designer, the activation of these input neurons are then passed to other downstream neurons, which are often referred to as "hidden" neurons. This process is repeated until an output neuron is activated. The activated output neuron determines which character was input. It should be appreciated that these same techniques can be applied in the case of localizing a target object referred by a compositional expression from an image set with similar visual elements as described herein.

[0063] The machine learning training and inference system **200** performs training **202** and inference **204**. During training **202**, a training engine **216** trains a model (e.g., the trained model **218**) to perform a task. Inference **204** is the process of implementing the trained model **218** to perform the task in the context of a larger system (e.g., a system **226**).

[0064] The training **202** begins with training data **212**, which can be structured or unstructured data. The training engine **216** receives the training data **212** and a model form **214**. The model form **214** represents a base model that is untrained. The model form **214** can have preset weights and biases, which can be adjusted during training. It should be appreciated that the model form **214** can be selected from many different model forms depending on the task to be performed. For example, where the training **202** is to train a model to perform image classification, the model form **214** can be a model form of a CNN (convolutional neural network). The training **202** can be supervised learning, semi-supervised learning, unsupervised learning, reinforcement learning, and/or the like, including combinations and/or multiples thereof. For example, supervised learning can be used to train a machine learning model to classify an object of interest in an image. To do this, the training data **212** includes labeled images, including images of the object of interest with associated labels (ground truth) and other images that do not include the object of interest with associated labels. In this example, the training engine **216** takes as input a training image from the training data **212**, makes a prediction for classifying the image, and compares the prediction to the known label. The training engine **216** then adjusts weights and/or biases of the model based on

results of the comparison, such as by using backpropagation. The training **202** can be performed multiple times (referred to as "epochs") until a suitable model is trained (e.g., the trained model **218**).

[0065] Once trained, the trained model **218** can be used to perform inference **204** to perform a task. The inference engine **220** applies the trained model **218** to new data **222** (e.g., real-world, non-training data). For example, if the trained model **218** is trained to classify images of a particular object, such as a chair, the new data **222** can be an image of a chair that was not part of the training data **212**. In this way, the new data **222** represents data to which the model **218** has not been exposed. The inference engine **220** makes a prediction **224** (e.g., a classification of an object in an image of the new data **222**) and passes the prediction **224** to the system **226**. The system **226** can, based on the prediction **224**, taken an action, perform an operation, perform an analysis, and/or the like, including combinations and/or multiples thereof. In some embodiments of the invention, the system **226** can add to and/or modify the new data **222** based on the prediction **224**.

[0066] In accordance with one or more embodiments of the invention, the predictions **224** generated by the inference engine **220** are periodically monitored and verified to ensure that the inference engine **220** is operating as expected. Based on the verification, additional training **202** can occur using the trained model **218** as the starting point. The additional training **202** can include all or a subset of the original training data **212** and/or new training data **212**. In accordance with one or more embodiments of the invention, the training **202** includes updating the trained model **218** to account for changes in expected input data.

[0067] With reference to FIG. **3**, a computer-implemented method **300** is provided for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon.

[0068] As shown in FIG. **3**, the computer-implemented method **300** initially includes building and training subject matter expert models to handle certain types of prompts from users (block **301**) and storing each subject matter expert models of a certain type on one of the multiple endpoints (block **302**). In addition, the computer-implemented method **300** also includes storing multiple datasets on a vector database (block **303**). Each of the multiple datasets stored on the vector database corresponds to and has a closest match with each subject matter expert model of the certain type stored on the one of the multiple endpoints.

[0069] Thus, in an exemplary case, one of the multiple endpoints can have subject matter expert models (i.e., foundation models and fine-tuned models) relating to medical prompts and a corresponding one of the multiple datasets stored on the vector database can include data and information corresponding to and having a closest match with the subject matter expert models relating to the medical prompts. In another exemplary case, one of the multiple endpoints can have subject matter expert models (i.e., foundation models and fine-tuned models) relating to financial prompts and a corresponding one of the multiple datasets stored on the vector database can include data and information corresponding to and having a closest match with the subject matter expert models relating to the financial prompts. In yet another exemplary case, one of the multiple endpoints can have subject matter expert models (i.e., foundation models and fine-tuned models) relating to technical prompts and a corresponding one of the multiple datasets stored on the vector database can include data and information corresponding to and having a closest match with the subject matter expert models relating to the technical prompts.

[0070] Of course it is to be understood that other types of subject matter expert models can be provided and that the types listed above are a non-exhaustive list.

[0071] The computer-implemented method **300** further includes receiving a prompt (block **304**), querying the vector database including the multiple datasets for an indication as to which one of the multiple datasets has a highest level of similarity with the prompt (block **305**), recognizing one of the multiple endpoints as having the set of the expert models stored thereon which have a closest match with the one of the multiple datasets (block **306**) and routing the prompt to the one of the multiple endpoints having the set of the expert models stored thereon which have the closest match with the one of the multiple datasets (block **307**). In accordance with embodiments, the receiving of the prompt of block **304**, the querying of the database of block **305**, the recognizing of the one of the multiple endpoints of block **306** and the routing of the prompt of block **307** are each executed by the serverless function router. In accordance with further embodiments, the serverless function router can include or be provided as a prompt aware router and the routing of block **307** can include prompt aware routing (block **3071**).

[0072] With reference to FIG. **4**, an architecture **400** is provided for an execution of the computer-implemented method **300** of FIG. **3**. The architecture **400** includes a prompt aware serverless function router **410**, vector database **420** and endpoints **430**, **440**, **450**. Each of the endpoints **430**, **440**, **450** can include or be provided as a server. Endpoint **430** has foundation models and fine-tuned models **431**, which have been built and trained previously for handling certain types of prompts (i.e., medical prompts), stored thereon. Endpoint **440** has foundation models and fine-tuned models **441**, which have been built and trained previously for handling certain types of prompts (i.e., financial prompts), stored thereon. Endpoint **450** has foundation models and fine-tuned models **451**, which have been built and trained previously for handling certain types of prompts (i.e., technical prompts), stored thereon. Datasets **421**, **422**, **423** are stored on the vector database **420**. Each of the datasets **421**, **422**, **423** has data or information that corresponds to or has a closest match with the foundation models and the fine-tuned models of one of the endpoints **430**, **440**, **450**. For example, dataset **421** can correspond to and closely match the foundation models and the fine-tuned models **431** of endpoint **430**, dataset **422** can correspond to and closely match the foundation models and the fine-tuned models **441** of endpoint **440** and dataset **423** can correspond to and closely match the foundation models and the fine-tuned models **451** of endpoint **450**.

[0073] When a prompt is received by the prompt aware serverless function router **410**, the prompt aware serverless function router **410** queries the vector database **420** for an indication as to which of the datasets **421**, **422**, **423** has the foundation models and the fine-tuned models that have the highest degree of similarity with the prompt. Upon receipt of the indication from the vector database **420**, the prompt aware serverless function router **410** determines which of

the endpoints **430**, **440**, **450** to send the prompt based on the indication and subsequently routes the prompt to that endpoint.

[0074] It is to be understood that, since each of the endpoints **430**, **440** and **450** only has subject matter expert models stored thereon and since the subject matter expert models are generally much smaller than an entire LLM, resource consumption on each of the endpoints **430**, **440** and **450** is limited (**100***s* or **10***s* of GBs for an LLM versus less than 10 GBs for a subject matter expert model). Thus, updates to the subject matter expert models and retraining, if needed, can be executed in relatively little time and with relatively little computing resources consumed.

[0075] With continued reference to FIG. **4** and with additional reference to FIGS. **5-7**, an operation of the prompt aware serverless function router **410** of FIG. **4** is at least partially based on a similarity of an incoming prompt and the foundation models and the fine-tuned models of endpoints **430**, **440** and **450**. In the example of FIGS. **5-7**, model A is specialized in medicine, model B is specialized in programming and model C is specialized in finance and the hypothetical prompts that are used for models A, B, and C represent the nature of their fields as shown in the table of FIG. **5**.

[0076] Given an incoming prompt "I don't feel comfortable", an embedding can be [0.5, 0.3, 0.4, 0.4, 0.4]. The Euclidean distance of similarity distance can then be calculated using the formula shown in FIG. **6**. This distance, which is the distance between the incoming prompt and the three hypothetical prompts are shown in the table of FIG. **7**. From FIG. **7**, it can be seen that the incoming prompt is closer to "Is fever a flu symptom" that represents model A. Therefore, in this example, the prompt aware serverless function router **10** will choose model A's server and route the prompt there.

[0077] Various embodiments of the present invention are described herein with reference to the related drawings. Alternative embodiments can be devised without departing from the scope of this invention. Although various connections and positional relationships (e.g., over, below, adjacent, etc.) are set forth between elements in the following description and in the drawings, persons skilled in the art will recognize that many of the positional relationships described herein are orientation-independent when the described functionality is maintained even though the orientation is changed. These connections and/or positional relationships, unless specified otherwise, can be direct or indirect, and the present invention is not intended to be limiting in this respect. Accordingly, a coupling of entities can refer to either a direct or an indirect coupling, and a positional relationship between entities can be a direct or indirect positional relationship. As an example of an indirect positional relationship, references in the present description to forming layer "A" over layer "B" include situations in which one or more intermediate layers (e.g., layer "C") is between layer "A" and layer "B" as long as the relevant characteristics and functionalities of layer "A" and layer "B" are not substantially changed by the intermediate layer(s).

[0078] The following definitions and abbreviations are to be used for the interpretation of the claims and the specification. As used herein, the terms "comprises," "comprising," "includes," "including," "has," "having," "contains" or "containing," or any other variation thereof, are intended to cover a non-exclusive inclusion. For example, a composition, a mixture, process, method, article, or apparatus that comprises a list of elements is not necessarily limited to only those elements but can include other elements not expressly listed or inherent to such composition, mixture, process, method, article, or apparatus.

[0079] Additionally, the term "exemplary" is used herein to mean "serving as an example, instance or illustration." Any embodiment or design described herein as "exemplary" is not necessarily to be construed as preferred or advantageous over other embodiments or designs. The terms "at least one" and "one or more" are understood to include any integer number greater than or equal to one, i.e. one, two, three, four, etc. The terms "a plurality" are understood to include any integer number greater than or equal to two, i.e. two, three, four, five, etc. The term "connection" can include an indirect "connection" and a direct "connection."

[0080] References in the specification to "one embodiment," "an embodiment," "an example embodiment," etc., indicate that the embodiment described can include a particular feature, structure, or characteristic, but every embodiment may or may not include the particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to affect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described.

[0081] For purposes of the description hereinafter, the terms "upper," "lower," "right," "left," "vertical," "horizontal," "top," "bottom," and derivatives thereof shall relate to the described structures and methods, as oriented in the drawing figures. The terms "overlying," "atop," "on top," "positioned on" or "positioned atop" mean that a first element, such as a first structure, is present on a second element, such as a second structure, wherein intervening elements such as an interface structure can be present between the first element and the second element. The term "direct contact" means that a first element, such as a first structure, and a second element, such as a second structure, are connected without any intermediary conducting, insulating or semiconductor layers at the interface of the two elements.

[0082] Spatially relative terms, e.g., "beneath," "below," "lower," "above," "upper," and the like, can be used herein for ease of description to describe one element or feature's relationship to another element(s) or feature(s) as illustrated in the figures. It will be understood that the spatially relative terms are intended to encompass different orientations of the device in use or operation in addition to the orientation depicted in the figures. For example, if the device in the figures is turned over, elements described as "below" or "beneath" other elements or features would then be oriented "above" the other elements or features. Thus, the term "below" can encompass both an orientation of above and below. The device can be otherwise oriented (rotated 90 degrees or at other orientations) and the spatially relative descriptors used herein interpreted accordingly.

[0083] The phrase "selective to," such as, for example, "a first element selective to a second element," means that the first element can be etched and the second element can act as an etch stop.

[0084] The terms "about," "substantially," "approximately," and variations thereof, are intended to include the

degree of error associated with measurement of the particular quantity based upon the equipment available at the time of filing the application. For example, "about" can include a range of +8% or 5%, or 2% of a given value.

[0085] The term "conformal" (e.g., a conformal layer) means that the thickness of the layer is substantially the same on all surfaces, or that the thickness variation is less than 15% of the nominal thickness of the layer.

[0086] The terms "epitaxial growth and/or deposition" and "epitaxially formed and/or grown" mean the growth of a semiconductor material (crystalline material) on a deposition surface of another semiconductor material (crystalline material), in which the semiconductor material being grown (crystalline overlayer) has substantially the same crystalline characteristics as the semiconductor material of the deposition surface (seed material). In an epitaxial deposition process, the chemical reactants provided by the source gases can be controlled and the system parameters can be set so that the depositing atoms arrive at the deposition surface of the semiconductor substrate with sufficient energy to move about on the surface such that the depositing atoms orient themselves to the crystal arrangement of the atoms of the deposition surface. An epitaxially grown semiconductor material can have substantially the same crystalline characteristics as the deposition surface on which the epitaxially grown material is formed. For example, an epitaxially grown semiconductor material deposited on a {100} orientated crystalline surface can take on a {100} orientation. In some embodiments of the invention, epitaxial growth and/or deposition processes can be selective to forming on semiconductor surface, and cannot deposit material on exposed surfaces, such as silicon dioxide or silicon nitride surfaces.

[0087] As previously noted herein, for the sake of brevity, conventional techniques related to semiconductor device and integrated circuit (IC) fabrication may or may not be described in detail herein. By way of background, however, a more general description of the semiconductor device fabrication processes that can be utilized in implementing one or more embodiments of the present invention will now be provided. Although specific fabrication operations used in implementing one or more embodiments of the present invention can be individually known, the described combination of operations and/or resulting structures of the present invention are unique. Thus, the unique combination of the operations described in connection with the fabrication of a semiconductor device according to the present invention utilize a variety of individually known physical and chemical processes performed on a semiconductor (e.g., silicon) substrate, some of which are described in the immediately following paragraphs.

[0088] In general, the various processes used to form a micro-chip that will be packaged into an IC fall into four general categories, namely, film deposition, removal/etching, semiconductor doping and patterning/lithography. Deposition is any process that grows, coats, or otherwise transfers a material onto the wafer. Available technologies include physical vapor deposition (PVD), chemical vapor deposition (CVD), electrochemical deposition (ECD), molecular beam epitaxy (MBE) and more recently, atomic layer deposition (ALD) among others. Removal/etching is any process that removes material from the wafer. Examples include etch processes (either wet or dry), and chemical-mechanical planarization (CMP), and the like. Semiconductor doping is the modification of electrical properties by doping, for example, transistor sources and drains, generally by diffusion and/or by ion implantation. These doping processes are followed by furnace annealing or by rapid thermal annealing (RTA). Annealing serves to activate the implanted dopants. Films of both conductors (e.g., polysilicon, aluminum, copper, etc.) and insulators (e.g., various forms of silicon dioxide, silicon nitride, etc.) are used to connect and isolate transistors and their components. Selective doping of various regions of the semiconductor substrate allows the conductivity of the substrate to be changed with the application of voltage. By creating structures of these various components, millions of transistors can be built and wired together to form the complex circuitry of a modern microelectronic device. Semiconductor lithography is the formation of three-dimensional relief images or patterns on the semiconductor substrate for subsequent transfer of the pattern to the substrate. In semiconductor lithography, the patterns are formed by a light sensitive polymer called a photo-resist. To build the complex structures that make up a transistor and the many wires that connect the millions of transistors of a circuit, lithography and etch pattern transfer steps are repeated multiple times. Each pattern being printed on the wafer is aligned to the previously formed patterns and slowly the conductors, insulators and selectively doped regions are built up to form the final device.

[0089] The flowchart and block diagrams in the Figures illustrate possible implementations of fabrication and/or operation methods according to various embodiments of the present invention. Various functions/operations of the method are represented in the flow diagram by blocks. In some alternative implementations, the functions noted in the blocks can occur out of the order noted in the Figures. For example, two blocks shown in succession can, in fact, be executed substantially concurrently, or the blocks can sometimes be executed in the reverse order, depending upon the functionality involved.

[0090] The descriptions of the various embodiments of the present invention have been presented for purposes of illustration, but are not intended to be exhaustive or limited to the embodiments described. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope and spirit of the described embodiments. The terminology used herein was chosen to best explain the principles of the embodiments, the practical application or technical improvement over technologies found in the marketplace, or to enable others of ordinary skill in the art to understand the embodiments described herein.

What is claimed is:

1. A computer-implemented method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon, the computer-implemented method comprising:

receiving a prompt;

querying a database comprising multiple datasets for an indication as to which one of the multiple datasets has a highest level of similarity with the prompt;

recognizing one of the multiple endpoints as having the set of the expert models stored thereon which have a closest match with the one of the multiple datasets; and

routing the prompt to the one of the multiple endpoints having the set of the expert models stored thereon which have the closest match with the one of the multiple datasets.

2. The computer-implemented method according to claim **1**, wherein the receiving of the prompt, the querying of the database, the recognizing of the one of the multiple endpoints and the routing of the prompt are executed by the serverless function router.

3. The computer-implemented method according to claim **1**, wherein the database comprises a vector database.

4. The computer-implemented method according to claim **1**, wherein each of the multiple datasets has a closest match with the subject matter expert models stored on one of the multiple endpoints.

5. The computer-implemented method according to claim **4**, wherein:

the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to medical subject matter,

the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to financial subject matter, and

the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to technical subject matter.

6. The computer-implemented method according to claim **4**, wherein the subject matter expert models comprise one or more foundation models and one or more fine-tuned models.

7. The computer-implemented method according to claim **1**, wherein the serverless function router is a prompt aware router and the routing comprises prompt aware routing.

8. A computer program product for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon, the computer program product comprising one or more computer readable storage media having computer readable program code collectively stored on the one or more computer readable storage media, the computer readable program code being executed by a processor of a computer system to cause the computer system to perform a method comprising:

receiving a prompt;

querying a database comprising multiple datasets for an indication as to which one of the multiple datasets has a highest level of similarity with the prompt;

recognizing one of the multiple endpoints as having the set of the expert models stored thereon which have a closest match with the one of the multiple datasets; and

routing the prompt to the one of the multiple endpoints having the set of the expert models stored thereon which have the closest match with the one of the multiple datasets.

9. The computer program product according to claim **8**, wherein the receiving of the prompt, the querying of the database, the recognizing of the one of the multiple endpoints and the routing of the prompt are executed by the serverless function router.

10. The computer program product according to claim **8**, wherein the database comprises a vector database.

11. The computer program product according to claim **8**, wherein each of the multiple datasets has a closest match with the subject matter expert models stored on one of the multiple endpoints.

12. The computer program product according to claim **11**, wherein:

the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to medical subject matter,

the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to financial subject matter, and

the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to technical subject matter.

13. The computer program product according to claim **11**, wherein the subject matter expert models comprise one or more foundation models and one or more fine-tuned models.

14. The computer program product according to claim **8**, wherein the serverless function router is a prompt aware router and the routing comprises prompt aware routing.

15. A computing system comprising:

a processor;

a memory coupled to the processor; and

one or more computer readable storage media coupled to the processor, the one or more computer readable storage media collectively containing instructions that are executed by the processor via the memory to implement a method for serving a large language model (LLM) application via a serverless function router communicative with multiple endpoints that each have a set of subject matter expert models stored thereon comprising:

receiving a prompt;

querying a database comprising multiple datasets for an indication as to which one of the multiple datasets has a highest level of similarity with the prompt;

recognizing one of the multiple endpoints as having the set of the expert models stored thereon which have a closest match with the one of the multiple datasets; and

routing the prompt to the one of the multiple endpoints having the set of the expert models stored thereon which have the closest match with the one of the multiple datasets.

16. The computing system according to claim **15**, wherein the receiving of the prompt, the querying of the database, the recognizing of the one of the multiple endpoints and the routing of the prompt are executed by the serverless function router.

17. The computing system according to claim **15**, wherein:

the database comprises a vector database, and

each of the multiple datasets has a closest match with the subject matter expert models stored on one of the multiple endpoints.

18. The computing system according to claim **17**, wherein:

the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to medical subject matter,

the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to financial subject matter, and

the subject matter expert models of a first one of the multiple endpoints is configured to handle prompts relating to technical subject matter.

19. The computing system according to claim **17**, wherein the subject matter expert models comprise one or more foundation models and one or more fine-tuned models.

20. The computing system according to claim **15**, wherein the serverless function router is a prompt aware router and the routing comprises prompt aware routing.

* * * * *