## CLAIMS

1. A task-oriented dialogue method, comprising:

in response to receiving dialogue information initiated by a user, sending, by a client, the dialogue information to a server;

in response to receiving the dialogue information, determining, by the server, intent information of the user through a large model, generating a tool execution plan according to the intent information, executing the tool execution plan to obtain a plan execution result, performing a check on the plan execution result, generating reply content based on the plan execution result that passes the check, and sending the reply content to the client, wherein generating the tool execution plan according to the intent information comprises:

retrieving at least one case from an execution plan experience repository through a predefined information retrieval model based on the intent information, and selecting at least one tool from a candidate tool list based on the at least one case to obtain the tool execution plan;

in response to determining that a matching degree between a case and the intent information is less than or equal to a preset threshold, configuring the case into a prompt template to select at least one tool from the candidate tool list based on the prompt template to obtain the tool execution plan; and

in response to determining that the plan execution result does not pass the check, generating feedback information based on the tool execution plan and a check result of the plan execution result, selecting at least one tool based on the feedback information and the intent information, and adding the at least one tool to the tool execution plan to obtain a modified tool execution plan; and

receiving and displaying, by the client, the reply content.

2. The method according to claim 1, wherein in response to receiving the dialogue information, determining, by the server, the intent information of the user through the large model comprises:

in response to receiving the dialogue information, acquiring, by the server, context information and predefined application scenario information corresponding to the dialogue information; and

determining, by the server, the intent information of the user based on the dialogue information and the context information by using the application scenario information as a constraint condition;

wherein determining, by the server, the intent information of the user based on the dialogue information and the context information by using the application scenario information as the constraint condition comprises:

rewriting, by the server, the dialogue information and the context information within a topic scope constrained by the application scenario information to obtain at least one piece of intent information of the user; and/or

retaining, by the server, information related to the application scenario information from the dialogue information and the context information, and generating at least one piece of intent information according to a theme of the retained content; and/or

retaining, by the server, time information and numerical information from the dialogue information and the context information in the intent information of the user; and/or

replacing, by the server, referential information from the dialogue information and the context information with name information, and retaining the name information in the intent information of the user; and/or

in response to identifying that the context information comprises a differentiated topic different from the dialogue information, deleting, by the server, content related to the differentiated topic from the intent information; and/or

in response to determining that the dialogue information and the context information are not within the topic scope constrained by the application scenario information, using, by the server, the dialogue information as the intent information of the user; and/or

retrieving, by the server, at least one case from an intent information experience repository through a predefined information retrieval model based on the dialogue information and the context information, and configuring the at least one case into a prompt template to determine the intent information of the user based on the prompt template, wherein the information retrieval model comprises at least one of a sparse retrieval model, a dense vector retrieval model, or a multi-vector retrieval model.

3. The method according to claim 1 or 2, wherein the intent information is generated by a pre-trained large model, and a pre-training method for the large model comprises:

performing, by the server, continued pre-training on a large model in a general domain based on a sample dataset in a predefined domain, wherein the sample dataset is provided with random noise and does not comprise sample user information;

performing, by the server, parameter fine-tuning on the large model after the continued pre-training based on an intent recognition case in the predefined domain; and

generating, by the server, intent recognition information of a predefined problem in the predefined domain based on the fine-tuned large model, and iteratively updating a parameter of the fine-tuned large model through a reinforcement learning algorithm based on positive and negative feedback information of the intent recognition information to obtain the pre-trained large model.

4. The method according to claim 1, wherein generating, by the server, the tool execution plan according to the intent information comprises:

acquiring, by the server, context information, and according to the context information, searching a tool library for a plurality of candidate tools related to the intent information to generate a candidate tool list; and

selecting, by the server, at least one tool from the candidate tool list based on the intent information, and arranging and combining the at least one tool to obtain the tool execution plan.

5. The method according to claim 4, wherein selecting, by the server, the at least one tool from the candidate tool list based on the intent information comprises:

performing, by the server, keyword matching on the candidate tool list based on the intent information to select at least one tool; and/or

in response to determining that intent information of a current round has a same meaning as intent information of a previous round, determining, by the server, a tool different from a tool of the previous round as a selected tool of the current round; and/or

determining, by the server, a tool mutually exclusive with an already selected tool as a disabled tool of the current round; and/or

in response to determining that no tool is selected based on the intent information, determining, by the server, a knowledge retrieval tool as a selected tool of the current round; and/or

wherein the information retrieval model comprises at least one of a sparse retrieval model, a dense vector retrieval model, or a multi-vector retrieval model; and

selecting, by the server, the at least one tool from the candidate tool list based on the at least one case comprises:

in response to determining that the matching degree between the case and the intent information is greater than the preset threshold, determining, by the server, a tool in the case as a selected tool.

6. The method according to claim 4, wherein arranging and combining the at least one tool to obtain the tool execution plan comprises:

arranging, by the server, tools in sequence and/or in parallel based on a dependency relationship between tool parameters of the tools; and

performing, by the server, operational connection on the tools arranged in sequence and/or in parallel through a predefined operator to obtain the tool execution plan; and/or

wherein the tool execution plan is generated by a pre-trained large model, and a pre-training method for the large model comprises:

performing, by the server, continued pre-training on a large model in a general domain based on a sample dataset in a predefined domain;

performing, by the server, parameter fine-tuning on the large model after the continued pre-training based on a tool invocation case in the predefined domain; and

converting, by the server, a parameter weight of the fine-tuned large model from a floating-point number to an integer with a preset number of bits to obtain the pre-trained large model.

7. The method according to claim 1, wherein executing, by the server, the tool execution plan to obtain the plan execution result comprises:

acquiring, by the server, context information, and extracting, from the dialogue information and the context information, tool parameters required by tools in the tool execution plan;

invoking, by the server, the tools sequentially and/or in parallel based on the tool parameters in accordance with an order of the tools as arranged and combined in the tool execution plan to obtain the plan execution result; and

wherein invoking, by the server, the tools sequentially and/or in parallel based on the tool parameters in accordance with the order of the tools as arranged and combined in the tool execution plan to obtain the plan execution result comprises:

invoking, by the server, tools arranged in sequence in the tool execution plan sequentially, and invoking tools arranged in parallel in the tool execution plan in parallel based on the tool parameters to obtain the plan execution result; and/or

wherein the method further comprises:

in response to identifying that there is a missing tool parameter, postponing, by the server, an execution order of a tool whose tool parameter is missing, and extracting the missing tool parameter from an execution result of an already executed tool; and/or

initiating, by the server, query information based on the missing tool parameter, and extracting the missing tool parameter from user response information to the query information; and/or

extracting, by the server, a time parameter from the tool parameters, and modifying relative time in the time parameter to absolute time; and/or

extracting, by the server, an order parameter from the tool parameters, and modifying a user expression in the order parameter to a standard expression in a predefined domain; and/or

the tools in the tool execution plan comprises at least one of a knowledge base search tool, an information query tool, a function configuration tool, or an information parsing tool.

8. The method according to claim 1, wherein performing the check, by the server, on the plan execution result comprises:

acquiring, by the server, context information, and performing a check on each tool in the tool execution plan and the plan execution result based on the dialogue information and the context information;

in response to determining that no tool is missing from the tool execution plan and/or the plan execution result is complete, determining, by the server, that the plan execution result passes the check; and/or

in response to determining that a tool is missing from the tool execution plan and/or the plan execution result is incomplete, cyclically executing, by the server, following process until a modified plan execution result passes the check, and/or a number of checks on the

plan execution result reaches a preset threshold: generating, by the server, feedback information based on the tool execution plan and a check result of the plan execution result, modifying the tool execution plan based on the feedback information and the intent information, and executing the modified tool execution plan to obtain the modified plan execution result.

9. The method according to claim 1, wherein generating, by the server, the reply content based on the plan execution result that passes the check comprises:

acquiring, by the server, context information, and integrating content related to the dialogue information in the plan execution result based on the dialogue information and the context information to obtain the reply content;

wherein integrating, by the server, the content related to the dialogue information in the plan execution result based on the dialogue information and the context information to obtain the reply content comprises:

deleting, by the server, content unrelated to the dialogue information in the plan execution result, and integrating content remained in the plan execution result to obtain the reply content; and/or

acquiring, by the server, at least one piece of link information based on the dialogue information, and adding the at least one piece of link information to the reply content; and/or

in response to identifying that the dialogue information comprises a negative emotion, generating, by the server, reassurance information based on the dialogue information, and adding the reassurance information to the reply content; and/or

in response to identifying that the dialogue information comprises human service requirement information, generating, by the server, guidance information based on the dialogue information, and adding the guidance information into the reply content; and/or

in response to identifying that the dialogue information does not match a predefined domain, generating, by the server, recognition rejection information based on the dialogue information, and using the recognition rejection information as the reply content; and/or

integrating, by the server, content related to the dialogue information in the plan execution result under a constraint condition of predefined word count and language style to generate structured reply content; and/or

retrieving, by the server, at least one case from a reply content experience repository through a predefined information retrieval model based on the intent information, and configuring the at least one case into a prompt template to determine the reply content based on the prompt template, wherein the information retrieval model comprises at least one of a sparse retrieval model, a dense vector retrieval model, or a multi-vector retrieval model.

10. The method according to claim 1, wherein the reply content is generated by a pre-trained large model, and a pre-training method for the large model comprises:

generating, by the server, a plurality of types of information reply cases based on a plurality of predefined dialogue scenarios in a predefined domain; and

performing, by the server, parameter fine-tuning on a predefined large model based on the plurality of types of information reply cases to obtain a pre-trained large model, wherein the predefined large model is a large model in a general domain or a large model obtained after continued pre-training with a sample dataset in a predefined domain; and/or

wherein the method further comprises:

generating, by the server, at least one type of case, and storing the case in a corresponding experience repository according to a type of the case, wherein the case comprises at least one of an intent recognition case, a tool invocation case, or an information reply case;

wherein generating the at least one type of case comprises:

receiving intent information, a work execution plan, and reply content generated during interaction of the user with the large model, and performing a feedback operation on the intent information, the work execution plan, and the reply content based on first positive and negative feedback information from the user regarding the intent information, the work execution plan, and the reply content to obtain the at least one type of case; and/or

receiving second positive and negative feedback information from the user regarding the reply content, and performing a feedback operation on intent

information, a work execution plan, and reply content generated in a process of generating the reply content based on the second positive and negative feedback information to obtain the at least one type of case.

11. The method according to claim 10, wherein performing, by the server, the feedback operation on the intent information, the work execution plan, and the reply content based on the first positive and negative feedback information to obtain the at least one type of case comprises:

in response to receiving positive feedback information from the user regarding the intent information, the work execution plan, and the reply content, performing, by the server, a positive feedback operation on a case corresponding to the intent information, the work execution plan, and the reply content corresponding to the positive feedback information, wherein the positive feedback operation comprises increasing an experience score of the case, and/or, adding a new case; and/or

in response to receiving negative feedback information from the user regarding the intent information, the work execution plan, and the reply content, performing, by the server, error correction processing on the intent information, the work execution plan, and the reply content corresponding to the negative feedback information to generate error correction information, and configuring the error correction information into a prompt template to regenerate intent information, a work execution plan, and reply content based on the prompt template;

wherein regenerating, by the server, the intent information, the work execution plan, and the reply content based on the prompt template comprises:

in response to identifying that a case corresponding to intent information, a work execution plan, and reply content that are generated through cases is erroneous, replacing, by the server, the erroneous case when regenerating intent information, a work execution plan, and reply content, and performing a negative feedback operation on the erroneous case, wherein the negative feedback operation comprises decreasing an experience score of the case and/or deleting the case; and/or

wherein performing, by the server, the feedback operation on the intent information, the work execution plan, and the reply content generated in the process of generating the reply content based on the second positive and negative feedback information to obtain the

at least one type of case comprises:

in response to receiving positive feedback information from the user regarding the reply content, performing, by the server, a positive feedback operation on a case corresponding to intent information, a work execution plan, and reply content that are generated through cases, wherein the positive feedback operation comprises increasing an experience score of the case; and/or

in response to receiving positive feedback information from the user regarding the reply content, adding, by the server, intent information, a work execution plan, and the reply content that are not generated through cases as a new case; and/or

in response to receiving negative feedback information from the user regarding the reply content, performing, by the server, a negative feedback operation on a case corresponding to intent information, a work execution plan, and reply content that are generated through cases, wherein the negative feedback operation comprises decreasing an experience score of the case and/or deleting the case; and/or

in response to receiving negative feedback information from the user regarding the reply content, deleting, by the server, intent information, a work execution plan, and reply content that are not generated through cases.

12. A task-oriented dialogue method, comprising:

in response to receiving dialogue information initiated by a user, sending the dialogue information;

receiving and displaying reply content, wherein a process of generating the reply content comprises:

determining intent information of the user through a large model based on the dialogue information, generating a tool execution plan according to the intent information, executing the tool execution plan to obtain a plan execution result, performing a check on the plan execution result, generating reply content based on the plan execution result that passes the check;

wherein generating the tool execution plan according to the intent information comprises:

retrieving at least one case from an execution plan experience repository through a predefined information retrieval model based on the

intent information, and selecting at least one tool from a candidate tool list based on the at least one case to obtain the tool execution plan;

in response to determining that a matching degree between a case and the intent information is less than or equal to a preset threshold, configuring the case into a prompt template to select at least one tool from the candidate tool list based on the prompt template to obtain the tool execution plan; and

in response to determining that the plan execution result does not pass the check, generating feedback information based on the tool execution plan and a check result of the plan execution result, selecting at least one tool based on the feedback information and the intent information, and adding the at least one tool to the tool execution plan to obtain a modified tool execution plan.

13. A task-oriented dialogue method, comprising:

in response to receiving dialogue information initiated by a user, determining intent information of the user through a large model;

generating a tool execution plan according to the intent information;

executing the tool execution plan to obtain a plan execution result;

performing a check on the plan execution result; and

generating reply content based on the plan execution result that passes the check;

wherein generating the tool execution plan according to the intent information comprises:

retrieving at least one case from an execution plan experience repository through a predefined information retrieval model based on the intent information, and selecting at least one tool from a candidate tool list based on the at least one case to obtain the tool execution plan;

in response to determining that a matching degree between a case and the intent information is less than or equal to a preset threshold, configuring the case into a prompt template to select at least one tool from the candidate tool list based on the prompt template to obtain the tool execution plan; and

in response to determining that the plan execution result does not pass the

check, generating feedback information based on the tool execution plan and a check result of the plan execution result, selecting at least one tool based on the feedback information and the intent information, and adding the at least one tool to the tool execution plan to obtain a modified tool execution plan.

14. A storage medium having a computer program stored thereon, wherein the computer program, when executed by one or more processors, implements the method according to any one of claims 1 to 13.

15. A computer device, comprising a storage medium, one or more processors, and a computer program stored on the storage medium and executable on the one or more processors, wherein the one or more processors, when executing the computer program, implement the method according to any one of claims 1 to 13.