Name: _____ Date: _____

How Does AI Work?

Reading level 11 (7th grade) · 1472 words

How does AI work? Start with a fill-in-the-blank game. "When I hear rain on my roof, I ___ in my kitchen." Maybe you would say cook soup. In an example table from Google's machine learning course, a language model gives "cook soup" a 9.4 percent chance and "warm up a kettle" 5.2 percent. Nothing was looked up. The numbers come from patterns in writing. Today, when people say AI, they often mean programs that learn from huge numbers of examples and then guess like this. This book follows the guess, from how it is trained to why it can be wrong.

Older programs followed rules that a person wrote. In 2002, Paul Graham tested rules like that for spam. Merely looking for the word "click" caught about 8 in 10 spams, but it also flagged some good mail. Machine learning changes the job. Google's guide says it means training a piece of software, called a model, to make predictions from data. To predict rain, you do not write out the physics. You give it enormous amounts of weather data until it learns which patterns bring rain.

What is inside a model? Numbers. A neural network is built in layers. Each unit takes values from the layer before, multiplies each by its own weight, and adds a bias. Together, the weights and biases are the model's parameters. Picture a huge control board of dials, with training turning each dial a tiny bit. A small network with three inputs and one layer of four units already has 21 parameters, by Google's count. The largest language models have hundreds of billions, or even trillions.

How do the numbers change? Through a loop. A new model starts with random numbers, so its first guesses are poor. It makes a guess, and a score called the loss tells how wrong it was. Then training works out which way to nudge each number to lower the loss, and nudges it a small step. Google's course calls the method gradient descent, and says the loss falls like a ball rolling down a hill. Then the loop starts again with the next example. It repeats an enormous number of times.

Now the chatbot. Common Sense Media's glossary lists ChatGPT, Gemini and Claude as examples of large language models. Such a model learns from a huge amount of writing. It works with tokens, which are small pieces of writing. In English, about 400 tokens make up 300 words. In the first phase of training, the model reads text and guesses the next token, as with the blank at the end of the rain sentence. Meta's 2024 paper on its Llama 3 models describes training this way. Nobody has to write the answers, because the real words are already there. The loss shows how close the guess came, and training adjusts the parameters.

Writing is that same game, played again and again. You type a question. The model rates the chance of each possible next token, picks one, adds it to the answer, and repeats. Each new token becomes part of what it reads before it picks the next. Google's course calls these models "essentially autocomplete mechanisms." The pick can be the likeliest token, or a random choice among the likely ones. That is one reason you can ask the same question twice and get two different answers.

A raw predictor is not yet a helpful assistant. Making a model bigger does not by itself make it better at following what a person wants, wrote OpenAI's team in 2022. So makers train it further. People wrote samples of good answers, and then ranked the answers against one another. Google's course calls a step like this instruction tuning. In that study, people preferred answers from a model with 1.3 billion parameters over answers from GPT-3, which had 175 billion. Even so, the new versions still made simple mistakes.

Why does it sound sure and still get things wrong? Scientists at OpenAI and Georgia Tech argued in 2025 that language models make things up partly because training and testing reward a guess over saying "I don't know," like a student guessing on a hard exam. They add that it keeps happening even in the best systems. NIST, a US standards agency, calls this confabulation: confidently stated but false content. In Common Sense Media's March 2026 survey of 1,204 US kids ages 9 to 17, only 35 in 100 gave the answer it counts as right: AI cannot tell true from false.

Why can it be unfair? It copies patterns in its examples, including bad ones. Common Sense Media says AI trained on data from all over the internet can pick up biased ideas, so its answers can reinforce stereotypes. In 2019, NIST tested 189 face-recognition programs and found that most made different mistakes for different groups of people. For one kind of mistake, the rates often differed by 10 to 100 times. The gaps were bigger for false alarms, where a program wrongly says two photos show the same person, than for misses. Programs have changed since 2019. Testing on many kinds of people is how gaps like these get found.

AI is strong at some jobs and weak at others. Stanford's 2026 AI Index says one AI system earned a gold medal at the International Mathematical Olympiad, yet the top system reads analog clocks correctly just 50.1 percent of the time. That score is from the 2026 report and will change. Scientists call this a jagged frontier. It means that a great result on one kind of task is no promise about the next. The only way to know is to check the answer you actually got, and to ask who could tell you if it were wrong.

AI is older than its tools. In 1950, Alan Turing asked, "Can machines think?" and swapped that question for a game he could test, called the imitation game. In its simplest form, a judge types questions to two hidden players, a person and a machine, and tries to tell which is which. In 1955, a proposal for a summer study at Dartmouth College used the words artificial intelligence in its title. It guessed that every part of learning could be described so exactly that

a machine could copy it. That was only a guess, and people have been testing it ever since.

Two game wins show two ways to build a player. In May 1997, IBM's Deep Blue won a six-game chess match against the world champion, and IBM says it got there by brute force, checking 200 million positions a second. In March 2016, AlphaGo beat a top Go player, Lee Sedol, four games to one. Its networks judged board positions and chose moves, and they had learned first from human expert games and then from playing against itself. One machine mostly searched very hard. The other learned from examples and practice, and searched too.

So does AI think? Turing called that question too meaningless to discuss. Scientists still split on a nearby one: could a model trained only on text understand language? In a 2022 survey of 480 language-technology researchers, 51 percent agreed that, given enough data and computing power, it could understand language in some non-trivial sense, and 49 percent disagreed. When the Nobel Prize explained its 2024 physics award for work on neural networks, its guide for the public wrote that computers cannot think, but that machines can now mimic functions such as memory and learning. The two sides look at the same programs. They disagree about what the good results show.

As of November 2026, here are some plain facts. Common Sense Media says chatbots can be wrong, repeat unfair ideas and show unsafe things. It advises ages 6 to 12 to use them a little, beside a parent or caregiver, and says most tools require users to be 13 or older. What you type may be kept: Common Sense says to avoid giving names, addresses or photos, and a researcher interviewed by UNICEF in 2025 said AI companion apps often collect photos, voice notes and location. Plain advice that goes past those pages: never type your full name, your address, your school or a photo. If an answer matters, ask how you could double-check it.

Back to the blank in the kitchen. A chatbot fills in every answer the way it filled in cook soup: one small piece at a time, from patterns in the writing it learned from. How much understanding, if any, sits behind that is the question scientists still argue about. It is also the question Turing set aside in 1950. Nobody has settled it. As of 2026, the answer you get comes from the same kind of guess, made one token after another, by a program that learned from a very large pile of examples.

Questions

1. What are the parameters of a model?

A. The questions people ask it
B. The weights and biases that training changes
C. The screen it runs on

2. Why can the same question get two different answers from a chatbot?

A. It forgets the question

B. Chatbots never repeat words

C. The pick can be a random choice among likely tokens

3. In this book, what does "loss" measure?

A. How much memory is left

B. How wrong a guess was

C. How long a chat lasted

4. According to the 2025 researchers, why do language models make things up?

A. They are built to lie

B. Training and testing reward a guess over saying "I don't know"

C. They have no text to learn from

5. How did AlphaGo's networks learn, according to the book?

A. From human expert games and then from playing itself

B. Only from rules typed by hand

C. By copying Deep Blue

6. What does the book say never to type into a chatbot?

A. The word please

B. A math problem

C. Your full name, address, school or a photo

Words to know

parameters: learned numbers

loss: error score

tokens: text pieces

confabulation: confident invention

autocomplete: finish the text

brute: raw force

Answer key: How Does AI Work?

1. B. The weights and biases that training changes
2. C. The pick can be a random choice among likely tokens
3. B. How wrong a guess was
4. B. Training and testing reward a guess over saying "I don't know"
5. A. From human expert games and then from playing itself
6. C. Your full name, address, school or a photo