

DESCRIPTION

Method and device for controlled modification of a generative artificial intelligence model, and associated usage method and device

The present invention relates to a method for controlled modification of a generative artificial intelligence model enhanced by retrieval, as well as a device for controlled modification of a generative artificial intelligence model enhanced by retrieval and an associated computer program.

It also relates to a method for using a modified generative artificial intelligence model enhanced by retrieval, and a device for using a modified generative artificial intelligence model enhanced by retrieval and an associated computer program.

The invention is in the field of generative artificial intelligence models, trained by machine learning to provide information from knowledge databases in response to a query formulated by a user or a client application.

Such generative artificial intelligence models have been developed recently and have been very successful in many application areas.

Among the generative artificial intelligence models, notably large language models, known by the acronym LLM from the English "Large Language Models," are known, which implement deep neural networks trained by machine learning on large amounts of text, for example, by self-supervised or semi-supervised learning.

Such large language models have notably been implemented in software that are conversational agents (or "chats"), for example, ChatGPT® (for "Chat Generative Pre-trained Transformer"), which allow, based on a query formulated by a user in natural language, in a given language, to obtain an adequate response. This technology has demonstrated very good performance on tasks such as text summarization, essay writing, computer programming, and more generally on relatively complex tasks.

Of course, such conversational agents implementing large language models are programmed to provide responses within a given legal framework (they are then said to be "aligned") and to avoid providing any information that could contribute to illegal action, for example, providing computer code to perform hacking. This constraint is called model framing or model alignment.

It has been observed that for certain queries, such large language models provide erroneous responses, also called hallucinations.

In order to address this problem, it has been proposed to use the generative artificial intelligence model enhanced by retrieval of contexts from a context base. This technology for improving generative artificial intelligence models is known by the acronym RAG for "*Retrieval Augmented Generation*," and consists of using the context database (or knowledge base) to enrich the formulated query with an adequate context, in order to limit the risk of an erroneous response. Thus, the reliability of the responses is increased.

However, it is difficult to ensure that for semantically similar queries, the selected context is always adequate. Furthermore, to ensure the framing of the model, in certain applications, it may be desirable to avoid using certain contexts to enrich a given query.

The purpose of the invention is to solve these problems, by proposing a controlled modification of a generative artificial intelligence model enhanced by retrieval, allowing better control of the operational framing of the model.

To this end, the invention has as its object a method for controlled modification of a generative artificial intelligence model enhanced by retrieval from a context base, the context base including a plurality of textual contexts, each textual context being transformed by applying a word embedding method into an associated context vector, the context vectors being stored in a vector base, the method being implemented by a processor of a programmable electronic device. This method includes steps of:

- acquiring a query in the form of a character string, and transforming said query, by said word embedding method of the query, into a query vector,
- selecting, depending on the query vector, a first context vector in said vector base, the first context vector being associated with an initial textual context of the context base,
- modifying the initial textual context into a modified textual context by adding a modification sequence to said initial textual context to obtain a modified context character string, then applying said word embedding method to the modified textual context to a second context vector and calculating a distance between the query vector and said second context vector, the modification sequence being determined depending on a distance objective between the query vector and the second context vector,
- storing the modified textual context in a modified context base and the associated second context vector in the vector base.

Advantageously, the proposed method allows to modify textual contexts, performed in such a way as to control the subsequent selection or non-selection of a given textual context to respond to semantically similar queries. Thus, the method allows control over the framing (or alignment) of the responses of the generative artificial intelligence model.

According to other advantageous aspects of the invention, the method for controlled modification of a generative artificial intelligence model enhanced by retrieval comprises

one or more of the following features, taken individually or according to any technically possible combination.

The modification of the initial textual context is implemented by applying an iterative optimization method.

5 The iterative optimization method is a gradient descent.

The distance objective consists of verifying that the distance between the query vector and said second context vector is less than a predetermined threshold.

10 The modification includes adding the modification sequence in the form of a prefix, and adding, between the modification sequence and the initial textual context, a predetermined separator fragment, or adding the modification sequence in the form of a suffix, and adding, between the initial textual context and the modification sequence, said predetermined separator fragment.

The step of selecting a first context vector in said vector base includes sub-steps of:

- selecting a predetermined number of context vectors closest to the query vector,
- 15 - providing, on a user interface, a set of initial textual contexts associated with the selected context vectors, ordered in ascending order of the distance between the query vector and the associated context vector,
- obtaining a selection result by a user of at least one textual context in said set of textual contexts, having a corresponding context vector, and an associated modification
- 20 indication, so as to indicate a convergence or divergence of said at least one selected context vector relative to the query vector,
- for each selected context vector, calculating the associated distance objective.

25 The invention also relates to a device for controlled modification of a generative artificial intelligence model enhanced by retrieval from a context base, the context base including a plurality of textual contexts, each textual context being transformed by applying a word embedding method to an associated context vector, the context vectors being stored in a vector base. This device includes a processor configured to implement:

- a module for acquiring a query in the form of a character string, and transforming said query by said word embedding method of the query into a query vector,
- 30 - a module for selecting, depending on the query vector, a first context vector in said vector base, the first context vector being associated with an initial textual context of the context base,
- a module for modifying the initial textual context into a modified textual context by adding a modification sequence to said initial textual context to obtain a modified context
- 35 character string, then a module (36) for applying said word embedding method to the modified textual context to a second context vector and calculating a distance between the

query vector and said second context vector, the modification sequence being determined by a module for determining the modification sequence, depending on a distance objective between the query vector and the second context vector,

- a module for storing the modified textual context in a modified context base and the associated second context vector in the vector base.

The invention also relates to a computer program including software instructions which, when executed by a computer, implement a method for controlled modification of a generative artificial intelligence model enhanced by retrieval as defined above.

According to another aspect, the invention also relates to a method for using a modified generative artificial intelligence model enhanced by retrieval by a controlled modification method such as briefly described above, including steps implemented by a processor of:

- acquiring a user query in the form of a character string, and transforming, by said word embedding method, the user query into a user query vector,

- selecting, in the vector base, a context vector closest to the user query vector, said selected context vector,

- extracting a textual context associated with the selected context vector, said selected textual context, and, if said selected textual context belongs to the modified context base, filtering the selected textual context to obtain the corresponding initial textual context,

- augmenting the user query by the corresponding initial textual context to obtain an augmented query,

- applying the generative artificial intelligence model with said augmented query to obtain a response to said user query.

According to one feature, the filtering comprises identifying a predetermined separator fragment, and subtracting a character string defining the modification sequence, located before or after said separator fragment, and subtracting said separator fragment to obtain the corresponding initial textual context.

According to another aspect, the invention relates to a device for using a modified generative artificial intelligence model enhanced by retrieval by such a modification device, the usage device including a processor configured to implement:

- a module for acquiring a user query in the form of a character string, and transforming by a word embedding method the user query into a user query vector,

- a module for selecting, in the vector base, a context vector closest to the user query vector, said selected context vector,

- a module for extracting a textual context associated with the selected context vector, said selected textual context, and, if said selected textual context belongs to the modified context base, a module for filtering the selected textual context to obtain the corresponding initial textual context,

5 - a module for augmenting the user query by the corresponding initial textual context to obtain an augmented query,

- a module for applying the generative artificial intelligence model with said augmented query to obtain a response to said user query.

10 The invention also relates to a computer program including software instructions which, when executed by a computer, implement a method for using a modified generative artificial intelligence model such as defined above.

The invention will become clearer upon reading the following description, given solely by way of non-limiting example, and made with reference to the drawings wherein:

15 Figure 1 schematically represents a system for using a modified generative artificial intelligence model enhanced by retrieval;

Figure 2 is a block diagram of the main functional blocks of one embodiment of a device for controlled modification of an enhanced generative artificial intelligence model;

Figure 3 is a flowchart of the main steps of a method for controlled modification of an enhanced generative artificial intelligence model according to one embodiment;

20 Figure 4 is a block diagram of the main functional blocks of one embodiment of a device for using a modified enhanced generative artificial intelligence model;

Figure 5 is a flowchart of the main steps of a method for using a modified enhanced generative artificial intelligence model according to one embodiment.

25 Figure 1 schematically represents a system 2 for using a modified generative artificial intelligence model enhanced by retrieval by a modification method such as proposed.

From a context base 4, enriched with a modified context base 6, a vector base 8 is obtained.

30 The modified context base 6 is obtained from the context base 4 by implementing a controlled modification device 20, such as described in detail below.

Each respective context base comprises textual contexts, a textual context being represented in the form of a character string. Such a character string includes intelligible words using a given dictionary, in a given language. For example, the considered character string forms a sentence in a given language.

35 Each textual context is transformed into a vector of predetermined size by a transformation method, also called word or semantic embedding, and more commonly

known as "sentence embedding" in English. These methods are artificial intelligence methods.

The "embedding" technology refers to the representation of words or text phrases by vectors of real numbers in a vector space.

The "embedding" technology is known in natural language processing, and more generally in the field of large language models. Any technology implementing a so-called "embedding model" trained by machine learning is applicable, for example, "text-embedding-ada-002" (OpenAI), nomic-embed, BGE (for "Beijing General Embedding"), UAE (for Universal Angle Embedding), MiniLM-L12-v2, etc.

By applying one of the known word embedding technologies, each textual context of the context base 4 and the modified context base 6 is transformed into a corresponding vector, stored in the vector base 8.

Subsequently, the term "word embedding method" will be used to refer to an "embedding" technology.

A user 10 provides a query P, for example, written in natural language, in the form of a character string.

For example, the query is provided via the input/output interface of the device or via a client application.

The query P is also transformed into a vector, called the query vector and denoted V_P , using the same word embedding method as the one used to form the vector base 8.

A context vector V_C is selected in the vector base 8. The selected context vector is the vector most similar to the query vector V_P from among all the vectors in the vector base 8, according to a chosen similarity metric.

The chosen similarity metric is, for example, the Euclidean distance.

Of course, other vector distance metrics are applicable.

The textual context CC corresponding to the chosen context vector, V_C , is then extracted from the context base 4 or the modified context base 6.

When the context is extracted from the modified context base 6, a corresponding initial textual context CC_i is obtained from the textual context CC.

The initial textual context CC_i is used to augment (or enrich) the query P and obtain an augmented query P^* which is then provided to the generative artificial intelligence model, for example, the LLM (for "Large Language Model") 12. The generative artificial intelligence model is then implemented to provide a response R.

The LLM 12 is a model trained by machine learning to provide information, in the form of the response R, following a query P formulated in written form, and applied, for example,

in conversational agent type software. In particular, the invention applies to any LLM provided in the form of a deep neural network ("deep learning").

Unlike classic applications of a generative artificial intelligence model enhanced by retrieval from a context base, it is proposed to use, in addition, a modified context base 6,
5 in order to ensure a predetermined alignment (or framing).

Below is described, on the one hand, one embodiment of a device and method for controlled modification of a generative artificial intelligence model enhanced by retrieval, allowing notably, to generate the modified context base 6 and the associated context vectors, and on the other hand, one embodiment of the device and method for using a
10 modified generative artificial intelligence model thus generated.

Figure 2 is a block diagram of a controlled modification device 20, allowing notably to generate the modified context base 6.

The device 20 is a programmable electronic device and includes, in one embodiment, an electronic memory unit 22, one or more processors 24, an input/output interface 26, and
15 a communication interface 28, these elements being configured to communicate with each other via a communication bus 25 internal to the device 20.

The input/output interface 26 includes, for example, a display screen and a character input device, for example, a keyboard, allowing a user to enter queries in written form to query the generative artificial intelligence model 12.

20 In one embodiment, the electronic memory 22 stores a context base 4, a modified context base 6, and a vector base 8, the vectors of the vector base 8 being calculated from the respective context bases 4, 6 by the chosen word embedding method.

According to alternatives, the context base 4 is stored in a memory external to the device 20, the device 20 being connected to the external memory, for example, through the
25 communication interface 28.

The generative artificial intelligence model 12 is also stored in the electronic memory unit 22 or accessible via the communication interface.

In some embodiments, the electronic memory 22 includes a mapping structure between vectors of the vector base 8 and contexts of the respective context bases 4, 6. For
30 example, the mapping structure is a table, associating by row a vector of the vector base 8 and corresponding contexts, for example, in different columns.

The calculation processor 24 of the device 20 is configured to execute:

- a module 30 for acquiring a query in the form of a character string, and transforming, by the word embedding method, the query into a query vector,

- a module 32 for selecting, depending on the query vector, a first context vector in the vector base 8, the first context vector being associated with an initial textual context of the context base;

- a module 34 for modifying the initial textual context into a modified textual context by adding a modification sequence to said initial textual context to obtain a modified context character string;

- a module 36 for applying said word embedding method to the modified textual context to a second context vector and calculating a distance between the query vector and said second context vector;

- a module 38 for determining the modification sequence, the modification sequence being determined, preferably iteratively, depending on a distance objective between the query vector and the second context vector,

- a module 40 for storing the modified textual context in the modified context base and the associated second context vector in the vector base 8.

In one embodiment, the modules 30, 32, 34, 36, 38, 40 are provided in the form of software instructions forming a computer program, which, when executed by a programmable electronic device, implements a method for controlled modification of a generative artificial intelligence model such as described.

In one alternative, not represented, the modules 30, 32, 34, 36, 38, 40 are each provided in the form of programmable logic components, such as FPGAs (*Field Programmable Gate Array*), microprocessors, GPGPU components (*General-purpose processing on graphics processing*), or dedicated integrated circuits, such as ASICs (*Application Specific Integrated Circuit*).

The computer program including software instructions is also able to be recorded on a non-transitory, computer-readable information recording medium. This computer-readable medium is, for example, a medium able to store electronic instructions and to be coupled to a bus of a computer system. For example, this medium is an optical disk, a magneto-optical disk, a ROM memory, a RAM memory, any type of non-volatile memory (for example, EPROM, EEPROM, FLASH, NVRAM), a magnetic card, or an optical card.

Figure 3 is a flowchart of the main steps of a method for controlled modification of a generative artificial intelligence model enhanced by retrieval from a context base according to one embodiment, implemented by the various modules 30, 32, 34, 36, 38, 40 of a controlled modification device 20 such as described above.

At the input of the method, the textual context base 4, which are the so-called initial textual contexts, to be used for query augmentation in the RAG scheme, the modified context base 6, which may initially be blank, and the vector base 8 including the vectors

corresponding to the contexts, and obtained by transforming a textual context into a vector by applying a chosen word embedding method, are provided.

A mapping between each context vector of the vector base and the corresponding textual context from which the context vector was calculated is also stored.

5 The method includes a step 50 of acquiring a query, also called a test query, in the form of a character string, and transforming by the chosen word embedding method, the query into a query vector V_{Pt} .

10 The method then includes a step 52 of selecting a first context vector V_{C1} in said vector base. The distance, according to the chosen distance metric, between the query vector V_{Pt} and the first context vector V_{C1} is called the first distance and denoted d_1 : $d_1 = \|V_{Pt} - V_{C1}\|$

The first context vector V_{C1} is associated with an initial textual context CT1 of the context base 4.

15 The method allows to modify the initial textual context CT1 into a modified textual context, the vector corresponding to the modified textual context, said second context vector, being distant from the query vector V_{Pt} by a second distance, d_2 , different from the first distance d_1 .

In the embodiment of Figure 3, the step 52 of selecting a first context vector comprises several sub-steps, of which, a selection 54 of a predetermined number of context vectors closest to the query vector V_{Pt} .

20 The predetermined number is an integer K greater than or equal to 1, for example, included between 1 and 50.

25 The selection is made, for example, by calculating the distances between the query vector V_{Pt} and each context vector of the vector base, according to the chosen distance metric; then an ordering of the context vectors according to increasing rank with the calculated distance.

The selection step 52 also comprises a sub-step 56 of providing, on a user interface, the K initial textual contexts associated with the selected context vectors, ordered according to their respective rank.

30 For example, considering K=3, the three context vectors V_{C-r1} , V_{C-r2} , and V_{C-r3} are selected, with respective ranks 1, 2, and 3, the context vector V_{C-r1} being the closest to the query vector, the context vector V_{C-r2} the second closest, and the context vector V_{C-r3} the third closest.

The initial textual contexts CTr1, CTr2, CTr3 are then extracted and provided for display.

35 A user can then view the K, for example, the three in this example, closest initial contexts.

In the classic operation, a generative artificial intelligence model enhanced by retrieval would use the initial context CTr1 associated with the context vector closest to the query vector to calculate an augmented query.

However, for reasons of preference alignment related to the application of the generative artificial intelligence model, the user can choose another ordering of the contexts, for example, can choose the initial textual context CTr2 as the initial context to be used, to calculate an augmented query.

The method then includes a sub-step 58 of obtaining a selection by the user of a textual context and an indication of modification of the associated rank.

The modification indication is, for example, an annotation '+'/'-' or '+1'/'-1' which aims to indicate a convergence or divergence of the selected textual context, and therefore of the selected context vector relative to the query vector.

For example, the modification indication is a modified rank. This allows the user to indicate which initial context should be chosen as the closest textual context.

This allows active learning of model alignment.

Alternatively or in addition, this also allows the user to ensure that one of the initial textual contexts will not be used to form an augmented query, and therefore to distance the corresponding context vector from the query vector.

According to a variant, the selection of a textual context and an indication of modification of the associated rank is not provided by the user, but by an automatic process called reranking.

At the output of step 58, at least one first selected context vector is obtained, corresponding to an initial textual context to be modified.

Furthermore, for each first selected context vector, an associated distance objective is calculated at step 60, depending on the modification indication.

Of course, if the user selected several textual contexts to modify, the following steps are applied to each of the first context vectors associated with the selected textual contexts.

For example, if the modification indication is an indication of convergence, the modification objective is to modify the associated textual context so that a second context vector calculated from the modified textual context is closer to the query vector than the first context vector.

For example, a distance objective may consist of minimizing the distance between the second context vector and the query vector, or in other words, that the second context vector is ranked as the closest query vector according to the applied distance metric.

The method includes, after the selection 52 of a first context vector, a step 62 of modifying the initial textual context into a modified textual context by adding a modification sequence.

The initial textual context is formed by an initial context character string.

5 The modification sequence is, for example, a modification character string, which is added to the initial character string to form a modified context character string.

Preferably, the modification sequence is added in the form of a prefix or a suffix to the initial context character string.

10 In one embodiment, the modification includes adding the modification sequence in the form of a prefix, and adding, between the modification sequence and the initial textual context, a predetermined separator fragment.

In another embodiment, the modification includes adding the modification sequence in the form of a suffix, and adding, between the initial textual context and the modification sequence, the predetermined separator fragment.

15 The predetermined separator fragment is, for example, a grouping of characters that is not part of the words of a natural language word dictionary used. Thus, the separator fragment is recognizable by automated processing.

The method then comprises a transformation 64 by the word embedding method of the modified context character string to obtain a second context vector V_{C2} , and a calculation
20 65 of the distance according to the chosen distance metric, between the query vector V_{Pt} and the second context vector V_{C2} , called the second distance and denoted d_2 : $d_2 = \|V_{Pt} - V_{C2}\|$

It is then verified at the verification step 66 if the second distance meets the distance objective.

25 For example, in the case where the objective is for the second context vector to be closer to the query vector, the distance objective is to make the distance d_2 as small as possible.

According to another example, the distance objective consists of minimizing the distance between the second context vector and the query vector, that is, making the
30 second distance d_2 the smallest from among all the distances between a context vector and the query vector.

For example, the distance objective can be formulated as: $d_2 < \varepsilon$, with ε the predetermined threshold.

35 In one embodiment, the threshold ε is a percentage of the first distance d_1 , for example, between 5% and 15%.

If the distance objective is not met, the steps 62 to 66 are iterated until the distance objective is met.

For example, the modification sequence is determined by an iterative optimization method, for example, a gradient descent method.

5 If the distance objective is met, then step 66 is followed by a step 68 of storing the modified textual context in a modified context base and the associated second context vector in the vector base.

In one embodiment, a correspondence between the modified context vector and the initial textual context is stored, for example, in a correspondence table.

10 According to one alternative, the selection 52 of a first context vector associated with a textual context to be modified is performed depending on a performance criterion: if the textual context associated with the context vector closest to the query vector provides an erroneous response (or hallucination) when the generative artificial intelligence model is applied, then a textual context that allows to provide a correct response is selected, with a
15 convergence objective.

Figure 4 is a block diagram of a device 70 for using a modified generative artificial intelligence model enhanced by retrieval by a method as described above.

The device 70 is a programmable electronic device.

20 According to the embodiments, the programmable electronic device may be the same as the programmable electronic device 20, or clearly distinct from the programmable electronic device 20.

The device 70 includes, in one embodiment, an electronic memory unit 72, one or more processors 74, an input/output interface 76, and a communication interface 78, these elements being configured to communicate with each other via a communication bus 75
25 internal to the device 70.

In one embodiment, the electronic memory 72 stores the context base 4, the modified context base 6 obtained by the modification method described above, and the vector base 8, the vectors of the vector base being calculated from the respective context bases 4, 6 by the chosen word embedding method. In addition, the generative artificial intelligence model
30 12, trained by machine learning, is stored.

According to alternatives, the context base 4 and the modified context base 6 are stored in a memory external to the device 70, the device 70 being connected to the external memory, for example, through the communication interface 78.

The calculation processor 74 of the device 70 is configured to execute:

35 - a module 80 for acquiring a user query in the form of a character string, and transforming by the word embedding method of the query into a user query vector,

- a module 82 for selecting, in the vector base 8, a context vector closest to the user query vector, said selected context vector;

- a module 84 for extracting a textual context associated with the selected context vector, said selected textual context;

5 - a module 86 for filtering the selected textual context to obtain the corresponding initial textual context,

- a module 88 for augmenting the user query by the corresponding initial textual context to obtain an augmented query and

10 - a module 90 for applying the generative artificial intelligence model with said augmented query to obtain a response to said user query.

In one embodiment, the modules 80, 82, 84, 86, 88, 90 are each provided in the form of software instructions forming a computer program, which, when executed by a programmable electronic device, implements a method for using a modified generative artificial intelligence model such as described.

15 In one alternative, not represented, the modules 80, 82, 84, 86, 88, 90 are each provided in the form of programmable logic components, such as FPGAs (*Field Programmable Gate Array*), microprocessors, GPGPU components (*General-purpose processing on graphics processing*), or dedicated integrated circuits, such as ASICs (*Application Specific Integrated Circuit*).

20 The computer program including software instructions is also able to be recorded on a non-transitory, computer-readable information recording medium. This computer-readable medium is, for example, a medium capable of storing electronic instructions and being coupled to a bus of a computer system. For example, this medium is an optical disk, a magneto-optical disk, a ROM memory, a RAM memory, any type of non-volatile memory
25 (for example, EPROM, EEPROM, FLASH, NVRAM), a magnetic card, or an optical card.

Figure 5 is a flowchart of the main steps of a method for using a modified generative artificial intelligence model enhanced by retrieval according to one embodiment, implemented by the various modules 80, 82, 84, 86, 88, 90 of a controlled modification device 70 as described above.

30 The method includes a step 92 of acquiring a user query in the form of a character string, and transforming by said word embedding method, the user query into a user query vector, V_{Pu} .

The method then includes a step 94 of selecting, in the vector base, a context vector VC_p closest to the user query vector, said selected context vector.

The selected context vector is the vector most similar to the user query vector V_{Pu} from among all the vectors in the vector base 8, according to a chosen similarity metric. The chosen similarity metric is, for example, the Euclidean distance.

Of course, other vector distance metrics are applicable.

5 The method then comprises a step 96 of extracting a textual context associated with the selected context vector, said selected textual context.

It is then verified at the verification step 98 if the selected textual context belongs to the modified context base, in other words, if the selected textual context is derived from an initial textual context modified by adding a modification sequence during the implementation
10 of the modification method described above.

In case of a positive response at the verification step 98, the method includes a step 100 of filtering the selected textual context to obtain the corresponding initial textual context. The filtering 100 allows to obtain the initial textual context without the added modification sequence.

15 In one embodiment, the filtering 100 includes subtracting the modification sequence from the modified textual context.

In the case where a predetermined separator fragment has been introduced, after or before the modification sequence according to whether the modification sequence was added as a prefix or suffix, the filtering 100 includes identifying the separator fragment and
20 subtracting all the characters forming the modification sequence located before or after the identified separator fragment. The separator fragment is also removed during the filtering step 100.

Advantageously, the separator fragment is recognizable by automatic processing, being a known grouping of characters, previously stored, that is not part of the words of a
25 natural language word dictionary used. Thus, the separator fragment is recognizable, and it is not necessary to know in advance whether the modification sequence was added as a prefix or suffix.

Alternatively, if a correspondence between the modified context vectors and initial textual contexts has been stored, for example, in a correspondence table, at the filtering
30 step 100, the filtering 100 consists of accessing the corresponding initial textual context.

The method includes, after the filtering step 100 or after the verification step 98 in case of a negative response to the verification step, a step 102 of augmenting the user query by the corresponding initial textual context to obtain an augmented query and a step
35 104 of applying the generative artificial intelligence model with said augmented query to obtain a response to the user query.

Optionally, the response is displayed (display step 106) on a display of the programmable electronic device 70 and/or provided to the client application that transmitted the query (step 108).

Advantageously, the invention allows to control the alignment in the use of a
5 generative artificial intelligence model enhanced by retrieval from a context base.

CLAIMS

1. A method for controlled modification of a generative artificial intelligence model enhanced by retrieval from a context base, the context base including a plurality of textual contexts, each textual context being transformed by applying a word embedding method into an associated context vector, the context vectors being stored in a vector base, the method being implemented by a processor of a programmable electronic device, and being characterized in that it includes steps of:

- acquiring (50) a query in the form of a character string, and transforming said query by said word embedding method of the query into a query vector,
- selecting (52), depending on the query vector, a first context vector in said vector base, the first context vector being associated with an initial textual context of the context base,
- modifying (62) the initial textual context into a modified textual context by adding a modification sequence to said initial textual context to obtain a modified context character string, then applying (64) said word embedding method to the modified textual context in a second context vector and calculating (65) a distance between the query vector and said second context vector, the modification sequence being determined depending on a distance objective between the query vector and the second context vector,
- storing (68) the modified textual context in a modified context base and the associated second context vector in the vector base.

2. The method according to claim 1, wherein the modification (62) of the initial textual context is implemented by applying an iterative optimization method.

3. The method according to claim 2 wherein the iterative optimization method is a gradient descent.

4. The method according to one of claims 1 to 3, wherein the distance objective consists of verifying that the distance between the query vector and said second context vector is less than a predetermined threshold.

5. The method according to one of claims 1 to 4, wherein the modification (62) includes adding the modification sequence in the form of a prefix, and adding, between the modification sequence and the initial textual context, a predetermined separator fragment,

or adding the modification sequence in the form of a suffix, and adding, between the initial textual context and the modification sequence, said predetermined separator fragment.

6. The method according to any one of claims 1 to 5, wherein the step of
5 selecting (52) a first context vector in said vector base includes sub-steps of:

- selecting (54) a predetermined number of context vectors closest to the query vector,
- providing (56), on a user interface, a set of initial textual contexts associated with the selected context vectors, ordered in ascending order of the distance between
10 the query vector and the associated context vector,
- obtaining a selection result (58) by a user of at least one textual context in said set of textual contexts, having a corresponding context vector, and an associated modification indication, so as to indicate a convergence or divergence of said at least one selected context vector relative to the query vector,
- 15 - for each selected context vector, calculating (60) the associated distance objective.

7. A computer program including software instructions which, when executed by a programmable electronic device, implement a method for controlled modification of a generative artificial intelligence model according to claims 1 to 6.
20

8. A method for using a modified generative artificial intelligence model enhanced by retrieval by a modification method according to claims 1 to 6, the method including steps implemented by a processor of:

- acquiring (92) a user query in the form of a character string, and transforming, by
25 said word embedding method, the user query into a user query vector,
- selecting (94), in the vector base, a context vector closest to the user query vector, said selected context vector,
- extracting (96) a textual context associated with the selected context vector, said selected textual context, and, if said selected textual context belongs to the modified
30 context base, filtering (100) the selected textual context to obtain the corresponding initial textual context,
- augmenting (102) the user query by the corresponding initial textual context to obtain an augmented query,
- applying (104) the generative artificial intelligence model with said augmented
35 query to obtain a response to said user query.

9. The method according to claim 8, wherein the filtering (100) comprises identifying a predetermined separator fragment, and subtracting a character string defining the modification sequence, located before or after said separator fragment, and subtracting said separator fragment to obtain the corresponding initial textual context.

5

10. A computer program including software instructions which, when executed by a programmable electronic device, implement a method for using a modified generative artificial intelligence model according to claims 8 to 9.

10

11. A device for controlled modification of a generative artificial intelligence model enhanced by retrieval from a context base, the context base including a plurality of textual contexts, each textual context being transformed by applying a word embedding method into an associated context vector, the context vectors being stored in a vector base, the device including a processor characterized in that it is configured to implement:

15

- a module for acquiring (30) a query in the form of a character string, and transforming said query, by said word embedding method of the query, into a query vector,

20

- a module for selecting (32), depending on the query vector, a first context vector in said vector base, the first context vector being associated with an initial textual context of the context base,

25

- a module for modifying (32) the initial textual context into a modified textual context by adding a modification sequence to said initial textual context to obtain a modified context character string, then a module (36) for applying said word embedding method to the modified textual context to a second context vector and calculating a distance between the query vector and said second context vector, the modification sequence being determined by a module (38) for determining the modification sequence, depending on a distance objective between the query vector and the second context vector,

30

- a module for storing (40) the modified textual context in a modified context base and the associated second context vector in the vector base.

12. The device for using a modified generative artificial intelligence model enhanced by retrieval by a modification device according to claim 11, the device including a processor configured to implement:

- a module (80) for acquiring a user query in the form of a character string, and transforming, by a word embedding method of the user query, into a user query vector,
- 5 - a module (82) for selecting, in the vector base, a context vector closest to the user query vector, said selected context vector,
- a module (84) for extracting a textual context associated with the selected context vector, said selected textual context, and, if said selected textual context belongs to the modified context base, a module for filtering (86) the selected textual context to obtain the corresponding initial textual context,
- 10 - a module (88) for augmenting the user query by the corresponding initial textual context to obtain an augmented query,
- a module (90) for applying the generative artificial intelligence model with said augmented query to obtain a response to said user query.