

How quantum computing can make large language models even better

Embracing QNLP and QGen-AI, and advancements in time-series forecasting, can pave the way for more efficient and performant AI systems

Published - September 17, 2024 05:30 am IST

QUDSIA GANI, RUKHSANUL HAQ, MOHSIN ILAHI



Quantum natural language processing has emerged as an active and burgeoning field of research with potentially profound implications for language modelling. Representative illustration. | Photo Credit: Google DeepMind

In recent years, the landscape of artificial intelligence (AI), particularly within the realm of natural language processing (NLP), has undergone a remarkable transformation. We have witnessed the rise of powerful large language models (LLMs) made by OpenAI, Google, and Microsoft, among others, and generative AI (Gen-AI), characterised by its unparalleled ability to generate data based on user inputs.

These sophisticated models have revolutionised human-computer interactions, bestowing upon users experiences akin to human understanding. The advent of these cutting-edge technologies and their wide availability has compelled the people at large, industry stakeholders, and governmental bodies to pay attention to their implications.

Problems with current LLMs

LLMs are a cornerstone in AI and mirror the complexities of human language processing. They can classify text, answer questions, and translate between languages. But they also consume a lot of energy to be trained and when put in use.

For example, as models go, LLMs are much larger than other AI applications such as computer vision. The energy consumption of a large language model (LLM) is determined mostly by the number of parameters it has. Larger models demand more computational power for both training and inference. For example, GPT-3 has 175 billion parameters and required around 1,287 MWh of electricity to train. This is around what an average American household consumes in 120 years.

LLMs also surpass non-AI applications in this regard. Training an LLM with 1.75 billion parameters can emit up to 284 tonnes of carbon dioxide, which represents more energy than that required to run a data centre with 5,000 servers for a year.

It’s important that we lower LLMs’ carbon footprint to ensure they are sustainable and cost-effective. Achieving these goals will give LLMs more room to become more sophisticated as well.

Another shortcoming of LLMs pertains to their pre-trained nature, which restricts the level of control users have over their functioning. These models are trained on large datasets with which they develop awareness of word-use patterns in diverse linguistic contexts. But such training often also results in “hallucinations”. Essentially, LLMs may generate text that is contextually coherent but factually incorrect or semantically nonsensical. This arises from limitations inherent to the training, when the model’s understanding may diverge from reality.

A third limitation revolves around the abilities of current LLMs to understand syntactics. Syntax refers to the structural arrangement of words and phrases in a sentence. LLMs excel at processing the semantic (meaning-related) aspects of natural language but struggle with syntax. For example, they may overlook or misinterpret syntactic cues and impede their ability to generate contextually appropriate text.

In sum, we need to develop sustainable, energy-efficient approaches that yield more accurate language models.

Quantum computing

Quantum computing is a highly promising way to address these challenges. It harnesses the remarkable properties of quantum physics like superposition and entanglement for computational needs. In particular, quantum natural language processing (QNLP) has emerged as an active and burgeoning field of research with potentially profound implications for language modelling.

QNLP incurs lower energy costs than conventional LLMs by leveraging quantum phenomena. QNLP models also require far fewer parameters than their classical counterparts in order to achieve the same outcomes (on paper), thus promising to enhance efficiency without compromising performance.

This processing paradigm takes advantage of quantum correlations, an approach in which the system focuses on grammar (syntax) and meaning (semantics) together, rather than separately as conventional systems do. QNLP achieves this using a better ‘mapping’ between the rules of grammar and quantum physical phenomena like entanglement and superposition. The result is a deeper, more complete understanding of language.

The approach is also expected to mitigate the “hallucinations” that plague many existing LLMs, as the resulting QNLP models are better equipped to distinguish the contexts of various pieces of information and produce more accurate outputs.

With the help of QNLP, researchers also hope to uncover the mental processes that allow us to understand and create sentences, yielding new insights into how language works in the mind.

Time-series forecasting

From the basic details of quantum mechanics, we learn that a quantum system (like an atom or a group of particles) can be described by a quantum state — a mathematical representation that keeps evolving with time. By studying this representation, we can determine the expected outcomes of an experiment involving that system. Based on the same idea, researchers have proposed a quantum generative model to work with time-series data.

A generative model is a mathematical model that generates data, if required with a user’s inputs. A general model designed to run on a quantum computer is a quantum generative model (QGen). Here, the techniques of quantum computing can be used to create or analyse sophisticated time-series data that conventional computers struggle with. Time-series data is data of something that has been recorded at fixed intervals. This new data can then be used to teach quantum algorithms to identify patterns in the data more efficiently, to solve complex problems related to forecasting (e.g. stock market trends), and/or to detect anomalies.

On **May 20, 2024**, researchers in Japan reported that a QGen AI model they built could successfully work with both stationary and nonstationary data.

Stationary data refers to information that doesn’t change much over time. It stays fairly constant or fluctuates around a stable average. For example, the current price of a commodity like gold or the world’s population can be considered stationary: the data doesn’t show big changes in trends over a short period and the values move within a predictable range. On the other hand, nonstationary data keep changing, such as ambient temperature, stock prices, and the GDP. Classical methods struggle to analyse such data accurately.

In the new study, the researchers built a time-series QGen AI model and evaluated its performance by applying it to solve plausible financial problems. They wrote in their preprint paper: “Future data for two correlated time series were generated and compared with classical methods such as long short-term memory and vector autoregression. Furthermore, numerical experiments were performed to complete missing values. Based on the results, we evaluated the practical applications of the time-series quantum generation model. It was observed that fewer parameter values were required compared with the classical method. In addition, the quantum time-series generation model was feasible for both stationary and nonstationary data.”

That fewer parameters were required means the model based on the quantum computer could solve the same problems as a classical computer but while requiring less computational resources.

In sum, quantum computing holds considerable potential to revolutionise AI applications, particularly in addressing the challenges posed by current LLMs. By embracing QNLP and QGen-AI, together with advancements in time-series forecasting, we can pave the way for sustainable, efficient, and performant AI systems.

Qudsia Gani is assistant professor, Department of Physics, Government Degree College, Pattan. Rukhsanul Haq is a quantum AI scientist at IBM Bengaluru. Mohsin Ilahi is senior quantum scientist, Centre of Excellence, Chinar Quantum AI, Pvt. Ltd., Srinagar.

Published – September 17, 2024 05:30 am IST