Chapter 1
Artificial Intelligence

Alessandro N. Vargas

Abstract: This chapter examines the evolution of artificial intelligence (AI) in language modeling, from early deterministic methods to probabilistic approaches and modern neural networks. The chapter revisits key concepts such as Markov chains, the Turing test, and natural language models. It discusses AI's societal impact and applications in language processing.

1 A new reality: artificial intelligence

So far, we have been told that only people can read, write, talk, listen, and communicate through language. These abilities characterize us as human beings. However, during the last few decades, researchers have succeeded in transferring some of these human traits—at least partially—to machines, in an incessant quest to make machines seem '*alive*.' As an idea largely propagated in sci-fi movies [145], a machine incorporating '*human intelligence*' seems impossible, even though what constitutes intelligence over machines has been a topic of intense debate [18, 32, 83, 102, 143].

The idea of machines having intelligence dates back to the 1950s, when a British mathematician named Alan Mathison Turing proposed a method to assess whether a machine is intelligent. Turing asked a person to become an evaluator, and the evaluator's task was to chat with an entity that could be either a computer or a person. That chat occurred only through text because Turing wished to evaluate only the linguistic abilities of the machine. The entity then generated pieces of text according to the evaluator's request.

Once the chat was completed, the evaluator guessed whether the piece of text came from a computer or a person. If the computer could systematically fool the evaluator, then it would pass the test and be considered '*intelligent*.' The test did not

require that a computer produces correct answers; what Turing investigated was the computer's capability of generating human-like content [47].

Turing test to determine what is an *intelligent machine* has raised a lot of controversy over the years because it does not consider certain traits like creativity and consciousness [47]. Even if we boil down the discussion to a specific area, for example, semantics and linguistics, we keep seeing strong debate [144]. This debate has its origin in understanding what defines human reasoning (e.g., [143, Ch. 10]).

If we turn our discussion into 'reading,' we can present some insights. As we all know, to grasp a sentence's meaning, we must get information about the writer's intention. A clue that helps us get the writer's intention is to read the entire sentence and connect it to a broader context. This context usually depends on what appears before and after the sentence.

We reason in this way. We create meaning by connecting information back and forth. Sometimes the reasoning follows in a serendipitous, inexplicable way. However, reasoning produced by machines is incomplete—as researchers claim. According to them, machines can pass the Turing test but do not have a conscious mind [47, p. 121], [144]. Still, a point of consensus is that the Turing test inspired many researchers to create smarter machines, thus increasing machines' abilities to mimic human language.

Not too long ago, researchers attempted to mimic human language through models that accounted for deterministic rules, like *if-then-else*, for parsing sentences (e.g., [62]). They soon realized, though, that language is too complex to be represented by a set of deterministic rules. Words may have two or more meanings. Sentences can be ambiguous. Ideas can be complex. As a result, computers running under deterministic rules were unable to capture the rich nuances and subtleties of language [99].

The rigidity of deterministic rules led researchers to think differently. In the early 90s, researchers changed how they approached language models: they shifted their investigation to statistics and probability models [143, Sec. 1.3.6]. Researchers realized that the nature of language is varying and random, and a probabilistic approach could better incorporate the randomness of language into machines (e.g., [39, 83, 108, 128]).

Researchers' shift to a probabilistic approach proved to be a wise move [143, Sec. 1.4]. Before presenting arguments in favor of this claim, let us first recall certain historical aspects of computer sciences.

In 1956, the term *artificial intelligence* (AI) was used for the first time in a two-month workshop held at Dartmouth College in Hanover, New Hampshire. The event was organized by John McCarthy, Marvin Minsky, Nathaniel Rochester, Claude Shannon, and others [143, Sec. 1.3.1]. The event aimed to bring together researchers to discuss *how machines can become smart*. Since that seminal workshop, AI has taken two steps: (i) from deterministic models to probabilistic models; and (ii) from probabilistic models to data-driven models; see the monograph [143] for a comprehensive history review of AI.

The term 'AI' has been used to refer to any machine that seems smart, often leading to misinterpretation [66, p. 1319], [83], [143]. Even so, AI is popular—it

is featured everywhere. News media provided widespread coverage of AI and its applications [32, 92], which has helped AI gain popularity [107].

Some news agencies and media organizations express fair concerns about AI and its use. For example, media agents often engage in heated discussions about the use of AI, even recalling ideas from sci-fi movies that depict machines taking over humans [107]. Although this scenario seems unrealistic, we have seen more and more cases in which AI approximates or even surpasses human competence [116].

Perhaps one of the most memorable milestones of AI surpassing human competence happened in 1997: IBM supercomputer Deep Blue, powered by an AI model, defeated the chess master, former world champion Garry Kasparov [73]. The defeat of Kasparov brought worldwide attention to the potential of AI for industrial applications. Since then, AI models have become more complex, requiring more investments in research and technology.

To expand AI applications, industries have invested around U$ 160 billion, a budget estimated for the year 2021 alone [184, Fig. 2, p. 3172]. Industries have then deployed AI applications to dominate their respective markets, in ways we hardly perceive [107]. For instance, `Amazon.com` uses AI to tailor specific products according to consumers' visited web pages; `Spotify.com` uses AI to create a list of songs that fits consumers' musical preferences; `Netflix.com` uses AI to suggest films that grab the viewer's attention; `Youtube.com` uses AI to maximize the number of viewers according to video content; `PayPal.com` uses AI to detect fraud in transactions; `Waymo.com` uses AI in autonomous vehicles; `Uber.com` uses AI to optimize distance and location between drivers and passengers; see [89] for the implications of AI for today's society, and [143] for other technical aspects of AI.

Industries have been laying off workers, diminishing their workforce, as certain AI applications can do the work previously done by humans (e.g., [92, 175]). More and more AI applications with the potential to replace humans have been reported, such as in speech recognition [69], face recognition [102], language translation [63], ancient-text recognition [98], command-by-voice activation of services [150], and even unveiling new drugs [66]. AI has reached a point where it can generate creative content on its own, like poems, stories, and music [26, 56, 128]. Curiously, all of these activities were once considered exclusively human.

The point of this book is not to claim AI as an enemy. This book's goal is to show how scientific authors can benefit from AI, as detailed in the sequence.

2 How language models work (and why it matters)

Language models are mathematical tools that help us understand how words and sentences stick together to form a meaningful piece of text. Language models have found applications in different fields, all of which attempt to decipher how people interpret and process information [18, 56, 176].

Where did the idea of a language model come from? To answer this question, we revisit an event that happened more than a century ago.

In 1913, Andrey Andreyevich Markov, a brilliant Russian mathematician, decided to play with one of the greatest novels of Russian literature. Markov came up with a curious conjecture: what if probability could be used to explain the statistical regularities of language? Markov was making progress on the theory of probability while contributing to real-world problems, like that of deciphering language [11].

To test his idea, Markov took a copy of the verse novel "Eugene Onegin," a masterpiece of Russian literature by Alexander Pushkin, published between 1825 and 1832. Markov took the first twenty thousand letters from the verse novel and wrote them down in a long string without any spaces or punctuation marks [11, p. 19]. Next, he classified the letters into two groups: consonants and vowels. His idea was to identify whether the letters followed any pattern—how consonants and vowels were distributed in the text.

After he annotated a letter, he observed that the probability of seeing the subsequent letter as a *consonant* depended on whether the current letter was a *consonant* or a *vowel*. In other words, the probability of observing a pair of letters, such as *consonant–consonant*, *consonant–vowel*, *vowel–consonant*, and *vowel–vowel*, was quite different from the probability of picking up a random *consonant* or *vowel*. What he found was that the chance of seeing a *consonant* or *vowel* depended on what was the previous letter. In summary, Markov saw dependence among successive letters.

Although this fact may seem intuitive to us today, it was not known at that time. Markov reported this finding in 1913 to the Royal Academy of Sciences, St. Petersburg, and published his findings in the paper [109].

Markov's finding was striking because it pointed out that the probability of an event happening now might be linked to what happened immediately before. In precise terms, Markov property (also known as '*memoryless property*') states that the entire history of previous events can be discharged because what matters in terms of probability was the previous event and the current one.

Markov formalized this finding in a probabilistic structure—this structure has been called simply as the *Markov chain* (e.g., [11, 19, 169]). A Markov chain takes values from a finite set, and the elements of this set are associated with known probabilities. For the particular study of letters, the set contains only two elements: *consonant* and *vowel*.

Markov's findings opened up new horizons for the scientific community—the discovery of the Markov chain changed how researchers interpret probabilistic events [74]. Markov chains have reached widespread use in today's society, even though we might be unaware of that. For instance, Markov chains have enabled Google to implement its search algorithm [178]; in other words, Markov has indirectly contributed to Google's success [170]. Other applications of Markov chains can be found in chemistry [14], music [19], biology [104], weather forecast [155], speech recognition [88], text generation [172], among others [169].

At this point, you might be questioning why it is necessary to discuss Markov's discoveries in this monograph. That is a fair question. Recalling Markov and his discoveries, we can contextualize better how researchers have used Markov's find-

ings to model the patterns of language. Attempts in this direction have consolidated a research area called *natural language modeling*.

3 Natural language modeling

Methods for natural language modeling emerged as a direct consequence of attempts to mimic the human ability to produce and interpret texts. During the 1980s, researchers considered the assumption that natural language obeys certain rules, and these rules could be translated into the underlying Markov chain. The literature contains thousands of papers documenting researchers' attempts to handle words and groups of words as elements of Markov chains [39, 86, 108]. However, they soon realized that Markov chains were inadequate to capture the intricacies and nuances of natural language.

In the middle of the 1990s, researchers moved investigation from Markov-chain models to a newer approach called *neural networks* [143, Sec. 1.3.5], [57]. Despite the name, 'neural networks' have nothing to do with the biological structure that makes our brains work. It has been a jargon created by computer scientists to refer to a particular nonlinear model [85].

This nonlinear model contains many scalar-weighted numbers that are interconnected to generate nonlinear functions. These numbers have received the fancy name '*neurons*,' and a model can have as many neurons as the computer scientist who is programming the model wishes [85]. The procedure computer scientists follow to determine those scalar-weighted numbers is called '*training*.' As expected, training requires a large amount of data and computer processing.

When applied to texts, neural networks tend to produce unnatural outcomes, even when the training considers groups of N consecutive words, where N is greater than one [57, Ch. 13, p. 158]. While increasing N can improve the results a little bit, it can create certain drawbacks. For example, increasing N makes the model more complex, requires more data, and consumes more computing power.

Despite the advances, researchers have not limited their investigations to neural networks. Indeed, they have been creating other kinds of sophisticated models that could benefit from the huge amount of data and texts available on the internet; see [39, Ch. 7.4], [2, 56, 108, 172] for further details; see also [143] for a comprehensive review of these models.

Note that these models have been categorized under tech-like names, such as '*machine learning*' [56, 83], '*deep learning*' [102], '*reinforcement learning*' [176, 139], '*transformers*' [128], etc. All of these names and their corresponding models have been categorized by the news media simply as '*artificial intelligence*' (AI) [143].

3.1 Application of AI: ChatGPT

One of the most striking advances in artificial intelligence (AI) was made public in November 2022 by a research company called OpenAI, based in San Francisco, CA. Founded by Sam Altman, Elon Musk, and other tech leaders, OpenAI created a groundbreaking language model known as the Generative Pre-trained Transformer (GPT). OpenAI later refined this model for interactive use through a chat-based system, releasing it as a product called ChatGPT [128].

OpenAI initially launched ChatGPT in version 3.5, free of charge, at `chat.openai.com`. Subsequently, newer versions of ChatGPT were introduced, though the most advanced version is available only via a paid subscription at the time of writing this chapter.

ChatGPT functions as a text-based conversational tool: users input queries, and the model generates responses based on the provided input. The chat interface is user-friendly, minimalistic, and retains a history of prior conversations.

ChatGPT's capabilities are far from simple. In fact, its underlying model is immense in size and astonishing in scope. It was trained on hundreds of billions of words, apparently sourced from the internet, resulting in a model with approximately 175 billion parameters in version 3.0 [49], and an estimated 100 trillion parameters in version 4.0 [18]. Within two months of its release in November 2022, ChatGPT had attracted 100 million users [164], an extraordinary milestone.

Training ChatGPT's model demanded substantial investment. For instance, Microsoft provided OpenAI with 1 billion in funding in 2019, followed by an additional 10 billion in 2023 to support its development, according to news reports.

ChatGPT demonstrates remarkable writing capabilities, generating sophisticated content that often appears indistinguishable from human-written text [65, 124]. In one study, scientists utilized automated code powered by ChatGPT to analyze data and independently write a scientifically sound paper [30], marking a significant achievement.

ChatGPT also outperforms humans in certain knowledge assessments. For example, it scored higher than 90% of test-takers in the USA Uniform Bar Examination, a crucial test for attorneys [128]. Additionally, it achieved 60% accuracy across a series of medical exams required for physician licensure in the United States [94].

Even more surprisingly, ChatGPT surpasses humans in identifying and describing people's emotions [42, p. 5]. This intriguing ability could encourage clinicians to adopt ChatGPT as a support in mental health treatment.

This book delves deeper into how ChatGPT can enhance scientific writing and its broader applications.

3.2 Transformers

A white paper distributed by OpenAI (not peer-reviewed) sparked significant excitement within the AI research community; this paper introduced the model known

as *transformers* [168]. According to the authors [168], transformer models greatly enhance the capabilities of neural networks by employing three key components: an *encoder*, a *decoder*, and *attention* mechanisms.

The encoder processes input text, converting it into numerical representations. The attention mechanism ensures both meaning and context are preserved by identifying relevant parts of the numerical representation and discarding less critical ones [48]. This information is stored in multiple layers, allowing for intricate processing. Finally, the decoder generates the desired output text by interpreting the processed data.

It seems that transformer models were first described by Google researchers in 2018, within a model called BERT [33]. Still, transformer models helped Google's competitor, OpenAI, to leverage ChatGPT-3.5 (see Section 3.1). OpenAI claims that ChatGPT is a large language model [128], and the 'T' in its name stands for *transformers*.

In conclusion, modern language models rely heavily on probabilistic principles, whether derived from Markov chains or other methods. Unsurprisingly, products like ChatGPT exhibit behavior akin to a Markov chain, where words are concatenated as if originating from a sequence of probabilistic events. The next chapter presents evidence suggesting that ChatGPT's text generation follows a pattern consistent with random chain sequences.