

A first approach to the generation of linguistic summaries from glucose sensors using GPT-4. *

Carmen Martinez-Cruz^{1[0000-0002-8117-0647]}, Juan F. Gaitan
Guerrero^{1[0009-0007-6872-1401]}, Jose L. Lopez Ruiz^{1[0000-0003-2583-8638]},
Antonio J. Rueda^{1[0000-0001-7692-454x]}, and M. Espinilla^{1[0000-0003-1118-7782]}

University of Jaen, Department of Computer Science, Jaen, Spain

Abstract. The use of activity monitoring sensors on users with some type of disease or dependence is very useful for health technicians, for family members or for the user himself. The knowledge of these values in real time allows alerting of a possible crisis or starting correcting actions to prevent a serious health problem. For this reason, many proposals have been made to summarize in words the huge amount of measures taken by these sensors in order to highlight only what is really important for the end user, family or medical staff. The emergence of new text generation tools based on Artificial Intelligence (AI), such as the latest GPT-4, is having a major impact in the healthcare field. In this article we analyze how the latest version of ChatGPT, allows the generation of linguistic summaries in natural language from glucose sensor measurements. We also learn how to ask the right questions to obtain the type of output adapted to the user, whether or not it is necessary to perform some kind of preprocessing on the data to be analyzed and what are the strengths and drawbacks of this technology.

Keywords: Linguistic summaries · Computing with words · Glucose level · Sensors · ChatGPT · Artificial Intelligence · GPT-4 · Large language models

1 Introduction

The Generative Pre-trained Transformer (GPT) family of Large Language Models, such ChatGPT o Bard, have revolutionized the current society and are going to set a new milestone in the technological race of the last decades[7]. These tools are able to produce themselves as humans providing answers to any question, in a fluent, agile, coherent, and natural way. In addition, they can respond to different roles, in different styles, whether conversational, written, requests, scientific texts, etc. and even in different spoken languages or programming languages.

* This work has been partially supported by the Government of Spain through the projects RTI2018-098979-A-I00 MCIN/ AEI/10.13039/501100011033/, ERDF “A way to make Europe”, B-TIC-744-UGR20 ADIM: Accesibilidad de Datos para Investigación Médica of the Junta de Andalucía and the University of Jaén under Action 1 with reference EL.TIC1.2021.

Since the ChatGPT chatbot was released last 2022, it has been used in numerous applications with amazing results. In the healthcare field, its use is also being evaluated, especially with the new GPT-4 engine released in 2023, which outperforms ChatGPT-3.5 on all the tasks and is also capable of analyzing images [1].

In recent decades, there has been a dramatic increase in the number of devices for monitoring subjects to control and even prevent their degree of well-being. These sensors generate such a large amount of information that it is very difficult to process by healthcare personnel, end users or their relatives, either because of the large volume of data or because they lack the technical training needed to understand it correctly. The generation of linguistic summaries of Time Series (TS) has provided a solution to these problems since it represents in text form the relevant information adapted to the needs of the end user. The generation of these summaries requires the supervision and knowledge of an expert who specifies the quality measures that the generated summary must meet.

In this proposal, an analysis of the performance of ChatGPT with GPT-4 in the generation of linguistic summaries of TS of blood glucose measurements is carried out. These measurements are taken through a monitoring device that controls glucose every 5 minutes. The values obtained are provided to this Artificial Intelligence (AI) and its conclusions are analyzed and different alternatives in the processing of the series are proposed to try to obtain the best possible output.

2 Chatgpt

ChatGPT is an artificial intelligence chatbot developed by OpenAI [1, 25, 20]. ChatGPT, using GPT-4 model, excels at creating fluent and natural texts, even in academic scenarios, where the incorporation of citations and references is quite acceptable. However, the bibliographic references and other sources of data, depending on the version used may be outdated, incorrect or invented [3]. Another problem of this AI is the generation of hallucinations [16], understood as the generation of seemingly realistic sensory experiences that do not correspond to any real-world input. Moreover, it exhibits pervasive biases making the use of this technology controversial in different areas. It also can not analyze audio or video or generate arithmetic errors that a simple calculator would never have [12]. Despite these problems, it has been tested in many scenarios with impressive results. In academics, GPT-4 achieves top scores in several exams for accessing to universities or certifications. It is able to write automatically various types of documents, including reports, summaries and even book reviews. In medicine it is being proved that this AI is able to provide assistance to the practitioners in many areas [7].

3 Generation of Linguistic Summaries

The process of generating linguistic summaries from Time Series, formally called Generation of Linguistic Descriptions of Time Series (GLiDTS), has been largely studied in the literature (see compilations of [18], [21], [17], [6]). Its main objective is to summarize raw data (mostly obtained from sensors) in a set of words, highlighting the most relevant phenomena identified in the TS, and adapting the language to the end user. Just as it would happen if it were a human interpreting these data. The purpose of the GLiDTS is very versatile, from the possibility of transmitting these data through text messages or to be audible, as the possibility of interpretation of technical data by users who have not expertness in the field of application of these data. There are many applications of these techniques in the literature, some examples are those related with summarizing weather reports using Natural Language Processing (NLP) as we can see in the works of Reiter et al. [22] and Ramos-Soto et al. [21]. Trivino et al. in [10], [4], [23], [24] generate summaries of streaming data in different application areas such as the energy consumption, gait analysis, etc. Banaee et al. [5] propose a partial trend detection algorithm to describe particular changes of health parameters in physiological TS data. Marin et al. in [9], [8] developed a mechanism that describes TS where time is expressed in different granularities. Finally, Martinez-Cruz et al. [19] presented a way of describing TS using the patterns detected in the TS, at different levels of abstraction. However, these techniques are complex as described in [18] since a knowledge model should be developed in order to obtain the most suitable summary according to the final user requirements. Moreover, the final text description must accomplish a set of quality standards specified by the domain experts. In this proposal, we try to avoid these steps by seeking the knowledge provided by the AI and, in particular, the one recently released, GPT-4.

4 Linguistic Summaries using ChatGPT

The process of generating linguistic summaries from TS or any other data source through the use of chatbot is a subjective process. Not only the process of generating summaries is opaque because of the nature of the technology, but the handling of natural language itself is subjective as well. For this reason, a methodology is presented here to allow us to analyze the results as objectively as possible. The proposal also describes different improvement processes that can be applied to increase the quality of the final text.

4.1 Methodology

The methodology used in this proposal consists in six phases illustrated in Figure 1 and described below:

Step 1. Data acquisition. A device embedded under the skin collects data that are transmitted to a database.

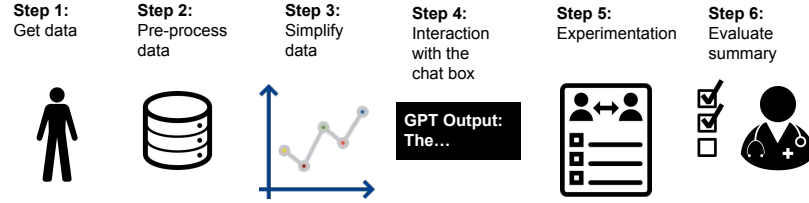


Fig. 1: System methodology.

Step 2. Preprocessing data. Data are processed to avoid inaccuracies generated by the sensors themselves, such as transmission errors or noise, empty periods of data, etc.

Step 3. Data simplification. This process allows highlighting the most relevant values of the TS while reducing the amount of data to be processed.

Step 4. Interaction with the chatbot. It consists of providing the chatbot with a textual description of the simplified TS and some prompting in the appropriate manner to obtain the best possible output.

Step 5. Experimentation. Analysis of the output obtained from a set of TS.

Step 6. Evaluate output. Evaluation of the linguistic summaries obtained by the chatbot.

4.2 Description of the problem

Data have to be acquired, processed and in some cases, preprocessed (steps 1 and 2). In this proposal we have used data obtained through an invasive glucose sensor connected via NFC (Near-Field Communication) and BLE (Bluetooth Low Energy). This sensor gets the interstitial fluid data, processes it and sends the samples continuously (one sample every five minutes) to a server running a MongoDB database. This data is accessible through a web application. In this proposal, glucose data are collected and preprocessed to fill spaces with no data below a threshold set at 30 minutes, to avoid any error caused by temporary connection problems between the sensor and the server. In this case, a moving average operation is performed on the data around the empty spaces, with a window of size 5.

4.3 Simplification process

When a Time Series contains many points, it is advisable to reduce this number of points so that its processing is simplified, in terms of efficiency and complexity. For example, the time series used in this experimentation are 288 points (with

some exceptions e.g. empty periods), resulting from monitoring a user's glucose every 5 minutes over the course of a day. This number may be too large to be processed by an artificial intelligence with satisfactory results. To simplify this number, a TS segmentation algorithm has been chosen using a technique similar to that performed by humans when observing a graph, where the most relevant points (peaks or valleys) are highlighted, and the rest are omitted. For this purpose, the Douglas-Peucker method has been used, whose algorithm can be seen in [11]. There are other segmentation methods, such as the Breakpoint Detection Algorithms [2], the Minimum Description Length [13], Moving Average Algorithm or those indicated in [14] [15]. However, it has been considered that geometric techniques are very useful in these cases, as discussed in [19].

This method filters those points whose distance is less than a given threshold (ϵ), so that only the most representative points of the series remain in a natural way. For this experiment, three thresholds have been used: $\epsilon = 0$ i.e. no threshold, $\epsilon = 0.2$ that generates a mild segmentation, and $\epsilon = 0.4$, which develops a more drastic segmentation. An example of this process is illustrated in Figure 2 from a TS of glucose data recorded on a person on December 31st, 2022.

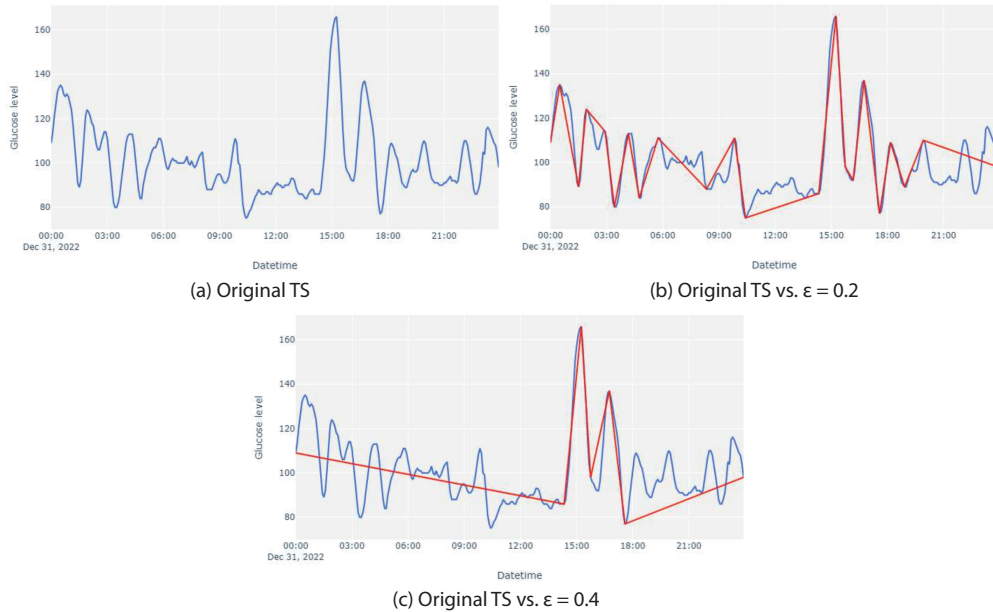


Fig. 2: TS Simplification (December 31st, 2022).

5 Experimentation

The goal of this experimentation is to obtain a textual description from a TS of sufficient quality to help the end user to understand the evolution of the measure, either with or without access to the graph. Moreover, the output must be short and highlight the most relevant events, which in the context of glucose, are related to relevant sugar highs or lows. To do so, in this experimentation we analyze:

- The choice of words as the prompt of the chatbot.
- The number of points included in the TS for obtaining the best description.

The interaction with ChatGPT (using GPT-4) consists in the creation of new conversations, where the AI uses a contextual memory to store the relevant details it needs to improve its answers. For this reason, we can perform multiple data insertions in the same conversation to obtain the desired result. Also, any response generated by this technology can be rephrased, so that the AI generates another output with the same purpose but using different words. However, the optimization of resources to obtain an answer must be taken into account, and performing a large number of interactions with the AI or generating different phrases for the same query is time and effort consuming.

We have try the following prompts once provided the TS¹ to the chatbot:

- Make a short linguistic summary of the following TS of glucose.
- Make a brief summary of the previous TS of glucose levels stressing the trends, variability and most relevant features.
- Generate a brief summary of the previous TS of glucose measures, highlighting the most relevant features.

The results obtained using any of these prompts are similar, highlighting the most representative elements of the TS, but varying the length and quality of the summary depending on the number of points entered. The prompt chosen for this experimentation is as follows:

`Make a summary of this Time Series in less than 100 words with the most relevant information.`

The output obtained from the AI has been analyzed for three sets of data: i) The original time series (described in Section 4.3) that is illustrated with an example in Figure 2 (a), ii) A “*medium*” simplification used with an $\epsilon = 0.2$, where the number of points obtained keeps the general trend of the graph, as can be seen in Illustration 2 (b), and iii) A “*strong*” simplification generated with $\epsilon = 0.4$, which gives us a very segmented TS where only the most significant points are maintained, as shown in Illustration 2 (c). The GPT-4 responses to those inputs are shown in Table 1.

¹ Although GPT-4 allows the analysis of figures, this is still a beta feature and therefore has not been used in our experimentation.

Table 1: Output of GPT-4 for a TS recorded on December 31st, 2022

Original TS($\epsilon = 0$). Number of points: 282.
<i>The glucose level time series for one person throughout the day shows a gradual decrease from 135.0 to 84.0 in the first few hours, reaching a minimum of 75.0. From there, the glucose levels gradually increase, peaking at 166.0 before stabilizing around 90.0-110.0. The most notable patterns include a decrease during the early morning hours, a rise during the afternoon, and a subsequent decrease during the evening. Overall, the glucose levels exhibit fluctuations within a range, with occasional spikes and dips throughout the day.</i>
“Medium” simplification ($\epsilon = 0.2$). Number of points: 22.
<i>The provided time series represents the glucose levels of an individual over a day. The range of glucose levels varies from 75.0 to 166.0 units. The average glucose level appears to be around 105.0 units. There are fluctuations throughout the day, with some notable spikes and drops. The lowest glucose reading was 75.0 units, while the highest was 166.0 units. It’s important to note that this summary is based on the limited information provided and does not account for potential patterns or underlying factors that may influence glucose levels. Consulting a medical professional for a thorough analysis is advised.</i>
“Strong” simplification ($\epsilon = 0.4$). Number of points: 7.
<i>The provided time series data represents a person’s glucose levels throughout the day. The initial measurement at midnight is 109.0, and it gradually decreases to 86.0 by 14:20. However, there is a significant spike at 15:15, with the glucose level reaching 166.0. Subsequently, it drops to 98.0 at 15:45 and then increases to 137.0 at 16:45. A sudden decrease is observed at 17:35 to 77.0. Towards the end of the day, at 23:55, the glucose level is recorded as 98.0 again. The data showcases fluctuations in glucose levels, with notable variations at specific time points.</i>

6 Evaluation

To assess the performance of this technology based on its outcomes, various criteria have been established to compare how effectively the AI system aligns with human-expected statements.

- C1 Highlight of relevant values (peaks, valleys, sharp increases or decreases, etc).
- C2 Highlight of more complex structures involving sets of peaks or valleys, or significant fluctuations.
- C3 Summarize the value of the majority of registered instances through statistical values.
- C4 Summarize the value of the majority of relevant registered instances.
- C5 Omit not significant measures (i.e.: normal glucose levels).
- C6 Indicate the day time a relevant situation is registered (hour or moment of the day)
- C7 Describe TS progressively in time.
- C8 Detect no-data-collection-intervals along the TS.
- C9 Describe the TS values in a semantic manner (i.e.: high glucose levels).
- C10 Include vague recommendations or superfluous information.

- C11 Include incorrect statements.
 C12 Omit relevant information.
 C13 Describe TS briefly.
 C14 Describe TS redundantly.
 C15 Warn about the need of more input data for the description.

For this experimentation, 5 days of data have been tested, i.e. a total of 15 Time Series with different amount of data. The results obtained and their analysis are discussed below. The scale defined in Figure 3 refers to the percentage of the TS for which the AI satisfies a certain criterion.



Fig. 3: GPT-4 performance scale.

Table 2: Checklist to validate the ChatGPT output.

Criterion	$\epsilon = 0$	$\epsilon = 0.2$	$\epsilon = 0.4$
C1	Always	Always	Always
C2	Never	Never	Never
C3	Rarely	Almost always	Rarely
C4	Never	Never	Never
C5	Rarely	Rarely	Never
C6	Always	Almost always	Always
C7	Always	Sometimes	Almost always
C8	Never	Never	Never
C9	Rarely	Rarely	Never
C10	Almost always	Almost always	Almost always
C11	Almost always	Rarely	Rarely
C12	Always	Almost always	Never
C13	Almost always	Always	Always
C14	Rarely	Almost always	Rarely
C15	Never	Sometimes	Never

The generated results allow the following conclusions to be drawn:

- Using the total number of glucose measures collected as input data leads the AI to generate incorrect sentences, as it tries to describe a huge amount of data in a few words.
- The generated output from a segmented TS with a “*strong*” simplification includes a more progressive linguistic description over time, due to the low number of points involved.

- All the linguistic descriptions generated by GPT-4 include values and times of the day which are quite accurate. It must be noticed that non-significant glucose measures are not filtered, i.e. normal glucose levels or irrelevant fluctuations which do not contribute to a transcendental knowledge.
- GPT-4 is unable to detect intervals of no-data, assuming a complete collection of registered instances in the input data set.
- The different outputs provided by GPT-4 contain trivial sentences or recommendations which move away from the main purpose of describing a graphical representation of a TS.
- GPT-4 randomly generates linguistic summaries following a correct time sequence but tends to describe the TS in a "discrete way" by jumping between relevant events. Therefore it does not provide a proper description of the trend and evolution of the TS over time.

7 Conclusion

The generation of linguistic summaries from time series facilitates the task of interpreting data obtained, for example from IoT devices, for family members or non-experts. Specifically, in the context of glucose, given the nature of the data, it is advisable to obtain only the relevant information, peaks or valleys with high or low values, and when they have taken place, avoiding redundancies or data that only result in extending the size of the final text. The use of ChatGPT to perform this type of summaries has shown us that the data output is totally dependent on the number of TS measures. If a large number of data is provided, ChatGPT will not return a correct summary, since it omits relevant information and performs a description that does not follow a temporal sequence. On the other hand, a TS with few points does not generate a good result either, since instead of an actual summary, it generates a straightforward text transcription of the TS values. A medium number of points is the most advisable option. However, it should be noted that in general, ChatGPT returns superfluous and not very relevant information, such as tips or statements that only lengthens a description that is expected to be short.

Also, the lack of open source code for GPT, in any of its versions, makes it difficult to know how to introduce data in this platform in order to obtain good results. However, it is possible to train the tool in certain areas of knowledge, so that a large data bank would be needed to obtain the desired results, in the form that the end user needs.

In any case, the results provided by this tool are impressive, and we believe that it is only a matter of time before ChatGPT or any LLM-based chatbot is used as a standard method for the generation of quality user-tailored summaries to the user's needs in a variety of different contexts.

References

1. Gpt-4 technical report (2023), <https://arxiv.org/pdf/2303.08774.pdf>

2. Ali, A., Aggarwal, J.: Segmentation and recognition of continuous human activity. In: Proceedings IEEE Workshop on Detection and Recognition of Events in Video. pp. 28–35 (2001). <https://doi.org/10.1109/EVENT.2001.938863>
3. Alkaissi, H., Mcfarlane, S.: Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus* **15** (02 2023). <https://doi.org/10.7759/cureus.35179>
4. Alvarez-Alvarez, A., Triviño, G.: Linguistic description of the human gait quality. *Eng. Appl. of AI* **26**(1), 13–23 (2013)
5. Banaee, H., Ahmed, M.U., Loutfi, A.: A framework for automatic text generation of trends in physiological time series data. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics. pp. 3876–3881 (Oct 2013)
6. Boran, F.E., Akay, D., Yager, R.R.: An overview of methods for linguistic summarization with fuzzy sets. *Expert Syst. Appl.* **61**, 356–377 (2016)
7. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: Early experiments with gpt-4 (2023)
8. Castillo-Ortega, R., Marín, N., Sánchez, D.: A fuzzy approach to the linguistic summarization of time series. *Journal of Multiple-Valued Logic and Soft Computing* **17**, 157–182 (2011)
9. Castillo-Ortega, R., Marín, N., Sánchez, D.: Linguistic query answering on data cubes with time dimension. *International Journal of Intelligent Systems (IJIS)* **26**(10), 1002–1021 (2011)
10. Conde-Clemente, P., Alonso, J.M., Triviño, G.: Toward automatic generation of linguistic advice for saving energy at home. *Soft Comput.* **22**(2), 345–359 (2018)
11. Douglas, D.H., Peucker, T.K.: Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization* **10**, 112–122 (1973). <https://doi.org/10.3138/FM57-6770-U75U-7727>
12. Etzioni, O.: Commentary: Openai’s gpt-4 has some limitations that are fixable — and some that are not (March 2023), <https://www.geekwire.com/2023/commentary-openais-gpt-4-has-some-limitations-that-are-fixable-and-some-that-are-not/>, [Online; posted 14-March-2023]
13. Farsi, N., Mahjouri, N., Ghasemi, H.: Breakpoint detection in non-stationary runoff time series under uncertainty. *Journal of Hydrology* **590**, 125458 (08 2020). <https://doi.org/10.1016/j.jhydrol.2020.125458>
14. chung Fu, T.: A review on time series data mining. *Engineering Applications of Artificial Intelligence* **24**(1), 164–181 (2011)
15. Höppner, F.: Time series abstraction methods - a survey. pp. 777–786 (01 2002)
16. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**(12) (mar 2023). <https://doi.org/10.1145/3571730>
17. Kacprzyk, J., Zadrozny, S.: Fuzzy logic based linguistic summaries of time series: a powerful tool for discovering knowledge on time varying processes and systems under imprecision. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **6**(1), 37–46
18. Marin, N., Sanchez, D.: On generating linguistic descriptions of time series. *Fuzzy Sets and Systems* **285**, 6 – 30 (2016), special Issue on Linguistic Description of Time Series
19. Martinez-Cruz, C., Rueda, A.J., Popescu, M., Keller, J.M.: New linguistic description approach for time series and its application to bed restlessness monitor-

- ing for eldercare. *IEEE Transactions on Fuzzy Systems* **30**(4), 1048–1059 (2022). <https://doi.org/10.1109/tfuzz.2021.3052107>
20. OpenAI: Chatgpt — release notes (2023), <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>
 21. Ramos-Soto, A., Bugarín, A., Barro, S.: On the role of linguistic descriptions of data in the building of natural language generation systems. *Fuzzy Sets Syst.* **285**(C), 31–51 (Feb 2016)
 22. Reiter, E., Sripada, S., Hunter, J., Davy, I.: Choosing words in computer-generated weather forecasts. *Artificial Intelligence* **167**, 137–169 (2005)
 23. Sanchez-Valdes, D., Eciolaza, L., Triviño, G.: Linguistic description of human activity based on mobile phone’s accelerometers. In: *Ambient Assisted Living and Home Care - 4th International Workshop, IWAAL 2012*, Vitoria-Gasteiz, Spain, December 3-5, 2012. *Proceedings*. pp. 346–353 (2012)
 24. Trivino, G., Sugeno, M.: Towards linguistic descriptions of phenomena. *International Journal of Approximate Reasoning* **54**(1), 22 – 34 (2013)
 25. Wang, F.Y., Miao, Q., Li, X., Wang, X., Lin, Y.: What does chatgpt say: The dao from algorithmic intelligence to linguistic intelligence. *IEEE/CAA Journal of Automatica Sinica* **10**(3), 575–579 (2023). <https://doi.org/10.1109/JAS.2023.123486>