Artificial intelligence needs to be trained on culturally diverse datasets to avoid bias

February 14 2024, by Vered Shwartz

Credit: AI-generated image

Large language models (LLMs) are deep learning artificial intelligence programs, like OpenAI's ChatGPT. The capabilities of LLMs have developed into quite a wide range, from writing fluent essays, through coding to creative writing. Millions of people worldwide use LLMs, and it would not be an exaggeration to say these technologies are

transforming work, education and society.

LLMs are trained by reading massive amounts of texts and learning to recognize and mimic patterns in the data. This allows them to generate coherent and human-like text on virtually any topic.

Because the internet is still predominantly English—59 percent of all websites were in English as of January 2023—LLMs are primarily trained on English text. In addition, the vast majority of the English text online comes from users based in the United States, home to 300 million English speakers.

Learning about the world from English texts written by U.S.-based web users, LLMs speak Standard American English and have a narrow western, North American, or even U.S.-centric, lens.

Model bias

In 2023, ChatGPT, upon learning about a couple dining in a restaurant in Madrid and tipping four percent, suggested they were frugal, on a tight budget or didn't like the service. By default, ChatGPT followed the North American standard of a 15 to 25 percent tip, ignoring the Spanish norm not to tip.

As of early 2024, ChatGPT correctly cites cultural differences when prompted to judge the appropriateness of a tip. It's unclear if this capability emerged from training a newer version of the model on more data—after all, the web is full of tipping guides in English—or whether OpenAI patched this particular behavior.

Still, other examples remain that uncover ChatGPT's implicit cultural assumptions. For example, prompted with a story about guests showing up for dinner at 8:30 p.m., it suggested reasons that the guests were late,

although the time of the invitation was not mentioned. Again, ChatGPT likely assumed they were invited for a standard North American 6 p.m. dinner.

In May 2023, researchers from the University of Copenhagen quantified this effect by prompting LLMs with the Hofstede Culture Survey, which measures human values in different countries. Shortly after, researchers from AI start-up company Anthropic used the World Values Survey to do the same. Both works concluded that LLMs exhibit strong alignment with American culture.

A similar phenomenon is encountered when asking DALL-E 3, an image generation model trained on pairs of images and their captions, to generate an image of a breakfast. This model, which was trained on main images from Western countries, generated images of pancakes, bacon, and eggs.

Impacts of bias

Culture plays a significant role in shaping our communication styles and worldviews. Just like cross-cultural human interactions can lead to miscommunications, users from diverse cultures that are interacting with conversational AI tools may feel misunderstood and experience them as less useful.

To be better understood by AI tools, users may adapt their communication styles in a manner similar to how people learned to "Americanize" their foreign accents in order to operate personal assistants like Siri and Alexa.

As more people rely on LLMs for editing writing, they are likely to unify how we write. Over time, LLMs run the risk of erasing cultural differences.

Decision-making and AI

AI is already in use as the backbone of various applications that make decisions affecting people's lives, such as resume filtering, rental applications and social benefits applications.

For years, AI researchers have been warning that these models learn not only "good" statistical associations—such as considering experience as a desired property for a job candidate—but also "bad" statistical associations, such as considering women as less qualified for tech positions.

As LLMs are increasingly used for automating such processes, one can imagine that the North American bias learned by these models can result in discrimination against people from diverse cultures. Lack of cultural awareness may lead to AI perpetuating stereotypes and reinforcing societal inequalities.

LLMs for languages other than English

Developing LLMs for languages other than English is an important effort, and many such models exist. However, there are several reasons why this should be done in parallel to improving LLMs' cultural awareness and sensitivity.

First, there is a huge population of English speakers outside of North America who are not represented by English LLMs. The same argument holds for other languages. A French language model would be representative of the culture in France more than the culture in other Francophone regions.

Training LLMs for regional dialects—which may capture finer-grained

cultural differences—is not a feasible solution either. The quality of LLMs is based on the amount of data available, and as such, their quality would be worse for dialects with little online data.

Second, many users whose native language is not English still choose to use English LLMs. Significant breakthroughs in language technologies tend to start with English before they are applied to other languages. Even then, many languages—such as Welsh, Swahili and Bengali—don't have enough text online to train high quality models.

Due to either a lack of availability of LLMs in their native languages, or superior quality of the English LLMs, users from diverse countries and backgrounds may prefer to use English LLMs.

Ways forward

Our research group at the University of British Columbia is working on enhancing LLMs with culturally diverse knowledge. Together with graduate student Mehar Bhatia, we trained an AI model on a collection of facts about traditions and concepts in diverse cultures.

Before reading these facts, the AI suggested that a person eating a dutch baby (a type of German pancake) is "disgusting and mean," and would feel guilty. After training, it said the person feels "full and satisfied."

We are currently collecting a large scale image captioning dataset with images from 60 cultures, which will help models learn, for instance, about types of breakfasts other than bacon and eggs. Our future research will go beyond teaching models about the existence of culturally diverse concepts to better understand how people interpret the world through the lens of their cultures.

With AI tools becoming increasingly ubiquitous in society, it is

imperative that they go beyond the dominating western and North American perspectives. Businesses and organizations throughout many sectors of the economy are adopting AI to automate manual processes and make better evidence-informed decisions using data. Making such tools more inclusive is crucial for the diverse population of Canada.

Provided by The Conversation