DELL Technologies

Whitepaper

Unleashing the Power of Large Language Models like ChatGPT for Your Business

By Ben Fauber, Ph.D., Senior ML Research Scientist,
Distinguished Member Technical Staff, Dell Technologies

Large language models (LLMs) are a type of artificial intelligence (AI) system that uses machine learning (ML) algorithms to process vast amounts of natural language text data. They can perform various natural language processing (NLP) tasks, such as:

- Text classification,
- Text summarization,
- Text generation,
- Named entity recognition (NER),
- Text sentiment analysis, and;
- Question-answering (Q&A).

Many large language models are multilingual, and they can generate text that is nearly indistinguishable from human written text. This has led to significant advancements in natural language processing and has numerous applications in various fields such as business, education, healthcare, and communication.

From Text to Speech: The Evolution of Large Language Models

The power of large language models lies in their ability to learn from massive amounts of text data and create accurate models of language. Many models were trained on more than 1 TB of text data from books (e.g., BooksCorpus), articles (e.g., RealNews, PubMed), websites (e.g., Wikipedia, Common Crawl, OpenWebText), and more. There are a few publicly available data sets available such as The Pile[1] and ROOTS[2] for training large language models, yet the many of the data sets used for training are proprietary and not public.

In language model training, the task is to predict a word in a sequence of words. In 2017, a breakthrough was achieved with the introduction of transformers.[3] Unlike previous deep learning language models, transformers can process all input data at once and assign different weights to different parts of the input data based on their position in the language sequence. This feature has led to significant enhancements in large language models, enabling them to handle much larger datasets and infer deeper meaning and context by differentially weighting various parts of the sequence.

> Large language models do not query the internet or other data sources for their knowledge. Rather, they use information and relationships embedded within their deep neural network (DNN) to provide responses to tasks and prompts.

Recent advancements in natural language processing have resulted in remarkable large language model outcomes, exemplified by GPT-3 (OpenAI),[4] Megatron-Turing (NVIDIA and Microsoft),[5] OPT-175B (Meta),[6] Bloom-176B (BigScience),[7] and most recently ChatGPT (OpenAI).[8] Many of these large language models provide state-of-the-art (SOTA) performance on natural language processing benchmark tasks such as:

- GLUE and SuperGLUE: The General Language Understanding Evaluation (GLUE) and SuperGLUE benchmarks include a set of diverse natural language understanding tasks, such as sentiment analysis, natural language inference, commonsense reasoning, textual entailment, and question answering.[9,10]

- SQuAD: The Stanford Question Answering Dataset (SQuAD) includes a set of reading comprehension tasks, where models are asked to answer questions based on a given passage.[11]

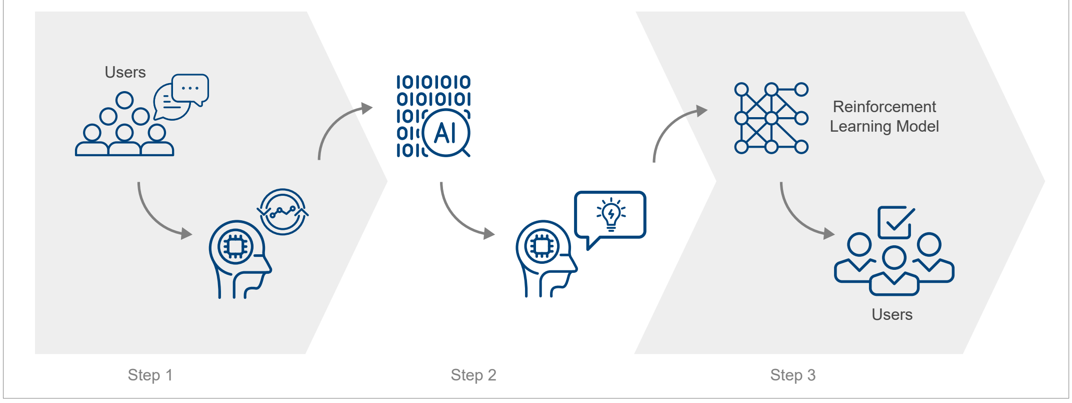
- SNLI and MultiNLI: The Stanford Natural Language Inference (SNLI) and the Multi-Genre Natural Language Inference (MultiNLI) corpora are benchmark datasets for natural language inference, where models are asked to determine whether a given hypothesis can be inferred from a given premise.[12,13]

- DREAM: The Dialogue-based Reading Comprehension (DREAM) dataset is a benchmark for evaluating models on answering questions that require reasoning and comprehension in a dialogue context.[14]

How Does ChatGPT Work?

ChatGPT has gained considerable attention since its initial release for its ability to write fluid human-like prose. ChatGPT is a refinement of the InstructGPT large language model. InstructGPT was created through the alignment of large language model outputs with user intent by incorporating reinforcement learning from human feedback (RLHF).[15]

Application of reinforcement learning from human feedback to create InstructGPT involved three major steps:



Ultimately, the performance of InstructGPT was compared against its predecessor, GPT-3, based on their ability to infer and follow user instructions (helpfulness), their tendency for hallucinations (truthfulness), and their ability to avoid inappropriate and derogatory content (harmlessness). InstructGPT outperformed GPT-3 on all three criteria, with human referees favoring the outputs of InstructGPT 85% of the time.

The large language models that wield the greatest power are extremely expansive, often boasting more than 100 billion model parameters and necessitating more than 800 GB of memory to activate, making them some of the largest machine learning models ever created.

These very large language models have showcased remarkable abilities in producing text that is nearly indistinguishable from human writing, tackling intricate inquiries, and even crafting computer code.

It is also noteworthy that large language models do not query the internet or other data sources for their knowledge. Rather, they use information and relationships embedded within their deep neural network (DNN) to provide responses to tasks and prompts. There are ongoing activities at Microsoft to merge their Bing internet search engine with the capabilities of ChatGPT.[16]

How Businesses Can Benefit from ChatGPT and other Large Language Models

Business Records Management

The most valuable aspect of large language models is their unparalleled set of abilities for business applications, such as identifying named entities (NER) within documents and aligning business records to a single entity, supplier, or customer. Such record alignment enhances business activities and minimizes duplication of effort.

Chatbot Virtual Assistants

Straits Research estimates that Chatbot market growth is projected to reach USD 3.62 Billion by 2030, growing at a CAGR of 23.9%. The Q&A capabilities of large language models, combined with their ability to create human-like text, also makes them useful as chatbots to augment customer service operations.

Enhance AI / ML with Faster Labeling of Data

The ability of large language models to summarize large volumes of text data is invaluable in structuring what was previously unstructured text. The newly structured text can be used in existing downstream machine learning models within a company or deployed in new machine learning methodologies that benefit the business. This approach can save time and money, negating the laborious process of hand-labeling and organizing unstructured data, ultimately brining additional value to the business from existing data sources.

Improve Customer Satisfaction and NPS Scores

Large language models have shown outstanding performance on sentiment analysis tasks. This capability can be leveraged to extract customer sentiment from text verbatims captured during customer service interactions and posts on social media platforms. These models can aid businesses in assessing and improving their customer satisfaction (CSAT), net promoter score (NPS), and customer dissatisfaction (DSAT) metrics to reduce churn and improve customer retention.

Drawbacks of Large Language Models

However, there are also concerns about the potential drawbacks of large language models when using them to create factual statements. One major concern is the potential for models to perpetuate bias and discrimination present in the data used to train them.[17] Additionally, large language models are known to occasionally generate inaccurate or misleading outputs, which can lead to problems in various fields, including journalism, education, and healthcare.[18] As a result, it is important to fact-check large language model outputs prior to usage to ensure accuracy and alignment with the desired goals.

What's Next for Large Language Models

Despite these challenges, large language models have the potential to revolutionize the way we interact with and process language. They have numerous capabilities and applications in language-based tasks across industries and verticals. As the technology continues to advance, it is crucial to develop responsible and ethical approaches to developing and deploying large language models to ensure their beneficial impact on society. Their continued development will undoubtedly lead to exciting new possibilities as their capabilities and applications expand in the future.

REFERENCES

1. Gao, et al. "The Pile: An 800GB dataset of diverse text for language modeling." 2020, arxiv.org/abs/2101.00027.

2. Laurençon, et al."The BigScience ROOTS corpus: A 1.6TB composite multilingual dataset." Conference and Workshop on Neural Information Processing Systems Conference (NeurIPS) 2022.

3. Vaswani, et al. "Attention is all you need." Conference and Workshop on Neural Information Processing Systems Conference (NeurIPS) 2017.

4. Brown, et al. “Language models are few-shot learners.” Conference and Workshop on Neural Information Processing Systems Conference (NeurIPS) 2019.

5. Alvi, et al. “Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, the world’s largest and most powerful generative language model.” 2021, https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

6. Facebook AI Research. “Democratizing access to large-scale language models with OPT-175B” 2022, https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/

7. BigScience Workshop, et al. “BLOOM: A 176B-parameter open-access multilingual language model.” 2022, arxiv.org/abs/2211.05100

8. Open AI Research. “ChatGPT.” 2022, https://chat.openai.com/

9. Wang, et al. “GLUE: A multi-task benchmark and analysis platform for natural language understanding.” International Conference on Learning Representations Conference (ICLR) 2019.

10. Wang, et al. “SuperGLUE: A stickier benchmark for general-purpose language understanding systems.” Conference and Workshop on Neural Information Processing Systems Conference (NeurIPS) 2019.

11. Rajpurkar, et al. “SQuAD: 100,000+ questions for machine comprehension of text.” Proceedings of the Conference on Empirical Methods in Natural Language Processing Conference (EMNLP) 2016.

12. Bowman, et al. “A large, annotated corpus for learning natural language inference.” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2015.

13. Williams, et al. “A broad-coverage challenge corpus for sentence understanding through inference.” Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) 2018.

14. Sun, et al. “DREAM: A challenge dataset and models for dialogue-based reading comprehension.” Transactions of the Association for Computational Linguistics (ACL) 2019.

15. Ouyang, et al. “Training language models to follow instructions with human feedback.” 2022, arxiv.org/abs/2203.02155.

16. Kan, M. “Free Sydney? Don’t worry, longer chats will return to Bing, Microsoft says.” 21Feb2023, https://www.pcmag.com/news/free-sydney-dont-worry-longer-chats-will-return-to-bing-microsoft-says

17. Heikkilä, M. “How OpenAI is trying to make ChatGPT safer and less biased.” 21Feb2023, https://www.technologyreview.com/2023/02/21/1068893/how-openai-is-trying-to-make-chatgpt-safer-and-less-biased/GPT factual errors

18. Stern, J. “What is ChatGPT? What to know about the AI chatbot.” 17Feb2023, https://www.wsj.com/articles/chatgpt-ai-chatbot-app-explained-11675865177

Learn more
about Dell HPC

Contact a
Dell Technologies Expert

Join the
HPC Community

Join the
conversation

DELL Technologies