



What Claude is and how it works

Models, tokens, context and limits: the basics for using it well

Gianluca Bonomo · www.gianlu.it · info@gianlu.it



Contents

How to use this handout	3
1. What Claude is.....	3
1.1 Where you use it.....	3
2. How it reads and writes: tokens and context.....	4
2.1 The context window.....	5
2.2 Why it does not remember past chats	5
3. The family of models.....	5
3.1 A criterion that lasts.....	6
3.2 Effort: how much it should think.....	7
4. The limits to know	8
4.1 Hallucinations.....	8
4.2 Knowledge has a date.....	8
4.3 It is not a person	8
Hands-on exercise	10
Questions that come up in class.....	11
Glossary.....	11
Further reading	11

How to use this handout

This handout gathers, in written and ordered form, everything covered in the lesson. You do not need to memorise it: it is there so you can find things again calmly, later, when the lesson is over and a question is left hanging.

GOAL

Understand what Claude really is, how it reads and writes text, which model to pick and which limits to keep in mind.

What you will be able to do

- Explain in plain words what a language model does.
- Recognize what a token and the context window are.
- Choose between Fable, Opus, Sonnet and Haiku based on the task.
- Spot the main limits: mistakes, outdated knowledge, confidentiality.

1. What Claude is

Claude is an AI assistant that understands and writes natural-language text: you talk to it the way you would talk to a colleague.

Claude is made by **Anthropic**, an AI research company based in San Francisco. The name «Claude» refers both to the assistant you use in chat and to the family of **models** behind it. A model is the program, trained on huge amounts of text, that generates the answers.

It is not a search engine. A search engine returns pages that already exist; Claude **writes** a new answer, word by word, from what you asked. That is why it can summarize, translate, write, explain, analyze spreadsheets and program.

IN PLAIN WORDS

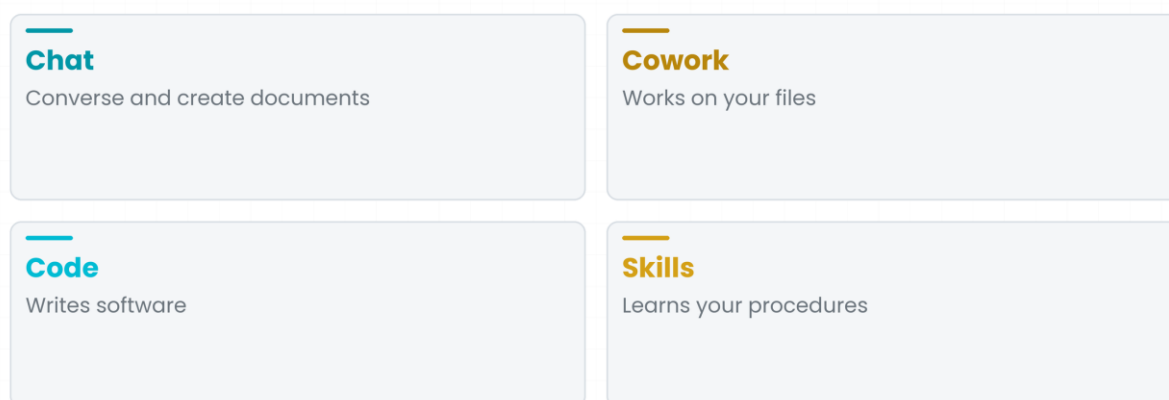
Think of a very well-prepared, lightning-fast, tireless colleague who knows nothing about your company until you tell them.

1.1 Where you use it

The same Claude lives in several places, and this course goes through all of them:

- **Chat** on claude.ai, in the desktop app and in the phone app: you converse, upload files, create documents.
- **Cowork**: Claude works on your behalf on the files and tools you allow it to use.
- **Claude Code**: an assistant that writes and edits software directly in your projects.
- **Skills, connectors and plugins**: modules that teach Claude your procedures and link it to your programs.

Where Claude is used



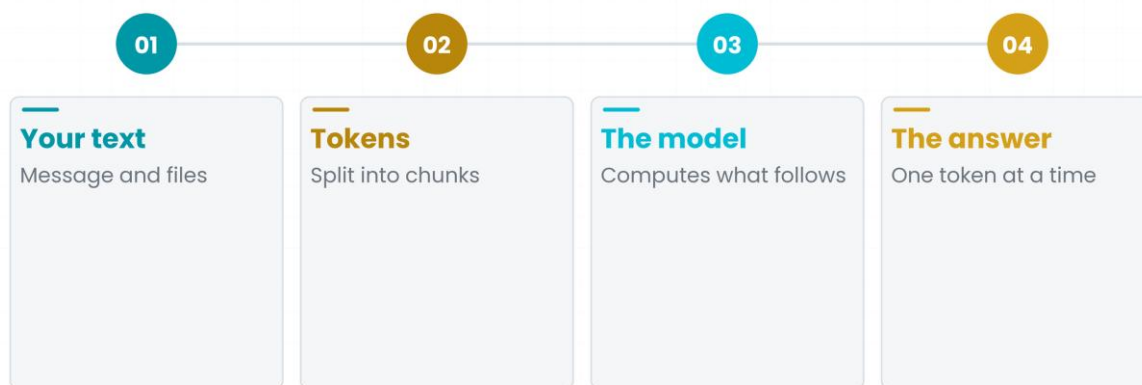
One Claude, four ways to use it.

2. How it reads and writes: tokens and context

Claude does not read letters or whole words: it reads small chunks of text called tokens.

A **token** is a fragment of text: sometimes a whole word, sometimes a syllable or a punctuation mark. The text you write is split into tokens, the model processes them and produces the most sensible answer one token at a time.

How an answer is born



From your message to Claude's answer.

2.1 The context window

The **context window** is the «working memory» of a conversation: everything Claude can keep in view at the same time. It includes your messages, its replies, the files you uploaded and the starting instructions.

According to the official documentation as of October 2026, the latest models (Fable 5.1, Opus 5.5 and Sonnet 5.5) have a window of **one million tokens**, roughly half a million words. Haiku 4.5 offers two hundred thousand tokens. When a conversation grows too long, older parts are summarized or lost.

CAREFUL

The fuller the context, the greater the chance of forgetfulness and mistakes. Open a new chat for each topic: it is the simplest advice for better answers.

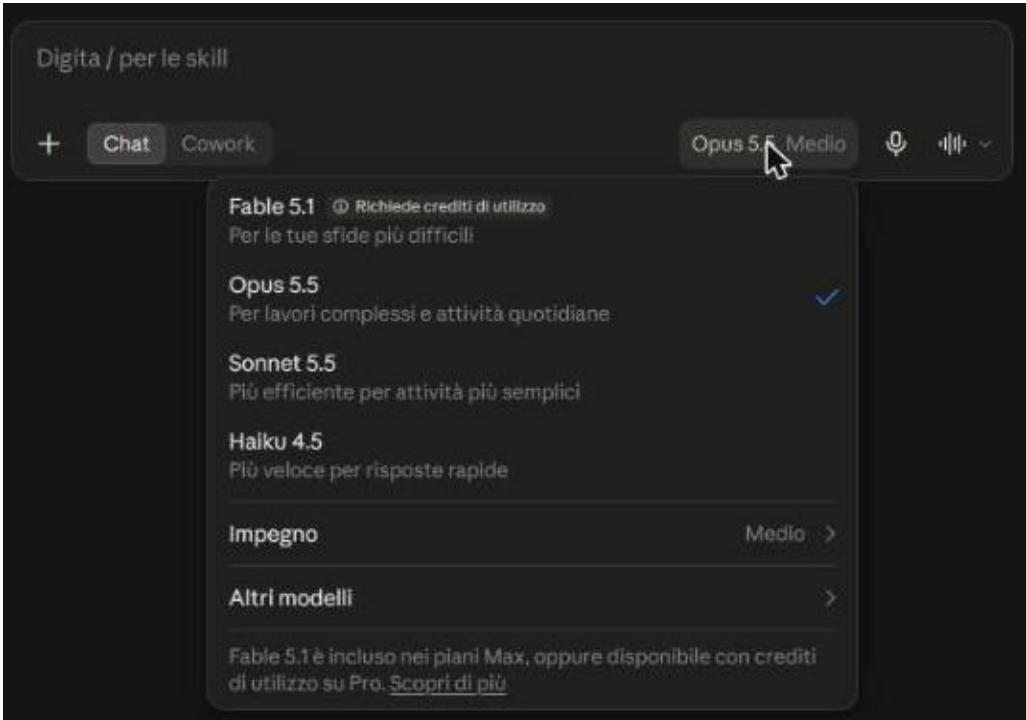
2.2 Why it does not remember past chats

Every conversation starts from zero: the model does not «remember» what you told it yesterday unless you turn on **memory** or work inside a **project** with fixed files and instructions. We will see this in chapter 2.

3. The family of models

There is no single «Claude»: there are several models, like different sizes of the same tool.

In the model menu, at the bottom of the writing box, you pick the one to use. Here is the lineup as of October 2026.



The model menu in claude.ai (Italian interface): Fable 5.1, Opus 5.5, Sonnet 5.5, Haiku 4.5, plus the «Effort» entry.

Model	What it is for	Speed
Fable 5.1	The hardest challenges: deep reasoning and long autonomous jobs	Slower
Opus 5.5	Complex jobs and everyday tasks: the default choice in chat	Medium
Sonnet 5.5	The best balance of speed and intelligence, for simpler tasks	Fast
Haiku 4.5	Quick replies and light tasks	Fastest

The four current models, from the official documentation and the claude.ai menu (October 2026).

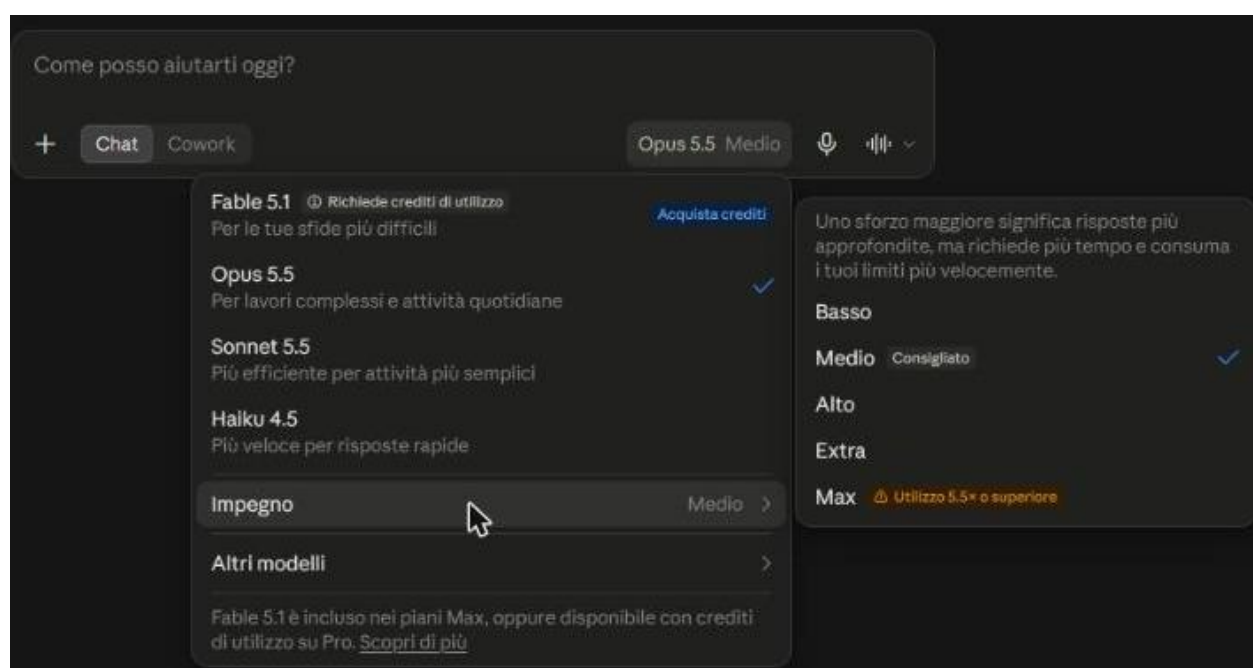
3.1 A criterion that lasts

Names change often, reasoning does not. Think of three levels: **light** (quick and cheap), **balanced** (fine for almost everything) and **powerful** (for hard cases). Start from balanced; go up only if the result is not enough, go down if the request is simple.

3.2 Effort: how much it should think

Next to the model you find the **effort** setting: the higher it is, the more Claude reasons before answering. In the screenshot it is set to «Medium». High effort gives more accurate answers on hard problems but takes longer and uses more of your usage limit.

Clicking «Effort» (in the Italian interface, «Impegno») in the model menu opens five levels: **Low, Medium** (recommended), **High, Extra and Max**. The menu itself warns that more effort means more thorough answers but more time and faster use of your limits; for Max it states usage of 5.5 times or higher.



The effort levels (Italian interface): Low, Medium (recommended), High, Extra, Max.

A PRACTICAL EXAMPLE

Summarizing an email of ten lines: Sonnet or Haiku, low effort.

Checking a contract or a long spreadsheet: Opus, medium or high effort.

Planning a project with many dependencies: Fable, high effort.

Which model to choose



Light, balanced, powerful: go up only if needed.

4. The limits to know

Claude is wrong with confidence: knowing when and why is the most important skill.

4.1 Hallucinations

A model generates the most plausible answer, not the truest one. Sometimes it invents a figure, a quotation, the name of a law: that is a **hallucination**. Always ask for sources, check important numbers and, for recent facts, turn on **web search**.

4.2 Knowledge has a date

The current models have reliable knowledge up to June 2026 (Haiku 4.5, older, up to February 2025). For anything after that you need web search or a document you provide yourself.

4.3 It is not a person

Claude may seem empathetic, but it has no feelings and does not know your real situation. For decisions about health, money or legal matters, use it to prepare and understand, then talk to a professional.

CONFIDENTIALITY

What you write is processed on Anthropic's servers. Do not paste passwords, card numbers or other people's personal data unless you are authorized. In chapter 8 we cover the settings that protect you.

Risk	What to do
Invented fact	Ask for sources and spot-check
Outdated information	Turn on web search or attach the updated document
Chat too long	Open a new conversation and summarize the state
Confidential data	Anonymize before writing

Four limits and the right behavior.

Hands-on exercise

Twenty minutes to test models, effort and limits on a real (anonymous) text.

1. Open claude.ai and start a new chat.
2. Paste a one-page text (an article or one of your own already-public emails) and ask: «Summarize it in five points».
3. Switch to Haiku and repeat the same request: compare speed and quality.
4. Now use Opus with high effort and ask: «Which claims in the text need to be verified?»
5. Write in three lines which combination seemed most useful for this task.

WATCH OUT

Did you use at least two different models?

Did you ask Claude to point out what needs checking?

Did you avoid personal data in the text?

Questions that come up in class

Is Claude the same thing as a chatbot?

Yes, but it is more than a chatbot: besides chat there are Cowork and Claude Code, which perform actions on files and software.

Which model should I normally use?

Start from the default (in the current chat it is Opus 5.5) and drop to Sonnet or Haiku for simple tasks, go up to Fable only for the hardest challenges.

What are the «tokens» I hear about for limits?

They are the small chunks of text Claude works with. Plan usage limits also depend on how many tokens you process.

Does Claude learn from my conversations?

It depends on your account's privacy settings: the Privacy page has a switch that decides whether chats can be used to improve the models. We will check it in chapter 8.

Does it understand Italian?

Yes, it writes and understands Italian and many other languages. You can also switch language within the same chat.

Glossary

Term	What it means
Anthropic	The company that develops Claude.
Context window	The amount of text Claude can keep in mind within the same conversation.
Effort	How long Claude reasons before answering.
Hallucination	A plausible-sounding answer with invented data.
Model	The trained program that generates the answers.
Token	A fragment of text (word, syllable, mark) used to read and write.

Further reading

Anthropic, Models overview (platform.claude.com/docs) – model table, context window and knowledge, accessed 3 October 2026

claude.ai interface, model menu - screenshot of 3 October 2026

WATCH OUT

Data verified on 3 October 2026: models, context windows and menus change often, always check the official models page.