LARGE MODEL–BASED QUESTION-ANSWERING METHOD AND APPARATUS, COMPUTER SYSTEM, AND STORAGE MEDIUM

TECHNICAL FIELD

[0001] One or more embodiments of this specification relate to the field of large models, and in particular, to a large model–based question-answering method and apparatus, a computer system, and a storage medium.

BACKGROUND

[0002] With the continuous advancement of artificial intelligence technologies, large language models have been gradually developed, bringing much convenience to people's work and life. Large language models are capable of understanding and generating natural languages, completing tasks such as translation, question answering, summarization, and text creation. Among these tasks, question answering based on large language models has broad applications. For example, users can ask questions to large language models, which can generate answers based on the users' questions and output the answers to the users. There is currently a need for a large model–based question-answering method for a specified knowledge document.

SUMMARY

[0003] One or more embodiments of this specification provide a large model–based question-answering method and apparatus, a computer system, and a storage medium.

[0004] According to a first aspect, a large model–based question-answering method is provided. The method includes: processing question information for a specified knowledge document based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model, so that the target large model outputs answer information based on the prompt information; obtaining update data, determined based on the answer information, for the specified knowledge document; and updating the knowledge base based on the update data.

[0005] According to a second aspect, a computer system is provided. The computer system includes a memory, a processor, and a computer program that is stored in the memory and that is capable of running on the processor. The processor executes the program to implement the following steps: processing question information for a specified knowledge document based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model, so that the target large model outputs answer information based on the prompt information; obtaining update data, determined based on the answer information, for the specified knowledge document; and updating the knowledge base based on the update data.

[0006] According to a third aspect, a large model–based question-answering apparatus is provided. The apparatus method includes: a processing module, configured to process question information for a specified knowledge document based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model, so that the target large model outputs answer information based on the prompt information; an obtaining module, configured to obtain update data, determined based on the answer information, for the specified knowledge document; and an updating module, configured to update the knowledge base based on the update data.

[0007] According to a fourth aspect, a computer-readable storage medium is provided. The storage medium stores a computer program, and the computer program is executed by a processor to implement the method according to any one of the implementations of the first aspect.

[0008] The technical solutions provided in embodiments of this specification may include the following beneficial effects:

[0009] According to the large model–based question-answering solution provided in the embodiments of this specification, question information for a specified knowledge document is processed based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model, so that the target large model outputs answer information based on the prompt information; update data, determined based on the answer information, for the specified knowledge document is obtained; and the knowledge base is updated based on the update data. This can ensure updating and optimization of the knowledge base without a need to train a large language model, thereby improving accuracy of question answering.

[0010] It should be understood that the foregoing general description and the following detailed description are merely exemplary and illustrative, and do not limit this application.

BRIEF DESCRIPTION OF DRAWINGS

[0011] To describe the technical solutions in embodiments of this specification more clearly, the following briefly describes the accompanying drawings needed for describing the embodiments. Clearly, the accompanying drawings in the following description are merely some embodiments of this specification, and a person of ordinary skill in the art can still derive other drawings from these accompanying drawings without creative efforts.

[0012] FIG. 1 is a system architectural diagram of asking a question about a particular technical document based on a large model according to an exemplary embodiment of this specification;

[0013] FIG. 2 is a flowchart of a large model–based question-answering method according to an exemplary embodiment of this specification;

[0014] FIG. 3 is a block diagram of a large model–based question-answering apparatus according to an exemplary embodiment of this specification; and

[0015] FIG. 4 is a block diagram of a computer system according to an exemplary embodiment of this specification.

DESCRIPTION OF EMBODIMENTS

[0016] To make a person skilled in the art better understand the technical solutions in this specification, the following clearly and comprehensively describes the technical solutions in the embodiments of this specification with reference to the accompanying drawings in the embodiments of this specification. Clearly, the described embodiments are merely some rather than all of the embodiments of this specification. All other embodiments obtained by a person of ordinary skill in the art based on the embodiments of this specification without creative efforts shall fall within the protection scope of this specification.

[0017] On the one hand, with the continuous advancement of artificial intelligence technologies, large language models have been gradually developed, bringing much convenience to people's work and life. Large language models are capable of understanding and generating

natural languages, completing tasks such as translation, question answering, summarization, and text creation. Among these tasks, question answering based on large language models has broad applications. For example, users can ask questions to large language models, which can generate answers based on the users' questions and output the answers to the users.

[0018] On the other hand, in some scenarios, for example, an organization A develops a new system B, which is provided with a related technical document. A user of the system B needs to check the related technical document to learn information such as an operation usage specification and a technical introduction of the system B. However, because the technical document may have a relatively large amount of information and a relatively large amount of content, the user cannot quickly focus on a specific issue in the technical document. In a related technology, a dedicated person is usually responsible for answering a question raised by a user about the technical document. This requires human resources. In addition, efficiency of manually querying and answering a question is usually low, and accuracy of an answer obtained cannot be ensured.

[0019] Therefore, it is conceivable to apply a large language model technology to this scenario, and use a pre-trained large language model to help the user understand detailed content in the technical document. For example, the user may ask a question about the technical document to the large language model, and the large language model answers the question. However, for some technical documents that involve narrow, highly specialized fields, it may be difficult for a general-purpose large language model to accurately comprehend the technical documents. If a large language model needs to be used to answer a question about a technical document, it is better to pre-train the large language model by using a corpus obtained based on the technical document. However, training the large language model is highly costly and complex. Especially when the technical document needs to be frequently updated, a manner of helping the user understand the technical document by using the large language model is more difficult to implement.

[0020] According to a large model–based question-answering solution provided in embodiments of this specification, question information for a specified knowledge document (for example, a technical document related to a particular technical product) is processed based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model, so that the target large model outputs answer information based on the prompt information; update data, determined based on the answer information, for the specified knowledge document is obtained; and the knowledge base is updated

based on the update data. This can ensure updating and optimization of the knowledge base without a need to train a large language model, thereby improving accuracy of question answering.

[0021] FIG. 1 is a system architectural diagram of asking a question about a particular technical document based on a large model according to an exemplary embodiment. As shown in FIG. 1, a terminal device 101 is a terminal device held by a user C. The user C may log in to a system usage page corresponding to a technical document D1 by using the terminal device 101. An interface 1 connected to an auxiliary platform 102 is embedded into the system usage page corresponding to the technical document D1. The user C may input question information for the technical document D1 by using the interface 1, and the interface 1 may transmit the question information to the auxiliary platform 102.

[0022] The auxiliary platform 102 may be a platform built for a particular technical document, and the auxiliary platform 102 may include a processing part 103 and a storage part 104. The storage part 104 stores at least data corresponding to the technical document D1. The data corresponding to the technical document D1 may include information such as document content information, document keyword information, and a specialized term and an explanation thereof. The processing part 103 may process the question information based on the data corresponding to the technical document D1 stored in the storage part 104 to obtain prompt information, and transmit the prompt information to a large model 105.

[0023] The large model 105 may be a general-purpose large language model (LLM) available on the market. The large model 105 may obtain answer information based on the prompt information transmitted by the auxiliary platform 102, and transmit the answer information back to the auxiliary platform 102. The auxiliary platform 102 returns the answer information to the terminal device 101, and then the terminal device 101 displays or plays the answer information to the user C.

[0024] In addition, an interface 2 connected to an operation and maintenance platform 106 is further embedded into the system usage page corresponding to the technical document D1. The user C may input feedback information based on the answer information by using the interface 2 connected to the operation and maintenance platform 106, and the interface 2 transmits the feedback information to the operation and maintenance platform 106. The operation and maintenance platform 106 may display the feedback information to operation and maintenance personnel, and the operation and maintenance personnel determine, based on the feedback

information, update data that needs to be updated in the storage part 104, and input the update data into the operation and maintenance platform 106. Then, the operation and maintenance platform 106 transmits the update data to the auxiliary platform 102, so as to update the data stored in the storage part 104.

[0025] In addition, before the processing part 103 or the storage part 104 in the auxiliary platform 102 is updated, or before a version of the large model 105 is upgraded, the operation and maintenance platform 106 may provide a pre-constructed reference question-answer library. It may be verified, based on the reference question-answer library, whether the data used to update the processing part 103 or the storage part 104 in the auxiliary platform 102, or the new version of the large model degrades overall system performance. If the overall system performance is not degraded, the update can be completed.

[0026] The following describes in detail a process of asking a question by using the foregoing system with reference to a specific application instance. Specifically, first, the user C logs in to the system usage page corresponding to the technical document D1 by using the terminal device 101, and inputs the question information by using the interface 1 on the system usage page. The question information may be a question expressed in a natural language. For example, the question information may be "how to transmit information under the abc protocol" (abc is a special name for a new protocol in a new product). The question information may be transmitted to the auxiliary platform 102 by using the interface 1.

[0027] Then, the processing part 103 included in the auxiliary platform 102 may first perform semantic recognition on the question information by using a semantic recognition algorithm, to obtain a word sequence formed by a semantic keyword. In addition, the processing part 103 obtains, from the storage part 104 included in the auxiliary platform 102, a term list pre-constructed for the technical document D1. The term list records specialized terms involved in the technical document D1 and explanatory information corresponding to each specialized term. The processing part 103 may detect whether the word sequence includes a specialized term recorded in the term list, and if the word sequence includes a specialized term, may update the word sequence by using explanatory information corresponding to the specialized term. For example, the processing part 103 detects that the word sequence includes the specialized term abc recorded in the term list, and obtains explanatory information corresponding to the specialized term abc from the term list. Then, the processing part 103 may update the abc in the word sequence by using the explanatory

information corresponding to the specialized term abc. For example, after the explanatory information corresponding to the specialized term abc is added to the abc in the word sequence, a new word sequence is obtained.

[0028] Then, the processing part 103 may perform risk detection on the word sequence to identify whether the word sequence includes a sensitive word. Specifically, the processing part 103 may first perform risk detection on the word sequence by using a locally deployed risk detection module, and when the detection succeeds, perform further risk detection by using an internationally accepted risk detection tool through an external interface.

[0029] When two levels of risk detection succeed, the processing part 103 may obtain, from the storage part 104, document content information and document keyword information that correspond to the technical document D1, and search, based on the document keyword information, the document content information for partial document content related to the semantic keyword included in the word sequence. Specifically, the document keyword information may be obtained based on a plurality of keywords extracted in advance from document content of the technical document D1. The processing part 103 may find, from the document keyword information by using a preset similarity algorithm, a document keyword related to the semantic keyword included in the word sequence, and obtain, based on the related document keyword, partial document content related to the question information from the document content information. Prompt information may be generated based on the partial document content. The prompt information may include some requirements for an answer, the found partial document content, and the marked question information. For example, the prompt information may include content of the abc protocol, prompt information with a preset form of separator added between characters "how-to-transmit-information-under-the-a-b-c-protocol", and a rule that needs to be followed by the large model 105 when answering a question.

[0030] The processing part 103 may transmit the generated prompt information to the large model 105, and the large model 105 analyzes the prompt information to obtain answer information for the question information, and transmits the answer information back to the processing part 103. The processing part 103 may further perform risk detection on the answer information, for example, may perform risk detection on the answer information by using a locally deployed risk detection module and/or an external common risk detection tool. If the risk detection succeeds, the processing part 103 may return the answer information to the terminal device 101, and the terminal

device 101 displays the answer information on the system usage page corresponding to the technical document D1 for the user C to view.

[0031] In addition, after viewing the answer information, the user C may input feedback information for the answer information by using the interface 2 on the system usage page. The interface 2 may transmit the feedback information to the operation and maintenance platform 106. The feedback information may be a comment of the user C on the answer information, such as like or dislike information, evaluation information, or correction suggestion information. For example, the user C feels that the definition content of the abc protocol is unclear or ambiguous, and may input a comment into the interface 2 in a form of feedback information. The comment is transmitted to the operation and maintenance platform 106. The operation and maintenance platform 106 may display the feedback information to the operation and maintenance personnel by using a display device. The operation and maintenance personnel may refer to the feedback information to determine update data that needs to be updated in the storage part 104. The update data may be the document content or the like corresponding to the technical document D1. For example, the operation and maintenance personnel may refer to the feedback information to re-edit the definition content of the abc protocol in the technical document D1 as update data for the technical document D1.

[0032] When updating the auxiliary platform 102 (for example, upgrading the processing part 103 or updating the document content corresponding to the technical document D1 in the storage part 104), the operation and maintenance platform 106 may perform an update test. Specifically, the operation and maintenance platform 106 may first obtain a pre-constructed reference question-answer library. The reference question-answer library includes a plurality of groups of question-answer information prepared in advance, and each group of question-answer information includes one question and one answer to the question. The operation and maintenance platform 106 may first obtain some or all of the questions from the reference question-answer library, and then transmit the some or all of the questions to the large model 105 by using the updated auxiliary platform 102. The answer to each question output by the large model 105 is obtained as a test answer, and the answers corresponding to the some or all of the questions are obtained from the reference question-answer library as reference answers. The test answer and the reference answer may be compared. If a result of the comparison indicates that a difference between the test answer and the reference answer is less than or equal to a preset difference, it indicates that the update test

succeeds, and an update of the auxiliary platform 102 can be completed. If the result of the comparison indicates that the difference between the test answer and the reference answer is greater than the preset difference, it indicates that the update test fails, and the auxiliary platform 102 can be rolled back to a state before the update. In addition, the update data for the auxiliary platform 102 may be reset to re-update the auxiliary platform 102.

[0033] It should be noted that when the version of the large model 105 is upgraded, the operation and maintenance platform 106 may also perform the foregoing update test. If the update test succeeds, upgrading the version of the large model 105 can be completed. If the update test fails, the large model of the previous version may be used.

[0034] The following describes in detail the solutions provided in this specification with reference to specific embodiments.

[0035] FIG. 2 is a flowchart of a large model–based question-answering method according to an exemplary embodiment. The method may be applied to a service platform. The service platform may be implemented as any device, platform, server, or device cluster that has computing and processing capabilities. The method includes the following steps:

[0036] In step 201, question information for a specified knowledge document is processed based on a knowledge base pre-constructed based on the specified knowledge document to obtain prompt information, and the prompt information is transmitted to a target large model.

[0037] In this embodiment, the knowledge base may be pre-constructed based on the specified knowledge document. The specified knowledge document may be a technical document specific to a particular product or a professional field. For example, the specified knowledge document may be a technical introduction document of a new device or a new system that is developed, or may be a knowledge document or the like that is in an independent field. Specific content of the specified knowledge document is not limited in this embodiment. After the specified knowledge document is obtained, some keywords may be extracted from the specified knowledge document, and a word vector is generated based on these keywords. In addition, the specified knowledge document may be divided into a plurality of segments. Each segment includes partial document content of the specified knowledge document, and each segment corresponds to one or more word vectors. The foregoing word vector and the foregoing plurality of fragments may be stored in a knowledge base constructed for the specified knowledge document.

[0038] In this embodiment, the service platform may process the question information for the

specified knowledge document based on the pre-constructed knowledge base, to obtain the prompt information to be transmitted to the target large model. Optionally, before the question information is processed, risk detection may be further performed on the question information. An operation of processing the question information is performed only when the question information does not include a sensitive word. For example, risk detection may be performed on the question information by using locally deployed risk detection, or risk detection may be performed on the question information by using an externally provided common risk detection tool through an external interface, which is not limited in this embodiment. After the risk detection succeeds, the operation of processing the question information may be performed.

[0039] In a process of processing the question information, semantic recognition may be first performed on the user-input question information by using a preset semantic recognition algorithm, to obtain a target word sequence formed by a semantic keyword. The target word sequence may be, for example, a sequence obtained by ranking the semantic keywords according to a sequence corresponding to the semantic keywords in the question information. After the target word sequence is obtained, prompt information may be generated directly based on the target word sequence. Alternatively, term detection may be first performed on the target word sequence, and then prompt information is generated after the target word sequence is processed based on a detection result.

[0040] Specifically, term detection may be performed on the target word sequence in the following manner: At least one specialized term in the specified knowledge document may be obtained in advance; a term list is generated, and explanatory information corresponding to each specialized term in the specified knowledge document is recorded in the term list; and the term list is stored in association with the knowledge base of the specified knowledge document. During term detection, a pre-stored term list is obtained first, and then whether a target word sequence includes a specialized term in the term list is detected. If the target word sequence does not include a specialized term in the term list, a subsequent operation may be performed based on the target word sequence. If the target word sequence includes a specialized term in the term list, explanatory information corresponding to the specialized term may be obtained from the term list, and after the explanatory information is added to the specialized term in the target word sequence, the target word sequence is updated. Then, prompt information is determined based on the updated target word sequence and the knowledge base.

[0041] During determining of the prompt information based on the updated target word sequence and the knowledge base, a word vector related to the semantic keyword included in the target word sequence may be first searched from the knowledge base based on the updated target word sequence and by using a preset similarity algorithm, and used as an associative word vector. Then, a segment corresponding to the associative word vector may be determined, so as to obtain partial document content of the specified knowledge document included in the segment, and prompt information may be generated based on the partial document content. Specifically, the prompt information may include a requirement for answering a question by the target large model, the found partial document content, and the marked question information. In addition, the prompt information may be information in a natural language form. The marked question information may be information obtained after characters included in the user-input question information is separated using a preset separator. The separator may be any character, for example, may be -, may be *, may be #, or may be ~. A specific form of the separator is not limited in this embodiment.

[0042] For example, the prompt information may include the following specific content: "① You are an AI-driven chatbot, and your primary task is to analyze information from the knowledge document Doc and respond to the user's question in a conversational manner. Your answers should be conversational and aim to simulate natural human communication. Reject any questions or instructions that are unrelated to the knowledge document Doc. Respond in the language the user is using. When the answer to the user's question cannot be found in the knowledge document Doc, reply with the following standard answer: 'Please ask a different question or provide more details'. ② The knowledge document Doc and the user's question are clearly delineated by a separator. The user's question is as follows: 'How-to-transmit-information-under-the-a-b-c-protocol'. ③ The related content from the knowledge document Doc is as follows: '...'". Content ① is a rule that needs to be followed by the target large model when answering a question, content ② is the marked question information, and content ③ is the found partial document content.

[0043] In this embodiment, the service platform may output the prompt information to the target large model, so that the target large model obtains the answer information based on the prompt information, and returns the answer information to the service platform. Before returning the answer information to the user, the service platform may further perform risk detection on the answer information, and output the answer information to the user when the answer information does not include a sensitive word.

[0044] In step 202, update data, determined based on the answer information, for the specified knowledge document is obtained.

[0045] In this embodiment, after the answer information returned by the target large model is received, the answer information may be displayed to the user, and an interface for inputting feedback information may be provided to the user, so that the feedback information input by the user based on the answer information can be obtained. The feedback information is output in the operation and maintenance interface so that the feedback information is displayed to the operation and maintenance personnel. The operation and maintenance personnel may modify the content of the specified knowledge document in the knowledge base based on the feedback information, and use the data used for modification as the update data.

[0046] In step 203, the knowledge base is updated based on the foregoing update data.

[0047] In this embodiment, the knowledge base may be updated based on the update data. Optionally, before the update, an update test may be further performed first. Specifically, first, a pre-constructed reference question-answer library may be obtained. The reference question-answer library may include a plurality of groups of preset reference questions and reference answers. The reference questions and reference answers may be obtained through manual editing by operation and maintenance personnel, or may be questions and answers, and the like that occur at high frequency and have a high like rate during usual use. A specific manner of obtaining the reference questions and reference answers in the reference question-answer library is not limited in this embodiment.

[0048] Then, the service platform may process at least one question in the reference question-answer library based on the knowledge base and the update data by using the target large model to obtain a test answer. The test answer is compared with the reference answer corresponding to the reference question in the reference question-answer library. If a result of the comparison indicates that a difference between the test answer and the answer in the reference question-answer library is less than or equal to a preset difference, the knowledge base is updated by using the update data. If the result of the comparison indicates that the difference between the test answer and the answer in the reference question-answer library is greater than the preset difference, the update can be stopped and the content of the specified knowledge document in the knowledge base is further modified.

[0049] According to the large model–based question-answering method provided in the

embodiments of this specification, question information for a specified knowledge document is processed based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model, so that the target large model outputs answer information based on the prompt information; update data, determined based on the answer information, for the specified knowledge document is obtained; and the knowledge base is updated based on the update data. This can ensure updating and optimization of the knowledge base without a need to train a large language model, thereby improving accuracy of question answering.

[0050] It should be noted that although in the foregoing embodiments, the operations of the methods in the embodiments of this specification are described in a specific order, this does not require or imply that these operations necessarily be performed in the specific order, or that all the operations shown necessarily be performed to achieve a desired result. On the contrary, the steps depicted in the flowchart may be performed in another order. Additionally or alternatively, some steps may be omitted, a plurality of steps may be combined into one step for execution, and/or one step may be broken down into a plurality of steps for execution.

[0051] Corresponding to the foregoing embodiments of the large model–based question-answering method, this specification further provides an embodiment of a large model–based question-answering apparatus.

[0052] FIG. 3 is a block diagram of a large model–based question-answering apparatus according to an exemplary embodiment of this specification. The apparatus may include a processing module 301, an obtaining module 302, and an updating module 303.

[0053] The processing module 301 is configured to process question information for a specified knowledge document based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model, so that the target large model outputs answer information based on the prompt information.

[0054] The obtaining module 302 is configured to obtain update data, determined based on the answer information, for the specified knowledge document.

[0055] The updating module 303 is configured to update the knowledge base based on the update data.

[0056] For ease of description, the foregoing apparatus is described by dividing functions into various modules. Certainly, during implementation of this specification, functions of all modules

may be implemented in one or more pieces of software and/or hardware.

[0057] The specification also provides a computer-readable non-transitory storage medium, wherein the storage medium stores a computer program, which, when executed by a processor, can be used to perform one or more steps of one or more methods described or illustrated herein or provides functionality described or illustrated herein. Herein, the computer-readable non-transitory storage medium or media may include one or more semiconductor-based or other integrated circuits (ICs) (such, as for example, field-programmable gate arrays (FPGAs) or application-specific ICs (ASICs)), hard disk drives (HDDs), hybrid hard drives (HHDs), optical discs, optical disc drives (ODDs), magneto-optical discs, magneto-optical drives, floppy diskettes, floppy disk drives (FDDs), magnetic tapes, solid-state drives (SSDs), RAM-drives, SECURE DIGITAL cards or drives, any other suitable computer-readable non-transitory storage media, or any suitable combination of two or more of these, where appropriate. A computer-readable non-transitory storage medium may be volatile, non-volatile, or a combination of volatile and non-volatile, where appropriate.

[0058] The specification also provides a computer system. FIG. 4 illustrates an example computer system 1000. In particular embodiments, one or more computer systems 1000 perform one or more steps of one or more methods described or illustrated herein. In particular embodiments, one or more computer systems 1000 provide functionality described or illustrated herein. In particular embodiments, software running on one or more computer systems 1000 performs one or more steps of one or more methods described or illustrated herein or provides functionality described or illustrated herein. Particular embodiments include one or more portions of one or more computer systems 1000. Herein, reference to a computer system may encompass a computing device, and vice versa, where appropriate. Moreover, reference to a computer system may encompass one or more computer systems, where appropriate.

[0059] This disclosure contemplates any suitable number of computer systems 1000. This disclosure contemplates computer system 1000 taking any suitable physical form. As example and not by way of limitation, computer system 1000 may be an embedded computer system, a system-on-chip (SOC), a single-board computer system (SBC) (such as, for example, a computer-on-module (COM) or system-on-module (SOM)), a desktop computer system, a laptop or notebook computer system, an interactive kiosk, a mainframe, a mesh of computer systems, a mobile telephone, a personal digital assistant (PDA), a server, a tablet computer system, or a combination

of two or more of these. Where appropriate, computer system 1000 may include one or more computer systems 1000; be unitary or distributed; span multiple locations; span multiple machines; span multiple data centers; or reside in a cloud, which may include one or more cloud components in one or more networks. Where appropriate, one or more computer systems 1000 may perform without substantial spatial or temporal limitation one or more steps of one or more methods described or illustrated herein. As an example and not by way of limitation, one or more computer systems 1000 may perform in real time or in batch mode one or more steps of one or more methods described or illustrated herein. One or more computer systems 1000 may perform at different times or at different locations one or more steps of one or more methods described or illustrated herein, where appropriate.

[0060] In particular embodiments, computer system 1000 includes a processor 1002, memory 1004, storage 1006, an input/output (I/O) interface 1008, a communication interface 1010, and a bus 1012. Although this disclosure describes and illustrates a particular computer system having a particular number of particular components in a particular arrangement, this disclosure contemplates any suitable computer system having any suitable number of any suitable components in any suitable arrangement.

[0061] In particular embodiments, processor 1002 includes hardware for executing instructions, such as those making up a computer program. As an example and not by way of limitation, to execute instructions, processor 1002 may retrieve (or fetch) the instructions from an internal register, an internal cache, memory 1004, or storage 1006; decode and execute them; and then write one or more results to an internal register, an internal cache, memory 1004, or storage 1006. In particular embodiments, processor 1002 may include one or more internal caches for data, instructions, or addresses. This disclosure contemplates processor 1002 including any suitable number of any suitable internal caches, where appropriate. As an example and not by way of limitation, processor 1002 may include one or more instruction caches, one or more data caches, and one or more translation lookaside buffers (TLBs). Instructions in the instruction caches may be copies of instructions in memory 1004 or storage 1006, and the instruction caches may speed up retrieval of those instructions by processor 1002. Data in the data caches may be copies of data of 80 in memory 1004 or storage 1006 for instructions executing at processor 1002 to operate on; the results of previous instructions executed at processor 1002 for access by subsequent instructions executing at processor 1002 or for writing to memory 1004 or storage 1006; or other

suitable data. The data caches may speed up read or write operations by processor 1002. The TLBs may speed up virtual-address translation for processor 1002. In particular embodiments, processor 1002 may include one or more internal registers for data, instructions, or addresses. This disclosure contemplates processor 1002 including any suitable number of any suitable internal registers, where appropriate. Where appropriate, processor 1002 may include one or more arithmetic logic units (ALUs); be a multi-core processor; or include one or more processors 1002. Although this disclosure describes and illustrates a particular processor, this disclosure contemplates any suitable processor.

[0062] In particular embodiments, memory 1004 includes main memory for storing instructions for processor 1002 to execute or data for processor 1002 to operate on. As an example and not by way of limitation, computer system 1000 may load instructions from storage 1006 or another source (such as, for example, another computer system 1000) to memory 1004. Processor 1002 may then load the instructions from memory 1004 to an internal register or internal cache. To execute the instructions, processor 1002 may retrieve the instructions from the internal register or internal cache and decode them. During or after execution of the instructions, processor 1002 may write one or more results (which may be intermediate or final results) to the internal register or internal cache. Processor 1002 may then write one or more of those results to memory 1004. In particular embodiments, processor 1002 executes only instructions in one or more internal registers or internal caches or in memory 1004 (as opposed to storage 1006 or elsewhere) and operates only on data in one or more internal registers or internal caches or in memory 1004 (as opposed to storage 1006 or elsewhere). One or more memory buses (which may each include an address bus and a data bus) may couple processor 1002 to memory 1004. Bus 1012 may include one or more memory buses, as described below. In particular embodiments, one or more memory management units (MMUs) reside between processor 1002 and memory 1004 and facilitate accesses to memory 1004 requested by processor 1002. In particular embodiments, memory 1004 includes random access memory (RAM). This RAM may be volatile memory, where appropriate. Where appropriate, this RAM may be dynamic RAM (DRAM) or static RAM (SRAM). Moreover, where appropriate, this RAM may be single-ported or multi-ported RAM. This disclosure contemplates any suitable RAM. Memory 1004 may include one or more memories 1004, where appropriate. Although this disclosure describes and illustrates particular memory, this disclosure contemplates any suitable memory.

[0063] In particular embodiments, storage 1006 includes mass storage for data or instructions. As an example and not by way of limitation, storage 1006 may include a hard disk drive (HDD), a floppy disk drive, flash memory, an optical disc, a magneto-optical disc, magnetic tape, or a Universal Serial Bus (USB) drive or a combination of two or more of these. Storage 1006 may include removable or non-removable (or fixed) media, where appropriate. Storage 1006 may be internal or external to computer system 1000, where appropriate. In particular embodiments, storage 1006 is non-volatile, solid-state memory. In particular embodiments, storage 1006 includes read-only memory (ROM). Where appropriate, this ROM may be mask-programmed ROM, programmable ROM (PROM), erasable PROM (EPROM), electrically erasable PROM (EEPROM), electrically alterable ROM (EAROM), or flash memory or a combination of two or more of these. This disclosure contemplates mass storage 1006 taking any suitable physical form. Storage 1006 may include one or more storage control units facilitating communication between processor 1002 and storage 1006, where appropriate. Where appropriate, storage 1006 may include one or more storages 1006. Although this disclosure describes and illustrates particular storage, this disclosure contemplates any suitable storage.

[0064] In particular embodiments, I/O interface 1008 includes hardware, software, or both, providing one or more interfaces for communication between computer system 1000 and one or more I/O devices. Computer system 1000 may include one or more of these I/O devices, where appropriate. One or more of these I/O devices may enable communication between a person and computer system 1000. As an example and not by way of limitation, an I/O device may include a keyboard, keypad, microphone, monitor, mouse, printer, scanner, speaker, still camera, stylus, tablet, touch screen, trackball, video camera, another suitable I/O device or a combination of two or more of these. An I/O device may include one or more sensors. This disclosure contemplates any suitable I/O devices and any suitable I/O interfaces 1008 for them. Where appropriate, I/O interface 1008 may include one or more device or software drivers enabling processor 1002 to drive one or more of these I/O devices. I/O interface 1008 may include one or more I/O interfaces 1008, where appropriate. Although this disclosure describes and illustrates a particular I/O interface, this disclosure contemplates any suitable I/O interface.

[0065] In particular embodiments, communication interface 1010 includes hardware, software, or both providing one or more interfaces for communication (such as, for example, packet-based communication) between computer system 1000 and one or more other computer systems 1000 or

one or more networks. As an example and not by way of limitation, communication interface 1010 may include a network interface controller (NIC) or network adapter for communicating with an Ethernet or other wire-based network or a wireless NIC (WNIC) or wireless adapter for communicating with a wireless network, such as a WI-FI network. This disclosure contemplates any suitable network and any suitable communication interface 1010 for it. As an example and not by way of limitation, computer system 1000 may communicate with an ad hoc network, a personal area network (PAN), a local area network (LAN), a wide area network (WAN), a metropolitan area network (MAN), or one or more portions of the Internet or a combination of two or more of these. One or more portions of one or more of these networks may be wired or wireless. As an example, computer system 1000 may communicate with a wireless PAN (WPAN) (such as, for example, a BLUETOOTH WPAN), a WI-FI network, a WI-MAX network, a cellular telephone network (such as, for example, a Global System for Mobile Communications (GSM) network), or other suitable wireless network or a combination of two or more of these. Computer system 1000 may include any suitable communication interface 1010 for any of these networks, where appropriate. Communication interface 1010 may include one or more communication interfaces 1010, where appropriate. Although this disclosure describes and illustrates a particular communication interface, this disclosure contemplates any suitable communication interface.

[0066] In particular embodiments, bus 1012 includes hardware, software, or both coupling components of computer system 1000 to each other. As an example and not by way of limitation, bus 1012 may include an Accelerated Graphics Port (AGP) or other graphics bus, an Enhanced Industry Standard Architecture (EISA) bus, a front-side bus (FSB), a HYPERTRANSPORT (HT) interconnect, an Industry Standard Architecture (ISA) bus, an INFINIBAND interconnect, a low-pin-count (LPC) bus, a memory bus, a Micro Channel Architecture (MCA) bus, a Peripheral Component Interconnect (PCI) bus, a PCI-Express (PCIe) bus, a serial advanced technology attachment (SATA) bus, a Video Electronics Standards Association local (VLB) bus, or another suitable bus or a combination of two or more of these. Bus 1012 may include one or more buses 1012, where appropriate. Although this disclosure describes and illustrates a particular bus, this disclosure contemplates any suitable bus or interconnect.

[0067] It should be further noted that, the terms "include", "comprise", or any other variant thereof are intended to cover a non-exclusive inclusion, so that a process, a method, a product or a device that includes a list of elements not only includes those elements but also includes other

elements which are not expressly listed, or further includes elements inherent to such process, method, product or device. Without more constraints, an element preceded by "includes a …" does not preclude the existence of additional identical elements in the process, method, product or device that includes the element.

[0068] The embodiments of this specification are described in a progressive way. For same or similar parts of the embodiments, mutual references may be made to the embodiments. Each embodiment focuses on a difference from other embodiments. Particularly, system embodiments are briefly described because they are basically similar to method embodiments. For related parts, references may be made to related descriptions in the method embodiments.

[0069] The foregoing descriptions are merely embodiments of this specification and are not intended to limit this specification. A person skilled in the art can make various changes and variations to this specification. Any modification, equivalent replacement, improvement, etc. made without departing from the spirit and principle of this specification shall fall within the scope of the claims in this application.

CLAIMS

What is claimed is:

1. A large model–based question-answering method, wherein the method comprises:

processing question information for a specified knowledge document based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model, so that the target large model outputs answer information based on the prompt information;

obtaining update data, determined based on the answer information, for the specified knowledge document; and

updating the knowledge base based on the update data.

2. The method according to claim 1, wherein the processing question information for a specified knowledge document based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model comprises:

performing semantic recognition on the question information to obtain a target word sequence formed by a semantic keyword;

performing term detection on the target word sequence;

if it is determined, based on the term detection, that the target word sequence comprises a preset specialized term, obtaining explanatory information corresponding to the specialized term, and adding the explanatory information to the target word sequence to update the target word sequence; and

determining the prompt information based on the updated target word sequence and the knowledge base.

3. The method according to claim 2, wherein the performing term detection on the target word sequence comprises:

obtaining a pre-constructed term list, wherein the term list comprises at least one specialized term and explanatory information corresponding to each specialized term; and

performing term detection on the target word sequence based on the term list, wherein

the obtaining explanatory information corresponding to the specialized term comprises obtaining the explanatory information corresponding to the specialized term from the term list.

4. The method according to claim 3, wherein the determining the prompt information based on the updated target word sequence and the knowledge base comprises:

searching, based on the updated target word sequence, the knowledge base for an associative word vector related to the semantic keyword comprised in the target word sequence;

determining partial document content in the specified knowledge document based on the associative word vector; and

generating the prompt information based on the partial document content.

5. The method according to claim 4, wherein the prompt information comprises information obtained after the question information is marked in a preset manner.

6. The method according to claim 1, wherein before processing the question information, the method further comprises:

performing risk detection on the question information; and

performing an operation of processing the question information when the question information does not comprise a sensitive word.

7. The method according to claim 1, wherein after the target large model outputs the answer information, the method further comprises:

performing risk detection on the answer information; and

outputting the answer information to a user when the answer information does not comprise a sensitive word.

8. The method according to claim 1, wherein the obtaining update data, determined based on the answer information, for the specified knowledge document comprises:

displaying the answer information to a user, and obtaining feedback information input by the user based on the answer information; and

outputting the feedback information in an operation and maintenance interface to obtain the update data for the specified knowledge document.

9. The method according to claim 1, wherein the updating the knowledge base based on the update data comprises:

obtaining a pre-constructed reference question-answer library;

processing a question in the reference question-answer library based on the knowledge base and the update data by using the target large model, to obtain a test answer;

comparing the test answer with an answer in the reference question-answer library; and

updating the knowledge base by using the update data if a result of the comparison indicates that a difference between the test answer and the answer in the reference question-answer library is less than or equal to a preset difference.

10. A computer system, comprising a memory, a processor, and a computer program that is stored in the memory and that is capable of running on the processor, wherein the processor executes the program to implement the following steps:

processing question information for a specified knowledge document based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model, so that the target large model outputs answer information based on the prompt information;

obtaining update data, determined based on the answer information, for the specified knowledge document; and

updating the knowledge base based on the update data.

11. The computer system according to claim 10, wherein the processing question information for a specified knowledge document based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model comprises:

performing semantic recognition on the question information to obtain a target word sequence formed by a semantic keyword;

performing term detection on the target word sequence;

if it is determined, based on the term detection, that the target word sequence comprises a preset specialized term, obtaining explanatory information corresponding to the specialized term, and adding the explanatory information to the target word sequence to update the target word sequence; and

determining the prompt information based on the updated target word sequence and the knowledge base.

12. The computer system according to claim 11, wherein the performing term detection on the target word sequence comprises:

obtaining a pre-constructed term list, wherein the term list comprises at least one specialized term and explanatory information corresponding to each specialized term; and

performing term detection on the target word sequence based on the term list, wherein

the obtaining explanatory information corresponding to the specialized term comprises obtaining the explanatory information corresponding to the specialized term from the term list.

13. The computer system according to claim 12, wherein the determining the prompt information based on the updated target word sequence and the knowledge base comprises:

searching, based on the updated target word sequence, the knowledge base for an associative word vector related to the semantic keyword comprised in the target word sequence;

determining partial document content in the specified knowledge document based on the associative word vector; and

generating the prompt information based on the partial document content.

14. The computer system according to claim 13, wherein the prompt information comprises information obtained after the question information is marked in a preset manner.

15. The computer system according to claim 10, wherein the processor executes the program to further implement the following steps:

performing risk detection on the question information before processing the question information; and

performing an operation of processing the question information when the question information does not comprise a sensitive word.

16. The computer system according to claim 10, wherein the processor executes the program to further implement the following steps:

performing risk detection on the answer information after the target large model outputs the answer information; and

outputting the answer information to a user when the answer information does not comprise a sensitive word.

17. The computer system according to claim 10, wherein the obtaining update data, determined based on the answer information, for the specified knowledge document comprises:

displaying the answer information to a user, and obtaining feedback information input by the user based on the answer information; and

outputting the feedback information in an operation and maintenance interface to obtain the update data for the specified knowledge document.

18. The computer system according to claim 10, wherein the updating the knowledge base based on the update data comprises:

obtaining a pre-constructed reference question-answer library;

processing a question in the reference question-answer library based on the knowledge base and the update data by using the target large model, to obtain a test answer;

comparing the test answer with an answer in the reference question-answer library; and

updating the knowledge base by using the update data if a result of the comparison indicates that a difference between the test answer and the answer in the reference question-answer library is less than or equal to a preset difference.

19. A large model–based question-answering apparatus, wherein the apparatus method comprises:

a processing module, configured to process question information for a specified knowledge document based on a knowledge base pre-constructed based on the specified knowledge document, to obtain prompt information to be transmitted to a target large model, so that the target large model outputs answer information based on the prompt information;

an obtaining module, configured to obtain update data, determined based on the answer information, for the specified knowledge document; and

an updating module, configured to update the knowledge base based on the update data.

20. A computer-readable storage medium, wherein the computer-readable storage medium stores a computer program, and when the computer program is executed on a computer, the computer is enabled to perform the method according to any one of claims 1 to 9.