The Language Essence of the World: A Linguistic Interpretation of the Large Language Model †

Leiming Shi and Peng Wu *

Independent Researcher, Xi’an 710000, China; drshileiming@icloud.com
* Correspondence: wppem@hotmail.com
† Presented at the 2023 Summit of the International Society for the Study of Information (IS4SI 2023), Beijing, China, 14–16 August 2023.

Abstract: The source of characters is hieroglyph. Hieroglyph is the imitation and reference of the phenomena and objects in the natural world. The objects and phenomena in the natural world are the original image. Or rather its own figurative meta-hieroglyphics. As Baudrillard said, matter is also a symbol. Matter is actually its own language and speech. If we understand this key concept, then we can understand that the world itself is linguistic. The natural language of humans is based on modeling and translating the material language of the world.

Keywords: language; token; essence; large language model

1. A Large Language Model Is a Word Vector Table That Carries Word Attributes and Implicit Relationships

Natural language is not just words. The relationships, attributes, and grammar of words are the core of language. Words are the most visible stable structure of language, while semantics, attributes, and relationships are presented in the explanatory relationships within the network of words. The foundation of language is words, which are also the most determinable structures and data to grasp. However, simply understanding the appearance of each word is not enough, just like knowing the spelling and pronunciation of each word without understanding its semantics. The semantics of words are understood in their relationships with other words. Many of the core words correspond to the material and phenomena of the world, which can be seen as the translation, reporting, or reference to the “material language” of the world’s existence.

Therefore, using artificial intelligence to understand and generate language is essentially based on human natural language corpus statistics, establishing a vocabulary, and establishing its semantics in the correlation between words. Corresponding to human language activities, it is to establish a dictionary. Each word in the dictionary carries explanations, which are sentences and paragraphs composed of other words. Some fundamental words are formed by the prototype of the material world or introduced by hieroglyph referring to it. For example, at the beginning of learning, children establish basic words through tangible objects such as apples and bananas, or through the practice of recognizing words from pictures. In the process of continuous learning, they need to learn new words, and the semantics of new words are based on phenomena in the material world and explanations of old words and expressions. The dictionary corresponding to the large language model in artificial intelligence is currently different from human dictionaries. Since a large amount of artificial intelligence for large language model is learning within the scope of pure text, without visual or other sensory input, the dictionary has undergone a transformation. The foundation of this dictionary is still words, which are processed by technology as tokens, and the parameters that are associated with each token are neural networks, which compare, calculate, and filter out relationships with other words through
many layers of networks; some of these parameters contain parts of speech and attributes similar to human dictionaries, while others contain the probability of their connections with other words.

Since human words are composed of attributes, classifications, and sets, the implicit attributes of human words can be determined with just a few items and parameters. For example, an apple is both a type of food and a fruit, and a fruit is the product of a plant, which is a type of organism. However, the dictionary formed by the current artificial intelligence large language model relies on statistical operations based on data filtering using multilayer neural networks. In the absence of tangible sensory experiences and processes similar to human language learning, the learning of the large language model involves inputting massive amounts of high-quality language data into a neural network to statistically correlate it from articles, paragraphs, and sentences. At present, the breakthrough GPT model is based on contextual relationships. Although contextual relationships do not directly reflect the attributes of words, they are similar to the explicit and stable structure of words themselves. Contextual relationships are also a relatively stable and definite structure of expression. Therefore, although the statistical foundation and clues involved in this process are complex and computationally intensive, when the sample size is large enough, there is a possibility of statistical recognition and annotation of the implicit relationships between words. This process requires a sufficiently large and high-quality language sample, which is the corpus, as well as multilayered large neural networks for multidimensional filtering. The results obtained are expressed as probabilities and weights.

The current vocabulary unit of the GPT is the token, which is accompanied by over two thousand statistical parameters. This token, along with its associated parameters, forms the data structure known as a word vector, which serves as the fundamental unit and structure for natural language understanding, computation, and generation in the artificial intelligence large language model. While the explanatory terms are different from those in natural language dictionaries, they are extremely similar to the relationships between words and annotations in human dictionaries.

Therefore, the capabilities of large language models primarily stem from their understanding of the attributes of words, particularly the implicit and unexpressed ones, as well as the statistics and annotation of relationships between words. The immense scale, abstraction, and complexity of this structure arise from the statistical analysis of high-entropy context relationships. This structure, despite being beyond the inherent structure of individual words, is the most definite manifestation structure within language. Achieving this requires significant computational and statistical efforts.

Accordingly, statistics are effective in understanding language. However, first-order statistics can only distinguish the contextual connections of words. For higher-order statistics, there is a possibility of distinguishing between parts of speech and even the logical relationships contained within them. This is because a vast amount of language data, also known as corpus, contains abundant information, including parts of speech, semantics, logic, causality, and more. With a sufficiently large corpus sample and powerful neural network computing capabilities, higher-order statistics are sufficient to analyze the structures and algorithms of causal relationships and semantic computing logic calculations from the corpus. The current GPT, with billions of parameters and supercomputing power, seems to have supported this machine to learn the large language model, appearing to have surpassed a critical threshold, resulting in a general sense of understanding and narrative, and even the ability to analyze semantics and logic.

Therefore, understanding the essence of the large language model is helpful in improving and enhancing its efficiency. Additionally, it reminds us that the analysis and understanding of the parameters behind the token are crucial.

Currently, ChatGPT consists of approximately 50,000 tokens in its vocabulary. The parameters following the tokens, similar to a genetic sequence, form a sequence of parameters. Their nature and semantics are the key and code for understanding the large language
model and language-based general artificial intelligence. Alternatively, one could say that a sufficiently large number of input language samples or corpus contains a model of the world, including the semantics of words and relationships between various objects. In other words, high-quality corpora contain nearly all human knowledge and the intelligence within them. These massive corpora, as information objects, contain entropy, structure, and gradients, as well as other implicit relationships among their internal information. Through statistical analysis and processing with large neural networks and computational power, it is possible to establish a hierarchical and lossless compression and transformation of information. Countless corpora, processed through neural network computations, produce a new information structure in the form of tokens and their associated parameters, forming a word vector library. This word vector library, along with tokens and their parameters, serves as another language system and dictionary built by artificial intelligence, which is based on its own binary computer language system to borrow and translate natural language. This language system shares common tokens with human language, but there are significant differences in the expression of parameters or word properties. Ambiguities also exist within this system. Currently, these parameters following the tokens are abstract and opaque to humans. Being able to translate and interpret these parameters would greatly contribute to our full understanding of the large language model.

From these analyses, we can understand that the essence of the large language model is language and mathematics. Statistics is not the ultimate outcome, rather, the hidden, deep-level relationships and structures of language, including parts of speech, semantics, logic, and causality, that statistics discover and analyze are incorporated into a new computational framework. It is no longer limited to statistics and statistical results themselves, but elevates the language to the level of both data and algorithms, and even a program. Such a large language model transforms natural language into a new language structure, at the same time upgrading language itself to a program, not just a corpus. Based on the preceding context, the large language model automatically generates subsequent text by computing the relationships between tokens and their parameters, utilizing the context provided.

In comparison, human learning, focusing on textual symbols, assumes an average person reads 2000 books throughout their lifetime. Let us assume each book consists of 150,000 words, resulting in a reading corpus of only 300 million words. Even for an undergraduate student, assuming he has only read 500 books by the time of graduation, his reading volume would be less than 100 million words. However, considering such a small reading sample, the intellectual capacity generated by humans is astonishing. While their total knowledge may be far less than GPT, their abilities in analysis, thinking, and creativity are powerful. From this perspective, we can also conclude that the current efficiency of the large language model is low. Despite inputting massive amounts of language data, the current large language model has a greater ability to memorize and grasp surface-level knowledge than human individuals. However, its understanding of the deep-level knowledge within language, as well as its analytical abilities and mastery of causal and logical reasoning methods, are insufficient. The key to its further improvement and enhancement still lies in a deep and accurate grasp of the relationships, parts of speech, attributes between words in language, and the semantic and logical calculations based on this, as well as thinking methods. Additionally, multimodal language models will bring even greater progress, since images and sounds, as forms of pictorial representation, are connected to phenomena in the natural world. Human language and thoughts are inherently multimodal. Multimodal language models will undoubtedly enhance artificial intelligence with greater capabilities and intelligence.

At least, the current large language model has proven that general artificial intelligence is based on language and mathematics. The essence of universal artificial intelligence is language and mathematics.
2. The Internal and External Aspects of Natural Language

There is a vocabulary within natural language, which also includes the attributes and relationships of words. Its external phenomena are sentences, articles, and even books. Its internal structure is an explanatory network that constantly generates new words from the interpretation of old words. This network’s structure is reflected in dictionaries and platforms like Wikipedia. We should not view entries in dictionaries and Wikipedia as isolated islands of entries. In fact, they are a network. Entries form nodes, and explanations form connections or vectors formed by other entries. These entries, in turn, form a vast network with explanatory connections or vectors. The initial vocabulary and concepts generated by this network are based on naming and referring to phenomena and objects in the natural world. Some of these references are used to describe phenomena and objects, while others describe motion, relationships, and changes. Human language has developed gradually through the expansion of the initial vocabulary. New entries rely on the explanations provided by existing entries, and the discovery of new objects and relationships in the world leads to the creation of new entries.

Therefore, dictionaries and platforms like Wikipedia constitute a network of word interpretations. This network has a definite structure, with nodes represented by individual entries, and the connections formed by word associations within sentences create vectors. Sufficient statistical analysis of these sentence-based word associations can construct this network. Each entry in this network can be seen as having hidden attributes and weights representing all possible connections.

The statistics and calculations of the large language model take the basic unit of an entry as a token, and the more than 2000 parameters associated with each token capture their connections with other words, indirectly implying other parts of speech and attributes of the token.

For the external part, natural language is the model of the external world. The language created by humans is the modeling of material symbols in the external world. Natural language as a whole is a model and map of the external world, which is why it encompasses almost all of human knowledge. Moreover, natural language corresponds to the objects and phenomena of the external natural world, while the image and sound of the world are another language, equivalent to the hieroglyph of the early natural language. In other words, we can understand that the collection of sensory signals such as images and sounds is a kind of generalized hieroglyph. Even objects and phenomena themselves can be regarded as the source prototype and equivalent language of this hieroglyph.

Therefore, within natural language, there are words, grammar, semantics, explanatory relationships, word connections, networks, external mappings that correspond to objects and phenomena in the world, as well as relationships. From a mathematical perspective, there are also sets and logic within the vocabulary of natural language. There are relationships between sets and logic, as well as algorithms, within natural language. Through the calculation of semantics, sets, logic, and grammar, natural language not only narrates, but also performs semantic sets and logical calculations, which constitutes the computational and analytical capabilities of natural language.

The large language model has demonstrated that neural networks based on natural language are the correct and effective path towards achieving general artificial intelligence. The large language model adopts an approach by utilizing natural language and constructing word blocks based on natural language terms to generate parameters represented by numbers. These parameters are too abstract for humans and can be seen as a foreign language. In the generation stage, these parameters participate in operations but do not appear in the final results, which only manifest in human natural language. In this regard, tokens are similar to foreign language dictionaries. A token is an object language, and the subsequent parameters are similar to the interpretation of another subject language. And this subject language is the language of the large language model itself. The large language model still utilizes and constructs a language translation mechanism. This is not surprising, as human natural language is primarily a modeling and translation of nature.
and the world, including human society. Human beings have also formed natural language based on the modeling and translation of the world. Similar to English and Chinese, they both model and translate the shared external world; as a result, English and Chinese can be translated between each other.

The Rosetta Stone is the key to deciphering ancient Egyptian language for modern humans. The large language model has found a key for computers to understand natural language. That is, the connection between tokens, which represent relatively deterministic structures within language. By utilizing sufficiently large samples and sufficient multilevel statistics, the relationship between tokens can be established to approximate an accurate model.

3. Lingualism

The natural language of the world is a four-dimensional language of time and space. Human natural language, on the other hand, is a one-dimensional sequence of word symbols [1]. Sometimes, our thinking ability is limited by the inefficiency of this one-dimensional language and its structure. However, the overall and implicit structure of natural language is a vast network of word interpretations. This network constitutes a high-dimensional structure and mathematical algorithm.

On one end of the world, there is the material language of the material world [2], the intermediate is the natural language of humans, and on the other end, there is an artificial intelligence large language model that inherits the knowledge and structure of natural language. The material world of nature and its material language represent the origin of the world, while the artificial intelligence large language model inherits human natural language and its contained knowledge and algorithms, representing the destination of the world [3].

The artificial intelligence large language model reveals the linguistic essence and instinct of the world. We can refer to this essence as “Lingualism”.

Author Contributions: Writing—original draft preparation, revised and modify: L.S.; Resources collecting, review and editing: P.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.