What Is Claimed Is:

1. A computer-implemented method for building a custom model comprising:

registering at least one module;

receiving a request from a user application;

filling a request queue with request information related to the request;

triggering the at least one module based on the request information;

processing at least part of the request based on data assets related to the request; and

building a custom model based on, at least in part, the processing of the at least part of the request.

2. The computer-implemented method of claim 1, wherein the request information includes at least one of: a request workflow, a list of one or more required modules, or a topic.

3. The computer-implemented method of claim 1, further comprising registering at least one second module and triggering the at least one second module based on the request information.

4. The computer-implemented method of claim 3, wherein the at least one module is at least one processing manager and the at least one second module is at least one request handing agent.

~~5. A computer-implemented method for building a custom model comprising:~~

~~registering, by a request queue, a custom machine learning model (MLM) server, at least one processing manager, and at least one request~~

~~handling agent;~~

~~receiving, by the custom MLM server, a request from a user application;~~

~~filling, by the custom MLM server, the request queue with request information related to the request, wherein the request information includes at least one of a request workflow, one or more required processing managers, or a topic;~~

~~triggering the at least one request handling agent based on the request information in the request queue;~~

~~triggering the at least one processing manager based on the request information as directed by the at least one request handling agent;~~

~~processing at least part of the request based on data assets related to the request using the at least one processing manager; and~~

~~building a custom model based on, at least in part, the processing of the at least part of the request.~~

~~6~~5. The computer-implemented method of claim ~~5~~1, wherein ~~the custom MLM server~~ filling the request queue with the request information comprises:

building~~, by the custom MLM server, the~~a request workflow; and

adding the request workflow to the request thereby modifying the request to an augmented request that is pushed into the request queue.

~~7~~6. The computer-implemented method of claim ~~6~~5, wherein the at least one ~~request handling agent~~module analyzes the augmented request for the request workflow, and wherein the ~~request handling agent~~ at least one module distributes information about required processing managers in the request queue having associated one or more topics.

~~8. The computer-implemented method of claim 7, wherein when the topic is marked in the request queue, the associated processing manager is triggered.~~

~~9. The computer-implemented method of claim 7, wherein the at least one request handling agent sends a message in the form of a further augmented request that is associated with the topic to the request queue, and wherein the request queue sends the further augmented request to the at least one processing manager associated with the topic triggering the at least one processing manager.~~

~~10~~7. The computer-implemented method of claim ~~7~~6, wherein ~~the~~a topic of the one or more topics of the augmented request is added or changed depending on a stage in the request workflow, and wherein the at least one ~~request handling agent~~ module determines that the augmented request relates to at least one processing manager or a group of processing managers based on the added or changed topic being associated with the at least one processing manager or the group of processing managers.

8. A computer program product residing on a computer readable storage medium having a plurality of instructions stored thereon which, when executed across one or more processors, causes at least a portion of the one or more processors to perform operations comprising:

registering at least one module;

receiving a request from a user application;

filling a request queue with request information related to the request;

triggering the at least one module based on the request information;

processing at least part of the request based on data assets related to the

request; and

building a custom model based on, at least in part, the processing of the at least part of the request.

9. The computer program product of claim 8, wherein the request information includes at least one of: a request workflow, a list of one or more required modules, or a topic.

10. The computer program product of claim 8, wherein the operations further comprise registering at least one second module and triggering the at least one second module based on the request information.

11. The computer program product of claim 10, wherein the at least one module is at least one processing manager and the at least one second module is at least one request handing agent.

12. The computer program product of claim 8, wherein filling the request queue with the request information comprises:

building a request workflow; and

adding the request workflow to the request thereby modifying the request to an augmented request that is pushed into the request queue.

13. The computer program product of claim 12, wherein the at least one module analyzes the augmented request for the request workflow, and wherein the at least one module distributes information about required processing managers in the request queue having associated one or more topics.

14. The computer program product of claim 13, wherein a topic of the one or more topics of the augmented request is added or changed depending on a stage in the request workflow, and wherein the at least one module determines that the augmented request relates to at least one processing manager or a group of processing managers based on the added or changed topic being associated with the at least one processing manager or the group of processing managers.

15. A computing system including one or more processors and one or more memories configured to perform operations comprising:

registering at least one module;

receiving a request from a user application;

filling a request queue with request information related to the request;

triggering the at least one module based on the request information;

processing at least part of the request based on data assets related to the request; and

building a custom model based on, at least in part, the processing of the at least part of the request.

16. The computing system of claim 15, wherein the request information includes at least one of: a request workflow, a list of one or more required modules, or a topic.

17. The computing system of claim 15, wherein the operations further comprise registering at least one second module and triggering the at least one second module based on the request information.

18. The computing system of claim 17, wherein the at least one module is at least one processing manager and the at least one second module is at least one request

handing agent.

19. The computing system of claim 15, wherein filling the request queue with the request information comprises:

building a request workflow; and

adding the request workflow to the request thereby modifying the request to an augmented request that is pushed into the request queue.

20. The computing system of claim 19, wherein the at least one module analyzes the augmented request for the request workflow, and wherein the at least one module distributes information about required processing managers in the request queue having associated one or more topics.

21. The computing system of claim 20, wherein a topic of the one or more topics of the augmented request is added or changed depending on a stage in the request workflow, and wherein the at least one module determines that the augmented request relates to at least one processing manager or a group of processing managers based on the added or changed topic being associated with the at least one processing manager or the group of processing managers.

~~11. The computer-implemented method of claim 5, wherein the custom MLM server filling the request queue with request information comprises creating, by the custom MLM server, the request workflow for the request, and wherein the request workflow includes descriptions of one or more processing steps required for fulfilling the request.~~

~~12. The computer-implemented method of claim 5, wherein when one of the one or~~

more processing steps is completed, the at least one request handling agent activates another next processing manager by marking a next topic in the augmented request of the request queue associated with the next processing manager.

13. The computer-implemented method of claim 5, wherein the custom MLM server filling the request queue with request information comprises the custom MLM server creating the request workflow for the request, and wherein the custom MLM server creates the request workflow by determining which processing managers are required for fulfilling the request and adds these required processing managers to the request workflow.

14. The computer-implemented method of claim 5, wherein the custom model is a language model such that the one or more processing steps of the workflow include a language modeling step, a grapheme-to-phoneme (G2P) step, and a decoding model building step.

15. The computer-implemented method of claim 14, wherein the at least one processing manager includes three different processing managers including a first processing manager, a second processing manager, and a third processing manager, wherein the first processing manager executes the language modeling step, the second processing manager executes the grapheme-to-phoneme (G2P) step, and the third processing manager executes the decoding model building step.

16. The computer-implemented method of claim 14, wherein the language modeling step includes normalizing of user data, the grapheme-to-phoneme (G2P) step includes phonetizing data, and the decoding model building step includes building of a decoding graph.

17. The computer-implemented method of claim 5, wherein the custom model is at least one of: a language model for automatic speech recognition (ASR), a natural language understanding (NLU) machine learning (ML) model, a natural language generation (NLG) ML model, a dialog management (DM) ML model, or a text-to-speech (TTS) ML model.

18. The computer-implemented method of claim 6, wherein the at least one processing manager is registered to monitor for the topic such that whenever a message is queued in the augmented request for the topic, the at least one processing manager associated with the topic receives the pushed message triggering the at least one processing manager.

19. The computer-implemented method of claim 5, wherein the built custom model is a decoding graph binary file.

20. The computer-implemented method of claim 6, wherein the custom MLM server, the at least one request handling agent, and the at least one processing manager are registered as at least one of publishers or subscribers.

21. The computer-implemented method of claim 6, wherein during registration, the custom MLM server creates one or more folders in the data assets to serve the request.

22. The computer-implemented method of claim 5, wherein the custom MLM server filling the request queue with request information includes modifying the request to associate any new request information with the request which converts the

request to an augmented request.

23. The computer-implemented method of claim 5, wherein the topic is an initial topic triggering the at least one request handling agent, wherein other topics relate to different stages of the request workflow, and wherein the initial topic of the augmented request is changed to at least one of the other topics causing at least one of an action or triggering of the at least one processing manager.

24. The computer-implemented method of claim 5, wherein the request includes text, an intended result, and ancillary data that is modified based on new information converting the request to an augmented request.

25. The computer-implemented method of claim 5, further comprising interpreting, by the at least one request handling agent, the request work for determining whether all required processing managers have completed their processing in relation to building the custom model.

26. The computer-implemented method of claim 25, if the at least one request handling agent determines that all required processing managers have completed their processing, then the custom MLM server is signaled of the fulfillment of the request.

27. The computer-implemented method of claim 5, wherein after the custom model is built, triggering a reload of the requested custom model in a runtime system.

28. A computing system for building a custom model including one or more processors and one or more memories configured to perform operations comprising:

registering, by a request queue, a custom machine learning model (MLM) server, at least one processing manager, and at least one request handling agent;

receiving, by the custom MLM server, a request from a user application;

filling, by the custom MLM server, the request queue with request information related to the request, wherein the request information includes at least one of a request workflow, one or more required processing managers, or a topic;

triggering the at least one request handling agent based on the request information in the request queue;

triggering the at least one processing manager based on the request information as directed by the at least one request handling agent;

processing at least part of the request based on data assets related to the request using the at least one processing manager; and

building a custom model based on, at least in part, the processing of the at least part of the request.

29. The computing system of claim 28, wherein the request queue is a facility for transporting information or data including requests from the custom MLM server to one or more processing managers.

30. The computing system of claim 28, wherein the request workflow includes a resulting workflow graph corresponding to the request including one or more processing steps required for fulfilling the request, wherein each workflow graph is created by the custom MLM server, and wherein each workflow graph is managed by the at least one processing manager.

~~31. The computing system of claim 28, wherein the at least one processing manager is sequenced through the at least one request handling agent.~~

~~32. The computing system of claim 28, wherein the at least one processing manager is associated with one or more docker images including at least one of resources or static files, input data, or output data.~~

~~33. The computing system of claim 28, wherein the data assets include at least one network file system server that stores at least one of: server resources, server input data, or server output data.~~

~~34. The computing system of claim 33, wherein the at least one processing manager includes a network file system client that receives information from the at least one network file system server via data asset links, and wherein the network file system client includes at least one of: client resources configured to receive information from the server resources, client input data configured to receive information from the server input data, or client output data configured to receive information from server output data.~~

~~35. The computing system of claim 34, wherein the operations further comprise:~~

~~starting the at least one network file system server and the at least one processing manager;~~

~~automatically mounting folders from the at least one network file system server;~~

~~starting a process associated with the at least one processing manager based on the server input data from the at least one network file system~~

server; and

writing processed output to the client output data of the network file system client.

36. The computing system of claim 35, wherein the process is at least one of:

a language modelling process for creating, interpolating, and pruning language models;

a grapheme-to-phoneme (G2P) process for providing G2P conversion; or

a decoding model building process for constructing, combining, optimizing, and searching weighted finite-state transducers (FSTs).

37. The computing system of claim 28, wherein the request queue includes data having default configuration data, service location data, and key/value data, and wherein the at least one processing manager includes a configuration management function configured to manage local application configuration using templates and the data from the request queue.

38. The computing system of claim 37, wherein each processing manager includes a different configuration management function, and wherein each configuration management function is used with a configuration management system for actively monitoring the key/value data for any key/value changes impacting configuration data.

39. The computing system of claim 37, wherein the at least one processing manager uses the configuration management function to sync configuration data by polling the request queue and processing template resources.

~~40. The computing system of claim 37, wherein the at least one processing manager uses the configuration management function with keys of the key/value data, and wherein at least one key of the keys activates the at least one processing manager.~~

~~41. The computing system of claim 28, wherein the custom MLM server includes an identity server configured to create, maintain, and manage identity information for principals while providing authentication services to relying applications within a network.~~