AI Linguistics: Chapter 1, Creating New Words for Humans and Designing New Interfaces for AIs

David Archer
Department of Computer Engineering
San Jose State University
San Jose, USA
guosheng.zhang@sjsu.edu

***Abstract*—AI operates continuously in a high-dimensional space, distinct from the discrete, lower-dimensional realm of human cognition. This fundamental difference, coupled with the finite vocabulary of human language, creates a barrier to accurately conveying AI-generated concepts, contributing to the phenomenon of AI "hallucinations" and potentially impeding future communication with advanced artificial superintelligence (ASI). To bridge this gap, we propose modifying Large Language Models (LLMs) to use continuous input and output through direct embeddings, liberating AI operations from the confines of human language and accelerating the evolution of AI. By leveraging FAISS and a translator for interpretation, we introduce a method to expand human vocabulary dynamically, based on proximity in embedding space. This approach aims to enhance the precision of AI-human communication, fosters clearer interactions with ASI, and facilitates the evolution of human language to better capture the nuances of AI thought.**

***Index Terms*—AI linguistics, artificial superintelligence, continuous operations, discrete operations, embedding space, large language models, translator, vocabulary**

I. Introduction

High-dimensional vectorized word operations are key to breakthroughs in large language models (LLMs), which also highlights the fundamental differences between AI thinking and human thinking. The primary distinction lies in the dimensional gap. Human thinking, constrained by the types of information transmitters between neurons [1], typically operates within a space of up to 100 dimensions. The everyday thinking of most people usually does not exceed 10 dimensions. To engage in higher-dimensional thinking, we often compress part of this space, thereby reducing its dimensionality. For example, when considering the overall affairs of a university, we typically compress the information from various departments into a single-dimensional overview. Detailed consideration of a department's internal dimensions arises only when focusing on that department's internal affairs. This nested structure of thought [2] extends upwards to the Department of Education, state affairs, national affairs, and even global human affairs. Similarly, it stretches downwards to the curricula of individual professors and the management of each class, as well as horizontally to include inter-university organization and the coordination of external affairs. Human beings utilize this nested structure of thinking spaces to efficiently address the world's complex problems within a constrained dimensional space. This approach economizes thinking effort but introduces a significant issue: information distortion during transformation and transmission across different nested spaces, making it challenging to consider local nuances in broader contexts. In contrast, AI thinking can construct a higher-dimensional space through word vectorization, enabling direct calculation of complex, global problems while attending to detailed local variables. Because there is no loss of information conversion in different spaces, AI thinking can be very precise.

Another fundamental difference is the mode of operation. The human thinking is largely based on language. Human language is made up of words, and a word is a tiny, defined area with fuzzy boundaries in the space of the human mind. The human thinking process involves jumping between different words in a discrete manner. Additionally, the vocabulary of human language is not evenly distributed within the mind space. When comparing various human languages, researchers find that different languages have differing numbers of words within the same domain. Similarly, within a single language, there are different numbers of words across various domains. Humans create new words and eliminate those that are no longer needed, adapting their mental lexicon to the demands of reality [3]. The discrete mode of operation of the human mind requires that the results of human thinking must be settled on a series of words. In contrast, AI thinking uses high-dimensional vectors as tools to perform continuous operations. The result of its thinking can be any point in the high-dimensional space, but that point may not have an exact human word. Currently, large language models (LLMs) attempt to approximate these results by selecting the nearest word within this vector space. This discrepancy highlights a fundamental conflict between the limitless potential outcomes of AI thinking and the finite vocabulary of human language. Furthermore, this misalignment contributes to the phenomenon of AI "hallucinations" in understanding and translation between AI-generated content and human languages. In each iteration where LLMs generate a word output, the generated approximate word is used as input for the next, causing errors to potentially accumulate [4] and the AI's otherwise very precise thinking to become more and more fuzzy and hallucinating.

The remainder of this paper is structured as follows: the second part reviews human-computer communication research; the third proposes our solutions; the fourth discusses the simulation experiments; and the fifth concludes our findings.

II. Related Work

Enabling computers to understand human language in order to communicate effectively between humans and machines has been a central topic since the beginning of computers. Specifying a set of transistor circuit switching modes for each character like "01010010" is the low-level solution [5]. This scheme has been used to this day due to its simplicity, and it is estimated that it will not change until quantum computers use quantum spin schemes [6]. But machines can't really understand the meaning of these characters, let alone the meaning of words, sentences, and articles composed of these characters. So the initial effort to get machines to understand human language was to count, comparing the number of times a word appears in two pieces of document to determine their similarity or "meaning" or classification. TF-IDF is a masterpiece [7][8]. People pick out "meaningful" words as "features" for machine learning, and count them as a number if a piece of text has those words, and as zero if they don't. These numbers become the characteristic values of a piece of text and are used for comparison and calculation. If we use keywords as feature names, we can also use TF-IDF to calculate the sparse vector of a text. For further refinement, this method was extended to the word level, which turned a word list into a one-hot matrix [9]. each word row only has a 1 at the intersection point with its own word column, and all the others are 0's. Such a matrix is an extremely sparse matrix. The next step is to evolve into a dense matrix with a greatly reduced dimensionality, which can be computed by PCA or other feature engineering methods, where each word can be represented by a set of decimal vectors of lower dimensions. Word2Vec[10][11] greatly improves this process by using a neural network to calculate a vector value for each word based on a given corpus, using only one layer of neural network, which contains not only the meaning of the word, but also its context and syntax, as well as all the information that the word presents in the given text. Because of its rich meanings and powerful functions, such a vector has far surpassed the previous PCA concept of laten features, and has evolved into a new species, so it is named word embedding, to distinguish it from all previous word representations.

Word embedding is widely used to train language models [12] [13], and it has shown amazing results because of its ability to understand the meaning of words in depth. Attempts to develop different embedding models [14] [15] to help machines understand human language more deeply continued, and the embedding competition on HuggingFace has been very hot [16]. New embedding training models are constantly emerging, and few people stay in the first place position for too long. In fact, this is a core competition for machines to understand human languages. The most powerful embedding models available today are already very complex large language models with more than billions of parameters [17].

In addition to developing better models, another effort is to change the way embedding is used from static to dynamic [18]. The trained embedding is used as the input of the language model, and then further trained during the language model training process to deeply fit the training text used. Dynamic embedding is currently the dominant use of generative AI models and is a key cornerstone of transformer architectures that can achieve amazing results [19]. After it was discovered that embedding and the hidden layers of language models have a substitution relationship [20], many models simply fused the two together, combining embedding models and language models into a supermodel with hundreds of billions or even trillions of parameters [21] [22] [23] [24]. This simplifies the input and output of the models. In this way, it is hoped that artificial general intelligence (AGI) will be trained. When the training of static embedding generally adopts 1000-4000 dimensions [25], the large language model using multiple expert modules is highly integrated with the architecture of dynamic embedding, and its comprehensive dimension has reached 5000-12000 [26]. Because it has been found that the higher the dimension of embedding, the better the model's understanding of the language, and the stronger the model's capabilities.

At the end of this path, we can expect AGIs that use higher-dimensional embedding, nest more expert modules inside, and have more complex structures and more parameters. For such an AGI, its output is still to select a suitable word from the human language vocabulary [27]. It may be more accurately selected than the current LLMs [28], and its knowledge boundaries can be closer to the boundaries of human knowledge. AGI is an encyclopedic generalist who has reached and approached the highest level of human beings in every professional direction, and is the sum total of all human knowledge. However, due to the limitation of input and output interfaces, the huge creativity of AI is locked by the limited language of humans, and the communication between AI and AI has become very difficult, because human languages are foreign languages to AI.

III. Solution

The language of AI is embedding, and the language of humans is English and other natural languages. Here we see that a key bottleneck in the evolution from AGI to ASI is that it cannot break through the limitations of human language, and even if its AI mind creates new knowledge, it is difficult for AI to accurately express its new concepts with the vocabulary of human language.

A. New interfaces

To enable AI to overcome the constraints of human language, enabling it to think freely and communicate its ideas seamlessly with humans, we have modified the language model's input and output to embeddings. This adaptation allows the AI model to operate entirely in its own language, from input through to output, facilitating learning, thinking, and communication in a manner intrinsic to AI.

So we need a translator that is responsible for translating between human words and embeddings. When the AI model

processes human language information, the translator converts this into embeddings. Upon generating thinking results, the model outputs these as embeddings too, which the translator then decodes back into human language, creating new words as needed. So we need a standard embedding dictionary. This dictionary is primarily for communication between humans and AI. This is because AI-to-AI communication can take place directly through their own language embeddings. Currently, AI primarily learns and communicates using human-created knowledge. Therefore, if AIs can adopt a shared dictionary to learn human knowledge, it also facilitates easier learning from one another, and allowing them to build upon each other's foundations.

Since embedding is a language that can be used for infinitely refined thinking, when all AIs are free to use embedding, the universal language of the AI world, to learn, think and communicate, the output of one AI can become the input of another AI without barriers, and all AIs can deeply cooperate to process more and more profound thinking and produce more and more profound insightful results, which will greatly accelerate the evolution of AI. It's like the reality of human experts to use the same language to communicate and stimulate new ideas and generate creative sparks for each other. Direct communication between AIs allows AIs to relay extremely fine thinking without stopping, translating embedding into human language only when it is necessary to communicate with humans. The current state is that AI models are reproductively isolated from each other and reproduce in solitary ways, and communication between AIs requires the intervention of human engineers or human language tools to convey information, which is unnatural to AI.

B. New words

The first problem faced by translators prepared for human-AI communication is the asymmetry between embedding and human language vocabulary. Embedding has infinite possible outcomes, while human language has only a limited vocabulary. In the high-dimensional thinking space, an embedding of the AI thinking result may fall near a word in human language, and we can directly use this word to translate this embedding. This is common before artificial general intelligence (AGI), because the knowledge that AI has is basically all that can be expressed in human language. However, after general artificial intelligence, especially artificial superintelligence (ASI), it will continue to create some new knowledge, and some of the embeddings in its thinking results may fall in areas where there are no human language words nearby, and the translator will not be able to find suitable words to translate these embeddings. That's when we need to create new words to help humans precisely understand the idea of AI.

We can set a distance threshold for embeddings. If the translator finds a human language word within this spatial threshold, it translates the embedding into the nearest word. If no human language word falls within this threshold, a new word is created.

There can be a variety of ways to create new words to suit different specific situations. Here are three basic methods.

The first is the neighborhood method. Simply choose the 2-3 words that are closest to the embedding. This method is simple and straightforward and can be used in any situation. The way to combine them can be as simple as arranging the words together according to the distance from closest to farthest.

The second is the central neighborhood method. Select a top_k word combination, whose center point can reach the threshold, to form a new word. This method is a more accurate way to translate, as embedding B in Figure 1, can choose Y, X, and Z to form the new word.

The third is the projection method. Sometimes, because all human language words are on one "side" of embedding, it is not possible to find a group of words whose center point is close to embedding. For example, embedding A, at this time, we can first project, project the two words X and Z that are closer to the other side of A with A as the center, form two temporary auxiliary words vi-X, vi-Z, and then use these four words to form a new word, and then simplify it to X-Z-vi.

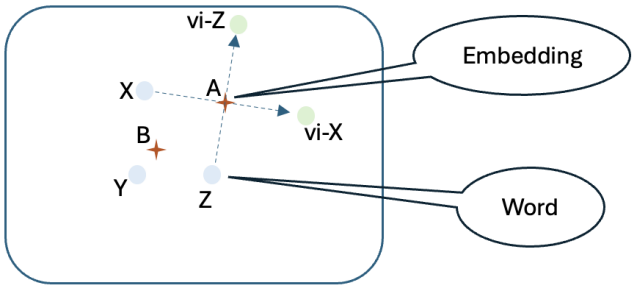

Fig. 1. Word formation

We can find some more ways to construct new words, but all of them should help people understand new words directly as much as possible, and grasp and use new words quickly. Human AI linguists and AI engineers can shape and recreate new words based on their meanings after the translator creates them, making them more in line with people's language habits.

The embedding value of the new word is readily available, so the new word can be immediately written into the standard embedding dictionary and immediately put into use in the communication between humans and AIs.

IV. Simulation Experiments

The main purpose of our simulation experiments was to demonstrate an executable operational process, so only 30 English words were selected as an example of a mini language. Experiments that can demonstrate its practical application effects need to be carried out on a model that has basically achieved artificial general intelligence. Our simulation procedure is as follows:

1. Using the word usage frequency distribution data from the Wordfreq [29] library, 30 words were randomly selected from the 150 most frequently used words weighted according to their frequencis. These words were:

"the, to, and, of, a, in, i, is, for, that, you, this, at, by, your, an, all, they, like, about, what, who, their, there, she, new, other, its, these, work"

2. On hugging face, the currently best open-source embedding model Salesforce/SFR-Embedding-Mistral [17] was selected to generate embeddings for these 30 words, dimension 4096, and a bidirectional translation dictionary was created as the basis of the translator. As a counterpoint, we also built a bidirectional translation dictionary of 100,000 words.

3. To build a phrase dataset, we randomly selected about 50B word sample documents from the large text databases COCA[30], Wikipedia Dumps[31], and American Stories[32], from which about 128,000 sentences or sentence fragments, that contain only the selected 30 words, were selected.

4. Designed a GPT model, embedding in and embedding out. This GPT model used the overall architecture of 4 layers of 8-head encoders, and added position embedding to the input embedding, which was processed by the model and finally outputted embeddings.

5. Used the above phrase dataset to train the GPT model.

6. Based on the translation dictionary and FAISS [33] search, we designed a translator. We experimented with some prompts and translated some of the embedding outputs of the model, giving the translator a chance to generate new words by adjusting the parameter of the high-dimensional spatial distance threshold. At the same time, we used 100,000 embedding dictionaries to find the English words that this new word might correspond to. Some interesting examples are shown in Table I.

TABLE I
NEOLOGISMS COINED BY AI

	Example embeddings output by AI and translated for the mini language	***Interesting words***	***Embeddings translated with the 100k-word vocabulary***
1	in and-by-of is for that	and-by-of	continued
2	and and-the-in	and-the-in	and-the-or
3	of a and-by-vi	and-by-vi	and-cont'd-vi
4	that you the-and-in	the-and-in	the-and-thew
5	this at by your	this at by your	it on with this

As can be seen from the examples, if a suitable word cannot be found to translate the embedding in a 30-word mini-language system, there is still a big chance that a ready-made word cannot be found in a larger 100,000-word language system. This proves the extreme sparseness of human language in high-dimensional space. We also noticed that the meaning of the larger language system translation is a bit different from that of the mini language system translation, which clearly explains part of the origin of the AI hallucinations.

A good translator can continue to create new words in the process of accelerating the evolution of AI, ensure that humans can always precisely understand AI and communicate clearly with AI in both directions, and help humans achieve co-evolution with AI in intelligence.

V. DISCUSSION

The language of AI is embedding, and the languages of humans are English and other natural languages, and if we look at the communication between humans and other agents from the perspective of intelligence-centrism rather than anthropocentrism, it is inevitable that AI will use its own language to learn, think, and communicate. And it's only natural that humans need a translator to communicate with AI. We propose to divide the large language model ecosystem into two parts. One is the AI model. AI models should be designed and optimized according to the embedding operation logic of the AI world, which requires us to conduct more in-depth research on the continuous computation in high-dimensional space. The second is the translator. The first translators were, of course, the translations between human languages, such as English, and AI language embedding. In addition, there are many physical, chemical and biological signals in the world that need to be translated into embeddings. In fact, what we need to do is embedding everything, completely translate our real world into the embedding world of AI, so that AI can truly understand the world and start scientific exploration in the real sense. Therefore, the development of various translators is an important prerequisite for the development of artificial superintelligence. Imagine that if everything in our world has a corresponding existence in the embedding world of AI, ASI is a matter of course.

All translators need to be standardized, including the dictionaries and the distance thresholds.

VI. CONCLUSION

Our research focuses on the two most fundamental issues of AI linguistics, one is to unravel the linguistic limitations of AI development, allowing AI to freely use its own language embedding to learn, think, and communicate. The second is to create new words to help humans understand AI more accurately and communicate and grow together with AI.

The accelerated development of AI has become inevitable, and how to maintain smooth communication between humans and AI has become increasingly urgent.

REFERENCES

[1] V. Di Maio and S. Santillo, "Information Processing and Synaptic Transmission," in Advances in Neural Signal Processing, R. Vinjamuri, Ed., IntechOpen, 2020. doi: 10.5772/intechopen.88405.

[2] J. L. S. Bellmund, P. Gärdenfors, E. I. Moser, and C. F. Doeller, "Navigating cognition: Spatial codes for human thinking," Science, vol. 362, no. 6415, p. eaat6766, Nov. 2018, doi: 10.1126/science.aat6766.

[3] T. Diyora, "Vocabulary As An Adaptive System And Understanding Language Development," Educational News Research in the 21st Century, vol. 2, no. 19, Art. no. 19, Mar. 2024.

[4] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith, "How Language Model Hallucinations Can Snowball." arXiv, May 22, 2023. doi: 10.48550/arXiv.2305.13534.

[5] J. V. Atanasoff et al., "The Invention of Electronic Digital Computing - Plenary Panel Summary," in 2023 IEEE John Vincent Atanasoff International Symposium on Modern Computing (JVA), Chicago, IL, USA: IEEE, Jul. 2023, pp. 8–8. doi: 10.1109/JVA60410.2023.00011.

[6] J. Singh and M. Singh, "Evolution in Quantum Computing," in 2016 International Conference System Modeling & Advancement in Research Trends (SMART), Moradabad, India: IEEE, 2016, pp. 267–270. doi: 10.1109/SYSMART.2016.7894533.

[7] H. R. Lewis, Ed., Ideas That Created the Future: Classic Papers of Computer Science. The MIT Press, 2021. doi: 10.7551/mitpress/12274.001.0001.

[8] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," J. Doc., vol. 60, no. 5, pp. 503–520, Jan. 2004, doi: 10.1108/00220410410560582.

[9] E. Arnaud, M. Elbattah, M. Gignon, and G. Dequen, "NLP-Based Prediction of Medical Specialties at Hospital Admission Using Triage Notes," in 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), Victoria, BC, Canada: IEEE, Aug. 2021, pp. 548–553. doi: 10.1109/ICHI52183.2021.00103.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space." arXiv, Sep. 06, 2013. Accessed: Mar. 22, 2024. [Online]. Available: http://arxiv.org/abs/1301.3781

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality." arXiv, Oct. 16, 2013. Accessed: Mar. 22, 2024. [Online]. Available: http://arxiv.org/abs/1310.4546

[12] L. Rheault and C. Cochrane, "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora," Polit. Anal., vol. 28, no. 1, pp. 112–133, Jan. 2020, doi: 10.1017/pan.2019.26.

[13] S. Jansen, "Word and Phrase Translation with word2vec." arXiv, Apr. 24, 2018. Accessed: Mar. 22, 2024. [Online]. Available: http://arxiv.org/abs/1705.03127

[14] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents." arXiv, May 22, 2014. Accessed: Mar. 22, 2024. [Online]. Available: http://arxiv.org/abs/1405.4053

[15] I. Banerjee, M. C. Chen, M. P. Lungren, and D. L. Rubin, "Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort," J. Biomed. Inform., vol. 77, pp. 11–20, Jan. 2018, doi: 10.1016/j.jbi.2017.11.012.

[16] "MTEB Leaderboard - a Hugging Face Space by mteb." Accessed: Mar. 23, 2024. [Online]. Available: https://huggingface.co/spaces/mteb/leaderboard

[17] "SFR-Embedding-Mistral: Enhance Text Retrieval with Transfer Learning," Salesforce AI. Accessed: Mar. 23, 2024. [Online]. Available: https://blog.salesforceairesearch.com/sfr-embedded-mistral/

[18] Y. Wang, Y. Hou, W. Che, and T. Liu, "From static to dynamic word representations: a survey," Int. J. Mach. Learn. Cybern., vol. 11, no. 7, pp. 1611–1630, Jul. 2020, doi: 10.1007/s13042-020-01069-8.

[19] J. Von Der Mosel, A. Trautsch, and S. Herbold, "On the Validity of Pre-Trained Transformers for Natural Language Processing in the Software Engineering Domain," IEEE Trans. Softw. Eng., vol. 49, no. 4, pp. 1487–1507, Apr. 2023, doi: 10.1109/TSE.2022.3178469.

[20] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." arXiv, Feb. 08, 2020. Accessed: Mar. 23, 2024. [Online]. Available: http://arxiv.org/abs/1909.11942

[21] "xai-org/grok-1." xai-org, Mar. 23, 2024. Accessed: Mar. 23, 2024. [Online]. Available: https://github.com/xai-org/grok-1

[22] M. Schreiner, "GPT-4 architecture, datasets, costs and more leaked," THE DECODER. Accessed: Mar. 23, 2024. [Online]. Available: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

[23] S. Jindal, "Meta To Release Llama 3 in July, Outperforms GPT-4 and Gemini," Analytics India Magazine. Accessed: Mar. 23, 2024. [Online]. Available: https://analyticsindiamag.com/meta-to-release-llama-3-in-july-outperforms-gpt-4-and-gemini/

[24] D. A. D. Thompson, "The Memo - Special edition: Claude 3 Opus," The Memo by LifeArchitect.ai. Accessed: Mar. 23, 2024. [Online]. Available: https://lifearchitect.substack.com/p/the-memo-special-edition-claude-3

[25] "Embeddings," Voyage AI. Accessed: Mar. 23, 2024. [Online]. Available: https://docs.voyageai.com/docs/embeddings

[26] kalyan, "GPT-4 Everything You Need To Know," MAKE ME ANALYST. Accessed: Mar. 23, 2024. [Online]. Available: https://makemeanalyst.com/gpt-4-everything-you-need-to-know/

[27] S. K. Routray, A. Javali, K. P. Sharmila, M. K. Jha, M. Pappa, and M. Singh, "Large Language Models (LLMs): Hypes and Realities," in 2023 International Conference on Computer Science and Emerging Technologies (CSET), Oct. 2023, pp. 1–6. doi: 10.1109/CSET58993.2023.10346621.

[28] M. T. Bennett, "Enactivism & Objectively Optimal Super-Intelligence." arXiv, Mar. 08, 2023. Accessed: Mar. 23, 2024. [Online]. Available: http://arxiv.org/abs/2302.00843

[29] E. R. L. (Robyn Speer), "rspeer/wordfreq." Apr. 02, 2024. Accessed: Apr. 02, 2024. [Online]. Available: https://github.com/rspeer/wordfreq

[30] "coca2020_overview.pdf." Accessed: Apr. 02, 2024. [Online]. Available: https://www.englishcorpora.org/coca/help/coca2020_overview.pdf

[31] "Data dumps - Meta." Accessed: Apr. 02, 2024. [Online]. Available: https://meta.wikimedia.org/wiki/Data_dumps

[32] "American Stories." Accessed: Apr. 02, 2024. [Online]. Available: https://dell-research-harvard.github.io/resources/americanstories

[33] "facebookresearch/faiss." Meta Research, Apr. 03, 2024. Accessed: Apr. 02, 2024. [Online]. Available: https://github.com/facebookresearch/faiss