

AI BEZ STRACHU - 2026

The Visual Encyclopaedia of AI for the Over-20s

60 practical chapters. 10 career-ready topics. AI for
your generation.

FREE 10-CHAPTER SAMPLE - AGE 20+

Inside this sample

The first ten chapters are reproduced from the complete edition. Download, read and share this sample; purchase links for the complete PDF and physical print-on-demand edition are available at aibezstrachu.matsiems.com/books.

01 What AI Actually Is

02 How LLMs Work

03 Training vs Inference

04 Context Windows

05 The Model Landscape

06 Multimodal AI

07 ChatGPT Deep Dive

08 Claude Deep Dive

09 Gemini and Google

10 Open Source Models

CHAPTER 01

What AI Actually Is

Cutting through the hype

Artificial intelligence, in 2026, means large language models and related systems that process and generate text, images, code, and more. The core idea: train a neural network on massive data, and it learns to predict useful outputs. No magic, no consciousness — just patterns at scale.

Key ideas

- Modern AI = large language models trained on massive datasets
- The core mechanism is predicting the next token in a sequence
- No consciousness, no understanding in the human sense — just patterns

Did you know?

GPT-4, the model behind ChatGPT, has an estimated 1.8 trillion parameters — numbers that encode the patterns it learned from the internet.

CHAPTER 02

How LLMs Work

The transformer architecture

Large Language Models (LLMs) use an architecture called the Transformer. It processes all the words in your message at once, weighs how much each word relates to every other word, and uses that to predict the best next word in a reply. Repeat millions of times — that is a conversation.

Key ideas

- Transformer architecture processes whole sequences at once
- Attention mechanisms weigh word relationships across the text
- Training on trillions of tokens builds general language ability

Did you know?

The Transformer architecture was invented by Google researchers in 2017. Their paper was titled 'Attention Is All You Need' — one of the most cited AI papers ever.

CHAPTER 03

Training vs Inference

Two very different things

Training is when a model learns — processing vast data over weeks or months using thousands of GPUs.

Inference is when a trained model answers your question — fast, cheap, and running in real time. You only ever experience inference. Training happens once, offline, by the AI company.

Key ideas

- Training: learning from data — slow, expensive, done once by the company
- Inference: using the trained model to answer questions — fast and cheap
- You only ever interact with the inference side of AI

Did you know?

Training GPT-4 is estimated to have cost over \$100 million. Running one conversation costs a fraction of a cent.

CHAPTER 04

Context Windows

How much AI can hold in mind

A context window is the maximum amount of text an AI can consider at once — its working memory. Models in 2026 have windows of 128K to 1M tokens (roughly 100K to 800K words). Longer contexts mean better multi-step reasoning and document analysis.

Key ideas

- Context window = how much the model can read at once
- Measured in tokens (about $\frac{3}{4}$ of an English word each)
- Larger context = can read whole books, codebases, or reports

Did you know?

In 2023 context windows were 4K tokens (about 3,000 words). By 2026 some models reached 1 million tokens — a 250× jump in two years.

CHAPTER 05

The Model Landscape

Who makes what

A handful of companies dominate the frontier: OpenAI (ChatGPT/GPT-4), Anthropic (Claude), Google (Gemini), Meta (Llama — open source), Mistral (open source, European). Each has a different philosophy: commercial, safety-focused, open, or efficient. Understanding the landscape helps you choose the right tool.

Key ideas

- OpenAI, Anthropic, Google, Meta, and Mistral lead the field
- Each model family has different strengths and pricing
- Open-source models (Llama, Mistral) can be self-hosted for free

Did you know?

Meta released Llama 3 openly in 2024, and it quickly became the most downloaded AI model ever — used in millions of apps worldwide.

CHAPTER 06

Multimodal AI

Beyond text

Modern AI handles not just text but also images, audio, code, and even video. You can upload a photo and ask questions about it, speak to your assistant and get spoken replies, or have AI analyse a chart. Multimodal means AI that works across all these modes at once.

Key ideas

- Multimodal AI handles text, images, audio, video, and code
- You can upload a photo and ask AI what it sees
- Models like GPT-4o and Claude 3.5 are strongly multimodal

Did you know?

GPT-4o ('o' for omni) can see, hear, and respond in real time — making live spoken conversations with AI possible for the first time.

CHAPTER 07

ChatGPT Deep Dive

The one everyone knows

ChatGPT is the world's most popular AI product with over 900M weekly users. The free tier uses GPT-4o Mini; paid (\$20/mo) unlocks GPT-4o and o1 reasoning models. It has a huge plugin and tool ecosystem, code interpreter, image generation via DALL-E, and strong third-party integrations.

Key ideas

- Free: GPT-4o mini · Plus \$20/mo: GPT-4o + o1 reasoning
- Has built-in code interpreter, image generation, and web search
- Largest user base and ecosystem of any AI product in 2026

Did you know?

ChatGPT went from 0 to 100 million users in 2 months — faster than Instagram, TikTok, and every other app in history.

CHAPTER 08

Claude Deep Dive

The one engineers trust

Anthropic's Claude is known for exceptional coding ability, long document analysis (200K context), and careful, nuanced reasoning. Claude 3.5 Sonnet and Opus are the top performers on many coding and writing benchmarks. Its Constitutional AI training makes it notably thoughtful on tricky ethical questions.

Key ideas

- Claude 3.5 Sonnet tops many coding and writing benchmarks
- 200K token context window — can read entire codebases
- Trained with Constitutional AI for safe, thoughtful responses

Did you know?

Anthropic was founded in 2021 by former OpenAI researchers — including the creators of the reinforcement learning from human feedback (RLHF) technique that made ChatGPT possible.

CHAPTER 09

Gemini and Google

The search giant enters the chat

Google's Gemini Ultra is competitive with GPT-4 and Claude Opus on most benchmarks. Its killer advantage: deep integration with the Google ecosystem — Gmail, Drive, Docs, and Search. Gemini Advanced (\$20/mo) includes a 1M token context and Workspace integration that no other model matches at scale.

Key ideas

- Gemini Ultra matches GPT-4 on most benchmarks
- Deep integration with Gmail, Docs, Drive, and Search
- 1M token context window in Gemini Advanced

Did you know?

Google has more AI compute than any other company — its TPU (Tensor Processing Unit) chips power both Google Search and Gemini.

CHAPTER 10

Open Source Models

Run AI on your own machine

Meta's Llama 3 and Mistral's models can be downloaded and run locally for free. With a mid-range GPU you can run a capable 8B or 13B parameter model entirely on your own hardware — no API costs, full privacy, and the ability to fine-tune for your specific needs.

Key ideas

- Llama 3 and Mistral can be downloaded and run locally for free
- No API costs and complete data privacy on your own machine
- Fine-tuning lets you specialise the model for specific tasks

Did you know?

Llama 3 8B running on a MacBook M3 Pro can generate 50 tokens per second — fast enough for real-time conversation, entirely offline.