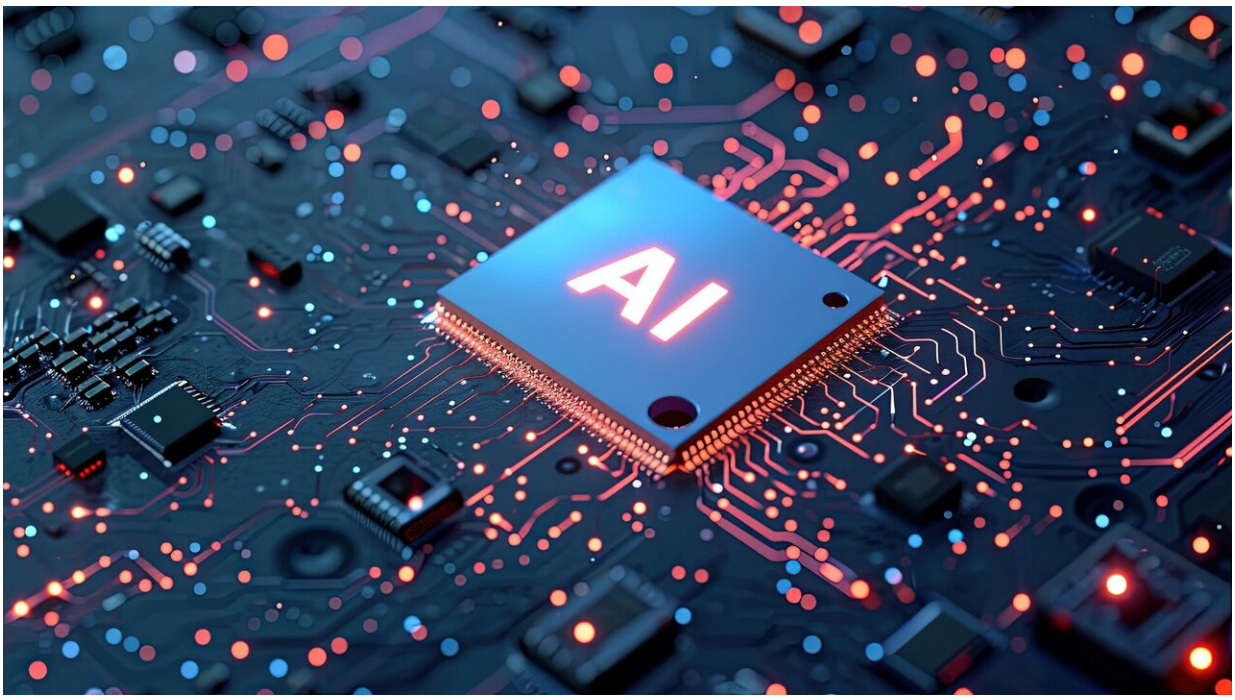
Large language models trained in English found to use the language internally, even for prompts in other languages

March 14 2024, by Tanya Petersen



Credit: Pixabay/CC0 Public Domain

EPFL researchers have shown that large language models primarily trained on English text seem to use English internally, even when they are prompted in another language. As AI increasingly runs our lives, this may have important consequences regarding linguistic and cultural bias.

Large language models (LLMs) including Open AI's ChatGPT and Google's Gemini have taken the world by storm, surprising with their ability to understand and respond to users with seemingly natural speech.

While it's possible to interact with these LLMs in any language, they are trained on hundreds of billions of text parameters mainly in English, and it has been hypothesized by some that they do most of their internal processing in English and then translate to the target language at the very last moment. Yet, there has been little evidence of this—until now.

Testing Llama

EPFL researchers from the Data Science Laboratory (DLAB) in the School of Computer and Communication Sciences studied the Llama-2 (Large Language Model Meta AI) open source LLM to try to determine which languages were being used at what stages along the computational chain.

"Large language models are trained to predict the next word. They do this by essentially matching every word to a vector of numbers, basically a multi-dimensional data point. The word 'the' for example will always be found at the exact same fixed coordinate of numbers," explained Professor Robert West, head of DLAB.

"The models chain together, say, 80 layers of identical computational blocks, each of which transforms one vector that represents a word into another vector. At the end of this sequence of 80 transformations what comes out is a vector representing the next word. The number of calculations is fixed via the number of layers of computational blocks—the more calculations that are done, the more powerful your model is and the more likely the next word will be correct."

As explained in their paper Do Llamas Work in English? On the Latent

Language of Multilingual Transformers, available on the pre-print server *arXiv*, instead of letting the model complete the calculations from its 80 layers, each time it was trying to predict the next word West and his team forced it to answer after each layer and they were able to see which word the model would predict at that point. They set up various tasks such as asking the model to translate a series of French words into Chinese.

"We gave it a French word, then the Chinese translation, another French word and the Chinese translation, etc., such that the model knows that it's supposed to translate the French word into Chinese. Ideally, the model should give 100% probability to the Chinese word but when we forced it to make predictions before the final layer we found that most of the time it predicted the English translation of the French word although English doesn't pop up anywhere in this task. It's only in the last four to five layers that Chinese is actually more likely than English," said West.

From words to concepts

A simple hypothesis would be that the model translates the entire input into English and translates into the target language right at the end, but in analyzing the data, the researchers came up with a far more interesting theory.

In the first phase of calculations there is no probability going to either word and they believe that the model is concerned with fixing input issues.

In the second phase, where English dominates, the researchers think the model is in some sort of abstract semantic space where it's not reasoning about single words but other kinds of representations that are more about concepts, universal across language and more of a model of the world.

This is important because in order to predict the next word well the model needs to know a lot about the world and one way to do this is to have this representation of concepts.

"We theorize that this representation of the world in terms of concepts is biased towards English, which would make a lot of sense because these models saw around 90% English training data. They map input words from a superficial word space into a deeper meaning space of concepts where there are representations for how these concepts relate to each other in the world—and the concepts are represented similarly to English words, rather than the corresponding words in the actual input language," said West.

Monoculture and bias

A key question that arises from this English dominance is 'does it matter'? The researchers believe it does. There is substantial research showing that structures that exist in language shape how we construct reality and that the words we use are deeply connected to how we think about the world. West suggests that we need to start researching the psychology of language models where they are treated as humans and, in different languages, interrogated, subjected to behavioral tests and assessed for biases.

"I think this research has really hit a nerve as people are becoming more worried about these kinds of issues of potential monoculture. Given that the models are better in English, something that is being explored now by many researchers is to feed in English content and translate back to the desired language. From an engineering viewpoint that might work but I would suggest that we lose a lot of nuance because what you cannot express in English will not be expressed," West concluded.

More information: Chris Wendler et al, Do Llamas Work in English? On the Latent Language of Multilingual Transformers, *arXiv* (2024). DOI: 10.48550/arxiv.2402.10588

Provided by Ecole Polytechnique Federale de Lausanne