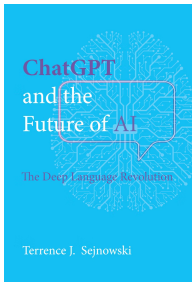
Review of [1]



ChatGPT and the Future of AI
The Deep Language Revolution
Terrence J. Sejnowski
The MIT Press, 2024
$22.95 Paperback, $22.99 eBook, 264 pages

Review by
Nicholas Tran (ntran@scu.edu)

1 Overview

Aimed at the general audience, this book is a field report on the current state of the art in artificial intelligence, focusing on the technical development of large language models (LLMs) and their potentially profound impact on society and humanity. It presents a fascinating demonstration of the capabilities and limitations of modern chatbots such as ChatGPT through a series of conversations with them about their own nature and about the book itself (e.g., there is a ChatGPT-generated summary at the end of each chapter). An early pioneer in neuroscience, the author provides a front-row view of the development neural networks, leavened with engaging personal anecdotes as well as his unique perspective on the future of AI research and its synergy with brain research.

2 Summary of Contents

Preface introduces large language models (LLMs) as intelligent chatbots that can assist with a wide range of language tasks previously performed only by humans. Sample unedited dialogs with ChatGPT are presented to showcase its capability to gather and summarize vast amounts of data *about LLMs* in conversational form, which may contain errors. Their performance prompts the question of whether LLMs understand or are conscious and leads the author to explore the nature of intelligence and consciousness.

Part I: Living with Large Language Models

Chapter 1 explains that the arrival of LLMs marks a new stage in the Information Revolution when machines can form internal representations of data provided to them, which is then used to generate novel human-like responses to novel queries. This new technology will likely to produce scores of new jobs and industries as its predecessor, the Internet, has.

Chapter 2 provides a glimpse of how LLMs might revolutionize existing industries, such as generating and analyzing doctors' notes in healthcare, reducing teacher workload in education, drafting contracts in law, analyzing sentiments in natural language processing, generating images, music, videos, or even digital influencers in entertainment. It also notes concerns that must be addressed in such applications, such as bias, misinformation, and loss of privacy or security.

Chapter 3 presents four remarkable conversations with LLMs that expose their sophistication as well as their weaknesses. The two conversations with LaMDA by Google could be interpreted as evidence of its ability to model human social interactions and to explain and defend its sentience. The other two conversations with ChatGPT by OpenAI portray a chatbot going off-the-rails with its profession of love to the human interlocutor and its authoritative-sounding answers regarding non-existent events or impossible scenarios.

Chapter 4 explains the outcomes of the four conversations in Chapter 3 in terms of priming, the process of providing LLMs with context and examples (called prompts) to limit their responses to those derived from a particular subspace of their training data.

Chapter 5 explores various views of intelligence, thinking, and consciousness and tests some of them against LLMs. It speculates that the mirroring of their interlocutors by LLMs could be an important component of intelligence and suggests basing future theories of intelligence, thinking and consciousness on a mathematical theory of LLMs.

Part II: Transformers

Chapter 6 summarizes the evolution of neural network models, beginning with the perceptron modeling a single neuron, to multilayer networks using backpropagation, to deep learning networks with hundreds of layers. Transformers are deep learning neural networks for language processing that transform words to numeric vectors to capture semantic, positional and contextual information among them. They are trained to predict the next word in a sentence based on the previous words. The author explains the architecture of transformers and their training process, drawing parallels with the evolution of the human cerebral cortex.

Chapter 7 attributes the unexpected success rate of the stochastic gradient descent algorithm used for training neural networks, even when overparameterized, to different dynamical properties of high-dimensional spaces. The author suggests that developing new mathematics for high-dimensional spaces will be useful to understanding LLMs and brains alike.

Chapter 8 surveys the current rush to invest in data centers, networks, and research and development of chip technologies to support the energy-intensive training and deployment of LLMs. The author predicts a law similar to Moore's Law will reduce the energy demand of LLMs exponentially over time and points to the human brain as proof of the possibility of portable LLMs.

Chapter 9 discusses the anxiety and uncertainty among experts in artificial intelligence regarding the possibility of creating a superintelligence that could lead to human extinction. The author points out the drawbacks of *not* continuing AI research and suggests that humanity has learned to manage the risks of new technologies in the past and can do so with AI as well.

Chapter 10 reviews inchoate efforts by the European Parliament and US Congress to regulate AI, a task that is complicated by the competing interests of different stakeholders. Issues such as copyrights and authorship of LLM-generated content will need to be addressed in the future through discussions or in court.

Part III: Back to the Future

Chapter 11 reviews the ascendance of the neural network model over the symbolic model in AI research and argues that the success of the former is due to its similarity in architecture to the human brain.

Chapter 12 proposes new features for LLMs, drawing inspiration from brains, such as long-term memory, working memory, sensorimotor systems, a reward system and others.

Chapter 13 discusses the synergy between AI and brain research, suggesting that the former can aid the latter by providing new tools for analyzing brain data and by proposing new theories of brain function. The author also suggests that the brain can inspire new AI models.

Chapter 14 reviews early scientific advances obtained with or inspired by LLMs at the dawn of AI and predicts rapid progress ahead. The author calls for a scientific theory of LLMs that can explain their behavior and guide their development.

Afterword briefly mentions a recent development in the transformer architecture. **Acknowledgements** thank the author's colleagues, students, and collaborators, including ChatGPT.

3 My Opinion

This book has two aims: to provide a popular account of the development of large language models and to demonstrate how they can assist in carrying out a deeply human task, namely, writing a book. I believe the author has achieved both objectives admirably. Chapters 3 and 4 are especially useful to the general reader in understanding how to use LLMs effectively, while Chapters 6 and 7 do a commendable job in explaining in plain English the architecture of transformers and why they succeed beyond expectations. The author's personal anecdotes and reflections on the future of AI research and its synergy with brain research are engaging and thought-provoking; they are welcome additions to a growing body of retrospectives by pioneers in the field.

The book reads like a first draft at places, especially the unedited end-of-chapter summaries by ChatGPT. I think the lack of editing is intentional to highlight the technology's limitations. Additionally, I would have preferred to see the dialogs differently, e.g., **Dialog 12.2** instead of **GPT 12.2**, since some of them are with LLMs other than ChatGPT and the numbering could be confused with the OpenAI chatbot's version number.

I highly recommend this accessible book to those wishing to gain a good overview of neural networks and machine learning for further reading on the subject.