HOW TO TALK TO AI

(AND HOW NOT TO)

JAMIE BARTLETT

WH ALLEN

UK | USA | Canada | Ireland | Australia
India | New Zealand | South Africa

WH Allen is part of the Penguin Random House group of companies
whose addresses can be found at global.penguinrandomhouse.com

Penguin Random House UK
One Embassy Gardens, 8 Viaduct Gardens, London SW11 7BW

penguin.co.uk

Penguin Random House UK

First published by WH Allen in 2026

007

Set in 12/15.5pt Bell MT Pro
Typeset by Six Red Marbles UK, Thetford, Norfolk

Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

The authorised representative in the EEA is Penguin Random House Ireland, Morrison Chambers, 32 Nassau Street, Dublin D02 YH68

A CIP catalogue record for this book is available from the British Library

ISBN 9780753561980

Contents

Chapter 1: The Rise and Rise of the 'Large Language Model'

The dream of intelligent machines

IT TOOK A LOT of time and work for these mysterious talking machines to become an overnight success. In the summer of 1956, a small group of researchers gathered at Dartmouth College to explore a radical idea: whether intelligence could be programmed into a machine. They believed human reasoning, or something approximating it, could be broken down into logical steps and rules. There was no reason a computer couldn't replicate it.

This approach is often called 'symbolic AI'. Engineers would make computers intelligent by teaching them ever more complicated rules about the world. If x, then y. And for certain, narrow tasks this worked amazingly well. Early translation software, medical diagnostic systems and IBM's Deep Blue, which defeated world chess champion Garry Kasparov, were all built using this approach.

But symbolic AI often ran into trouble when

confronted with messy, ambiguous real-world reality. Especially language. You could teach a computer to understand 'I'm hungry'. But what about 'I could eat a horse', or 'My stomach's rumbling' or 'Oh great, *another* salad'? Even for the most powerful machines, trying to program in nuance, context, humour, irony, cultural nuance, was too much.

Another group of researchers looked at the problem a different way. The brain is not a book of rules, they figured, but a dense web of neurons that strengthen or weaken their connections in response to experience. When a baby coos at her mother and the mother coos back, neurons fire and the connections between them strengthen. Do this a thousand times and the baby learns that certain sounds bring comfort. These 'connectionists' believed machine intelligence might emerge in the same way and began designing something called a 'neural network'. Instead of programming rules, feed a machine examples and let it figure out the rules itself. Show it a million pictures of cats – tabby cats, fat cats, sleeping cats, angry cats – plus a little guidance and some clever maths known as 'back-propagation', and let the machine infer what 'cat' means.

Although there were overlaps, for years, symbolic AI was the dominant approach. But around the

turn of the century, two things changed. First, the cost of computing power fell dramatically. Second, the internet appeared, bringing almost infinite words and images with which to train these neural networks. Suddenly the connectionist approach could be tested at scale.

In 2012, a British-Canadian cognitive psychologist called Geoffrey Hinton, along with two of his students, entered a neural network into an image recognition competition. It won by such a wide margin that everyone realised this approach wasn't crazy after all. Researchers started to explore more, and in 2017 a group of engineers at Google developed something called the 'transformer'.[1] This clever little machine could process language in a new way, by looking at the relationship between every word across entire paragraphs, working out which part was the most important. AI researchers realised this could improve the performance of neural networks, and allow them to scale faster.

From that moment, there was only one game in town to develop machine intelligence: bigger data sets, more computing, more training and a dream that some form of intelligence might emerge. This pattern-matching approach explains why these

systems work, and why they fail in the strange ways they do.

The arrival of the large language model

When OpenAI was founded in 2015, it wasn't planning to create the world's most popular chatbot. It was aiming much higher: to build 'Artificial General Intelligence' before Google, solve humanity's greatest challenges and make it available to all. An LLM was just one of many speculative experiments the company ran to get there. The hypothesis was simple: feed a transformer-based neural network enough text and see what comes out.

When OpenAI released GPT-3 in 2020 – probably the moment you first heard of these new machines – even the researchers who'd built the model were stunned. It could write coherent essays, hold conversations and answer questions with a fluency that seemed impossibly human-like for a machine. When they released ChatGPT two years later, it became the fastest-growing consumer app in history, reaching 100 million users in two months. To many people it felt like machine consciousness had fallen from the sky.

An arms race began immediately, and the

number of language models multiplied. Google launched Gemini. Former OpenAI researchers, increasingly worried about the safety of these systems, started Anthropic and built Claude. Meta released Llama as open source. Elon Musk (one of the original OpenAI founders) built Grok. Perplexity offered integrated web search, making its answers up to date (some models are 'locked' at the moment of public release). Chinese companies launched DeepSeek and Qwen – which were far cheaper to make than their American rivals. Smaller firms began building applications on top of these base models: creating specialised bots for specific tasks, from legal advice to 'companion' bots.

Most of these companies now believe they are in a winner-takes-all race – to be better than their rivals, maybe even to reach Artificial General Intelligence. The result is a mad dash for more investment, more data, more computing power. Each model is regularly updated: better reasoning, multimodal functions (text to image, text to video), better memory, fewer errors, new safety filters. Because of this race, models are sometimes developed and released before the social or economic consequences are thought through. The result is that we users are part of a fast-moving mass experiment, and no one

knows how it will end. There are many other types of AI out there – protein-folding machine-learning algorithms, specialised chess-playing algorithms, vision systems for self-driving cars – but now for most people, the LLM *is* AI. It is already woven into email, customer service systems, search engines, creative editing software, and their applications and uses are multiplying every month.

How to build a smart machine

Building an LLM from scratch is incredibly expensive, complicated and involves some of the smartest statisticians and engineers in the world. But it can be boiled down to a few simple steps that are worth you understanding.

First comes 'pretraining'. The model is given an enormous dataset, sucking up more words than the human brain can comprehend: books, websites, academic articles, Wikipedia entries, blogs, Reddit posts. Everything the engineers can legally, or sometimes maybe illegally, get their hands on. Some models are trained on hundreds of billions, maybe even a trillion, words. No one knows, because the companies don't say what goes into their 'training data'. The task is deceptively simple: given a sequence of words, try to predict what comes next.

(Technically speaking, they don't think in words, but 'tokens', which are words or pieces of words.) The model makes a guess. Gets it wrong. It adjusts its internal parameters and tries again. The model does this trillions of times and gradually builds a gigantic multidimensional linguistic universe which maps the relationship between every word and every other word.

What emerges from this pretraining phase is often chaotic and offensive. (Like the internet.) The model is then refined through human feedback, testing and tuning. Most use a process called 'reinforcement learning from human feedback'. Thousands of human annotators – often sitting in offices in the Philippines or Kenya – spend their days rating the model's answers. It's painstaking, sometimes traumatic, work. But gradually the model learns to produce responses that humans tend to prefer: more helpful, more polite and less likely to create a lawsuit.

Finally, safety and 'alignment' layers get added. Dedicated teams write policies about what the model should and shouldn't say, and design filters and rules to stop the model pumping out hate speech, self-harm content and other offensive or illegal material. Given most of its training data is from the internet, that's a big job.

And after all that, you have your LLM. There are important differences between them: for example, models created by Anthropic are widely considered to be safer. Some are better at certain tasks than others. But all are massive, sophisticated, next-word prediction machines with a friendly interface. Capable of remarkable feats and bizarre failures.[2] You type, the machine predicts and something comes out that usually looks fluent.

What they can and can't do

With enough data and computational power, pattern matching can produce astonishing results. LLMs are tested on an ever-changing battery of 'benchmark' exams: general knowledge, coding challenges, reasoning puzzles. On the brutal 'Massive Multitask Language Understanding' exam, which covers 57 varied subjects, most models now exceed expert human performance. They do similarly well on tests examining coding skills, mathematical reasoning, common sense. They pass the bar and medical exams with flying colours. (And at a speed unimaginable to a human.)[3] The leading models – Claude Sonnet 4.5, ChatGPT-5, Google's Gemini 3 – are usually separated by just a few percentage points.

But these formal benchmark tests don't quite capture what's going on. For years the litmus test for machine intelligence was the Turing Test. This famous thought experiment, created by Alan Turing in 1950, said that a machine could be considered to have human-like intelligence if, over the course of a conversation, a human couldn't tell it apart from a fellow human. According to some analysts, one model recently passed a modern version of that test.[4] But these models are geniuses one moment and fools the next.

The same design that makes LLMs powerful also makes them imperfect. They don't 'know' things the way you or I do. If I ask you the capital of France, you retrieve a fact stored in your memory. You have a mental model of what capital cities are and how they fit into the order of things. Maybe you've been to Paris. An LLM does not have a theory of the world like this and doesn't maintain an official database of capital cities that it refers to when asked. Instead, it calculates that 'Paris' is the statistically most likely word to follow the question, based on the patterns it's seen. (To be precise: the model predicts a series of highly likely next words and then samples from across them. This introduces a degree of randomness into every output, which is