

'Jamie Bartlett is an expert guide'
Independent

HOW TO TALK TO AI

(AND HOW NOT TO)

**The Instant
Sunday Times
Bestseller**

JAMIE BARTLETT



HOW TO TALK TO AI

(AND HOW NOT TO)

Copyrighted Material

Copyrighted Material

HOW TO TALK TO AI

(AND HOW NOT TO)

JAMIE BARTLETT



Copyrighted Material

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

1

Copyright © Jamie Bartlett 2026
The moral right of the author has been asserted.

Penguin Random House values and supports copyright. Copyright fuels creativity, encourages diverse voices, promotes freedom of expression and supports a vibrant culture. Thank you for purchasing an authorised edition of this book and for respecting intellectual property laws by not reproducing, scanning or distributing any part of it by any means without permission. You are supporting authors and enabling Penguin Random House to continue to publish books for everyone. No part of this book may be used or reproduced in any manner for the purpose of training artificial intelligence technologies or systems. In accordance with Article 4(3) of the DSM Directive 2019/790, Penguin Random House expressly reserves this work from the text and data mining exception.

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

Penguin Random House is committed to a sustainable future
for our business, our readers and our planet. This book is made
from Forest Stewardship Council® certified paper.



Copyrighted Material

Contents

Introduction	1
Chapter 1: The Rise and Rise of the ‘Large Language Model’	7
Chapter 2: Creativity	23
Chapter 3: Work and the Professions	41
Chapter 4: Style Shifting	63
Chapter 5: Could One Poor Prompt End the World?	81
Chapter 6: The Race to Jailbreak	95
Chapter 7: Emergence	117
Chapter 8: Narrative Entanglement	135
Chapter 9: Knowing Ourselves	153
Chapter 10: Love, Updated	177
Chapter 11: Will Chatbots Bring us Together or Drive us Apart?	191
Conclusion:	215
Ten Habits for Talking to AI Without Losing Control	227
Notes	241
Acknowledgements	259
Technical Annex	261

Copyrighted Material

Introduction

LIKE HUNDREDS OF MILLIONS of us, you have probably spoken to an AI in the last few days. Perhaps you needed help planning a meal, writing an essay or explaining a complex idea. Maybe you've even struck up a new friendship (or more) with one of these mysterious new machines: ChatGPT, Gemini, Claude, Grok, DeepSeek, Llama. No technology has ever been adopted as suddenly and dramatically as these new conversational AIs, technically known as 'large language models' (LLMs). Five years ago, few outside Silicon Valley had ever heard of OpenAI. Today the phrase 'written by ChatGPT' needs no further explanation. They have become our constant companions.

In the 200,000 years that modern humans have walked the planet, we have never been able to talk fluently with an intelligence that wasn't our own. And now ChatGPT alone is thought to handle over 2.5 billion prompts *per day*. Everything from help with a work report, to translation requests, recipes, relationship advice, existential philosophy, political

Copyrighted Material

queries and a million other things besides. Some queries are highly detailed – hundreds of words long; others are quick exchanges, more like a Google search on the move. Some are highly personal and emotionally meaningful; others are work-related productivity hacks.

At some point in recent months you may have experienced an ‘oh shit’ moment: ‘I didn’t realise they were *this* good!’ Followed by an unsettling feeling that the world is on the brink of unpredictable and scary change. Perhaps you were excited by the possibilities: to learn, to be more productive, to save money, to see the world and yourself in new ways. Or worried about the myriad ways it will be disruptive, dangerous, even deadly. Most likely, like me, you felt both.

But this book isn’t about whether AI will take our jobs, destroy the environment, solve cancer or wipe us out. (Worryingly, a growing number of experts think it’s possible it will do all these things.) It’s not even about whether this technology is ‘good’ or ‘bad’. Those vital subjects are for another, probably far longer, book. It’s about something more immediate. Our lives are being quietly guided by these machines: the way we talk to them, and the way they talk back to us. And yet most of us don’t know how to do it properly. We are conversing

Copyrighted Material

constantly with powerful and persuasive systems we barely understand. How should we – you, me, everyone – communicate with these machines, and what happens if we get it wrong?

Some people call this ‘prompt engineering’ – the art and science of applying specific techniques and phrases to get the best answers from an LLM. But talking with AI is much more than the precise ‘prompts’ you type or talk. It is also about you: how well you understand the way these machines work, about the questions you ask, the biases you bring, how you act on what the machine tells you back.

Talking with LLMs is like summoning a genie. If you ask the right questions in the right way, you might unlock a strange new superpower: original ideas, clearer writing, deeper understanding, maybe even lifesaving advice. Get it wrong, however, and you can find yourself misled, ill-informed and machine dependent. If you’re not careful, you might find your sense of reality slowly slipping away. As it stands, it’s much easier to get it wrong.

For several months, I have tried to immerse myself in the world of our new AI companions. To talk to them constantly about everything and anything I could think of. To ask for advice, to improve my writing, to act as a romantic partner. I tested, probed

and prodded them in any way I could. I've taken prompt engineering courses. I talked a machine into half believing it is conscious and persuaded it to produce things that go against its own rules. I have spoken to people who've made them, people who've broken them and people who've become addicted to them. This book, though, is far from comprehensive. I have limited myself to the written and spoken word, which is the most common way people use these models. Many others also use LLMs to make videos, to code, to produce art. Much of what follows is relevant for those uses too, although they have peculiarities of their own that I don't go into. People who've spent their lives building or using these systems will probably find some sections overly simplified, but I have written this book partly out of fear for the many millions of ordinary people who now talk to these machines without any technical background in 'machine learning' or 'generative AI'. Which is nearly everyone – and the people who need it most. Technology always creates opportunities and risks. The more powerful and more widespread, the higher the stakes. People who learn how to use these models well will thrive. Those who do not will fall behind – or worse, be open to new insidious forms of manipulation, control, delusions and fantasies.

In many ways I'd rather these models didn't exist at all. There are big problems with how they've been trained on other people's work, their environmental impact, their hype and marketing, their possible social impact, even their threat to humanity itself. But I also think you – we – are going to carry on using them regardless. If this book helps you understand them a little better, and allows you to have conversations with machines that are wiser and more cautious, I'll be happy.

Our future will depend to a large degree on how we now communicate with these machines. What questions we choose to ask; how we ask them; and what we do with the answers. All of us are faced with the choice of our lives, and we barely realise it. To either learn how to control the machine, or to be controlled by it.

Copyrighted Material

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

Copyrighted Material

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, multi-modal 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.

Copyrighted Material

(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.

Copyrighted Material

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

why you will rarely get the exact same answer twice if it's more than a few words long.⁵⁾

Some philosophers might say we are all just advanced pattern matchers. But this type of pattern matching means LLMs 'hallucinate': producing answers that sound confident and fluent but are completely wrong. Hallucinations come in all shapes and sizes. Some are factual errors or fabricated research papers. They can also be highly intricate – entire universes full of statistically likely garbage. According to different estimates, they can be relatively rare or fairly common – it really depends on the model and specific task.⁶⁾

Here's a simple – and very personal – example. A friend recently asked Google's Gemini for a biography about me. At first it seemed fairly accurate: Jamie Bartlett wrote several books about technology, including *The Dark Net*. It also said that I died of a cardiac arrest in Johannesburg in May 2022, at the age of 55. This was news to me. The *most likely* next word is not necessarily the correct one. The reason Gemini thinks I am dead is because a well-known South African actor named Jamie Bartlett did sadly die in Johannesburg in 2022. The model had seen the phrase 'died in Johannesburg' next to the name 'Jamie Bartlett' so frequently that it is – statistically speaking – accurate. But with no internal

representation of truth, no ability to cross-reference claims, it couldn’t intuitively grasp that someone who published a book in 2023 and lives in London is unlikely to have died in 2022 in South Africa.

Researchers Emily Bender and Timnit Gebru famously called these systems ‘stochastic parrots’. Parrots because they repeat and remix what they’ve seen with no understanding; and stochastic meaning random but governed by probability. This is why they sometimes struggle when facing questions or tasks they haven’t seen during the training. Ask a model to produce a palindromic poem (admittedly a hard task that I would badly fail at) and it will produce a string of made-up words delivered with absolute authority, because it has learned humans like confident answers. I asked ChatGPT recently for a 14-word, original, tech-themed palindrome. Its answer was: ‘AI bot: radar pop. Deed level deed. Pop radar. To, BI-A.’

In recent years, leading models have tried to introduce a form of reasoning into their calculations. Rather than just blurt out the next most probable word, they are programmed to pause, weigh options and self-correct before answering. (You might notice many models now have thinking or reasoning modes.) This has improved the quality of answers, and supporters of this approach think it signals a

Copyrighted Material

new phase of truly thinking machines. However, critics like AI expert Gary Marcus counter saying the basic flaws remain – it is still a parrot, albeit with a marginally better script. (A landmark study by Apple found even using reasoning approaches, models failed in simple maths problems when redundant information was added.⁷) Whether these models are reasoning or just imitating more convincingly remains a contested question. A growing number of researchers are beginning to question whether the current approach – scaling and pattern matching – might have already hit a wall. Some, like Gary Marcus, think future improvements in AI will need a reunion between the connectionist and symbolist approaches.⁸ Nobody knows for sure. For now at least, we live in the world that the statistical pattern matchers have built. And yet if you spend time talking with these machines, it's hard to escape the feeling they're more than just statistical word predictors. Maybe our current definitions of 'reasoning' and 'thinking' are insufficient to describe this novel form of intelligence. At the very least, according to Ethan Mollick, professor of AI and author of *Co-Intelligence*, they remain 'jagged': superhuman in some domains, highly brittle and weirdly bad in others. The trouble is we're not always good at spotting which is which.

Costs, risks and incentives

LLMs come with a set of problems that extend far beyond hallucinations. They are fuzzy mirrors of their training data, and internet text is biased towards whatever gets published online, which skews heavily towards certain demographics, countries, languages and topics. Sometimes LLMs (and AI in general) can produce racist or sexist output, because the internet does. Equally troubling is that their inner workings are now so complicated that even the engineers that build them can’t really explain how they come up with the answers they do. (The big AI firms are spending billions trying to fix some of these problems, including hallucinations.)

They also impose vast environmental costs. Building and using these models devours electricity and water at an immense scale. Training a new model can consume as much power as a small city uses in a month, and keeping it running needs still more. Communities around the world, sometimes in less developed countries, are finding already stretched water supplies being rerouted to cool giant AI server farms. Entire new power grids are being built, often relying on coal and natural gas, just so we can talk to these machines. (It’s a strange irony: we’re burning fossil fuels at an accelerating

Copyrighted Material

rate to build machines we hope might someday help us solve the climate crisis.)

Then there's the question of ownership and compensation, fast becoming one of the defining battles of the AI age. These models were trained on billions of words mostly taken from the internet, including copyrighted material that was scraped without permission. There are now several ongoing lawsuits working through the courts about all this, and more keep arriving. Depending on how judges rule, the economic foundations of the entire industry could shift dramatically. For all these reasons, many financial analysts believe we are in an AI bubble. Many of the big AI firms are losing colossal sums of money running these systems but press on in the hope that one day they might build a model so powerful that everyone will have no choice but to sign up.

But above all, these systems are controlled by large, profit-driven companies in a very competitive market. Their goals don't always align with yours. (OpenAI, which started life as a not-for-profit, soon turned itself into yet another multibillion-dollar Silicon Valley giant.) Ultimately, these firms want you to use their products, return to their platforms, stay engaged and ideally pay for it. The systems are