

What is a token in artificial intelligence?

The little bricks AI uses to read and write, explained with practical examples and a few laughs



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



Contents

How to use this handout	3
1. The problem: AI does not read like us	3
2. What is a token?	4
2.1 How many tokens is a text?	5
2.2 From text to numbers	5
3. How AI writes: one token at a time	6
3.1 Why the same question gives different answers.....	7
4. Four answers built token by token.....	7
4.1 Example 1: an easy question	7
4.2 Example 2: a piece of advice.....	9
4.3 Example 3: the rare word	10
4.4 Example 4: the silliest excuse	10
5. Funny tokens: the museum of pieces.....	12
6. Why it is worth knowing.....	14
6.1 Cost.....	14
6.2 Limits and memory.....	14
6.3 Funny mistakes	14
Hands-on exercise	15
Questions that come up in class.....	16
Glossary.....	16
Further reading	17

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 a token is, how AI turns text into tokens and builds an answer by adding one at a time, so you can use chatbots better and make sense of limits and costs.

What you will be able to do

- Explain in your own words what a token is.
- Describe how AI generates an answer one token at a time.
- Roughly estimate how many tokens a text contains.
- Understand why tokens matter for cost, limits and some funny mistakes.
- Use a token counter to check your own texts.

1. The problem: AI does not read like us

You see words and sentences. Inside, the computer sees only numbers.

When you write «Hello, how are you?» to a chatbot, nothing inside the machine «understands» those letters the way you do. First the text has to be taken apart into small pieces, and each piece turned into a number. Only then can the model work with it.

Those pieces are called **tokens**. Think of toy building bricks: with just a few kinds of brick you can build castles, spaceships and even a platypus. With a limited set of tokens, AI builds any sentence.



Tokens are the bricks AI uses to build sentences

THE IDEA IN ONE LINE

A token is a small piece of text: the unit AI uses to read what you write and to write what it answers.

2. What is a token?

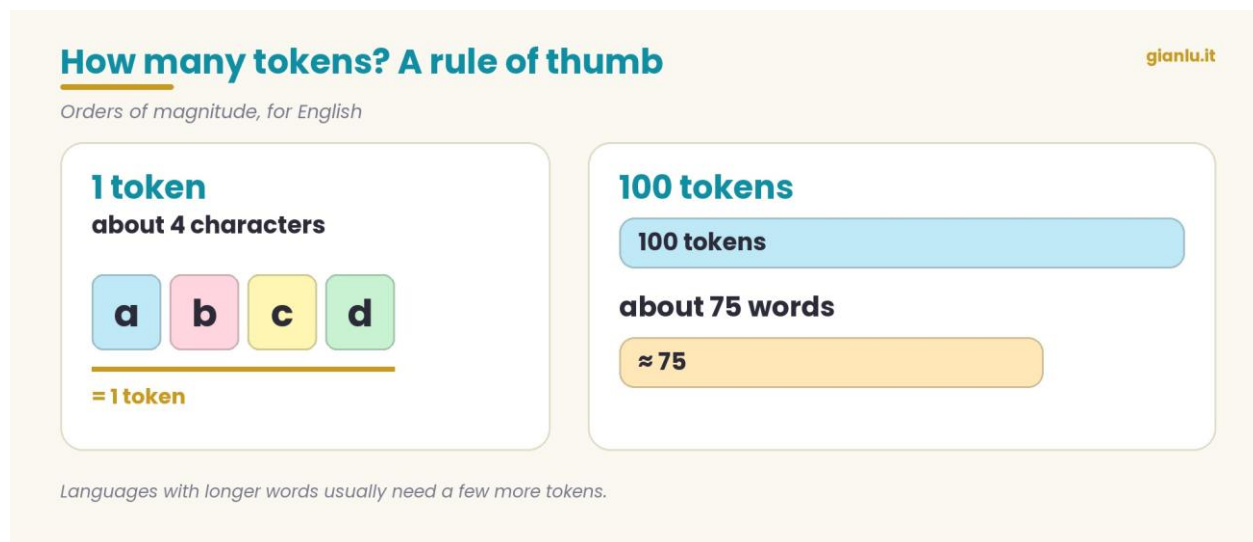
A token is **not always a word**. It can be a whole word, part of a word, a punctuation mark, or a space attached to the word that follows. Common words, which AI has seen millions of times, are usually a single token. Rare words are split into several pieces.

Text	How it might be split	Tokens
cat	«cat»	1
Hello, how are you?	«Hello» «.» «how» «are» «you» «?»	6
platypus	«pl» «aty» «pus»	3

Illustrative examples: every model has its own text «cutter», so real splits can differ. The dot · stands for the space attached in front of a word.

2.1 How many tokens is a text?

A handy rule of thumb, especially for English: **one token is about four characters**, and one hundred tokens are about seventy-five words. Languages with long words and many endings, like Italian, usually use somewhat more tokens than the same text in English.



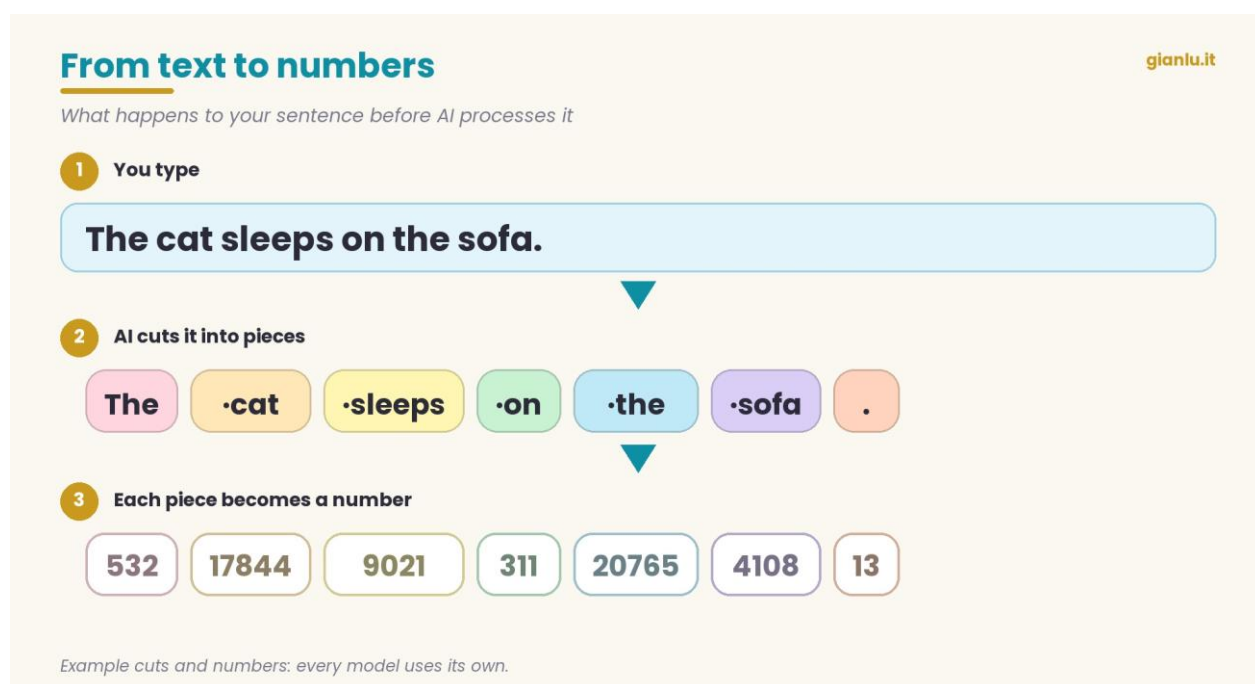
A rule of thumb: one token is about four characters

CAREFUL

These are estimates. Every model has its own way of splitting text, so the same text can have a different token count from one service to another. For an exact number, use the counter of the service you are using.

2.2 From text to numbers

AI has a closed list of tokens it knows, the **vocabulary**: tens of thousands of pieces, each with a code number. The text you write is cut into pieces from that list, and each piece is replaced by its number. That series of numbers is what goes into the model.



From the text you type to the numbers AI processes: colourful pieces

The same thing happens in reverse: the model produces a number, and the number is translated back into the piece of text you see appear on the screen.

3. How AI writes: one token at a time

AI does not write the answer all at once: it builds it one piece at a time, like when you compose a message on your phone.

The mechanism always repeats the same way. The model reads all the text so far, calculates how likely each possible token is to come next, picks one and adds it. Then it starts again, with one more token to read.

How AI writes: one token at a time

gianlu.it

Four moves that repeat, like your keyboard's autocomplete



Example: «The capital of Italy is ...»

Rome		97%
Milan		<1%
Naples		<1%

Example probabilities.

The loop AI uses to write: it repeats for every token

THE TRICK

It is like the autocomplete on your phone keyboard, but trained on an enormous amount of text, and starting over after every single piece.

3.1 Why the same question gives different answers

The model does not always pick the most likely token: sometimes it also picks among less likely ones. A setting called **temperature** controls how adventurous it is. Low: predictable, uniform answers. High: more imagination and more surprises. That is why asking the same thing twice gives you two different texts.

4. Four answers built token by token

Let us see it live. Each example has a question and the answer that is born **one token at a time**. Read the tables from top to bottom: each row adds a piece, and the last column shows the text that exists at that point. The splits are illustrative, but the mechanism is the real one.

4.1 Example 1: an easy question

QUESTION

«What is the capital of Italy?»

Step	Token added	Text built so far
1	«The»	The
2	«·capital»	The capital
3	«·of»	The capital of
4	«·Italy»	The capital of Italy
5	«·is»	The capital of Italy is
6	«·Rome»	The capital of Italy is Rome
7	«.»	The capital of Italy is Rome.

Seven tokens, seven steps. The final text is the last row.

Example 1 · An easy question

gianlu.it

Question: What is the capital of Italy?

▼ AI answers one token at a time:

1

2

3

4

5

6

7

The

·capital

·of

·Italy

·is

·Rome

.

7 tokens in total

FINAL ANSWER

The capital of Italy is Rome.

Example 1: «The capital of Italy is Rome.» in seven tokens

At step 6 the model chooses among many candidates, each with its own probability. This is how that moment might look:

Candidate for step 6	Probability (illustrative)
«·Rome»	97%
«·Milan»	1%
«·Naples»	less than 1%
all other tokens	the rest

When the answer is a very well-known fact, one candidate dominates and the choice is almost forced.

4.2 Example 2: a piece of advice

REQUEST

«Give me one sentence of breakfast advice.»

Step	Token added	Text built so far
1	«Start»	Start
2	«·with»	Start with
3	«·a»	Start with a
4	«·coffee»	Start with a coffee
5	«·and»	Start with a coffee and
6	«·a»	Start with a coffee and a
7	«·slice»	Start with a coffee and a slice
8	«·of»	Start with a coffee and a slice of
9	«·bread»	Start with a coffee and a slice of bread
10	«·and»	Start with a coffee and a slice of bread and
11	«·jam»	Start with a coffee and a slice of bread and jam
12	«·.»	Start with a coffee and a slice of bread and jam.

Twelve tokens. Notice that at step 1 the model did not know how the sentence would end: it found out by writing it.

Example 2 · A piece of advice

gianlu.it

Request: Give me one sentence of breakfast advice.

▼ AI answers one token at a time:

1 Start

2 ·with

3 ·a

4 ·coffee

5 ·and

6 ·a

7 ·slice

8 ·of

9 ·bread

10 ·and

11 ·jam

12 .

12 tokens in total

FINAL ANSWER

Start with a coffee and a slice of bread and jam.

Example 2: the breakfast advice in twelve tokens

4.3 Example 3: the rare word

QUESTION

«Which Australian animal has a duck-like bill?»

Step	Token added	Text built so far
1	«It»	It
2	«'s»	It's
3	«the»	It's the
4	«·pl»	It's the pl
5	«aty»	It's the platy
6	«pus»	It's the platypus
7	«.»	It's the platypus.

A rare word is built in pieces: «platypus» alone can take three tokens, a lot for one word.

Example 3 · The rare word

gianlu.it

Question: Which Australian animal has a duck-like bill?

▼ AI answers one token at a time:



7 tokens in total

FINAL ANSWER

It's the platypus.

Example 3: «platypus» built with three tokens

WHAT YOU LEARN

Common words cost one token, rare ones cost more. The model rebuilds them brick by brick.

4.4 Example 4: the silliest excuse

REQUEST

«Write the silliest excuse for being late.»

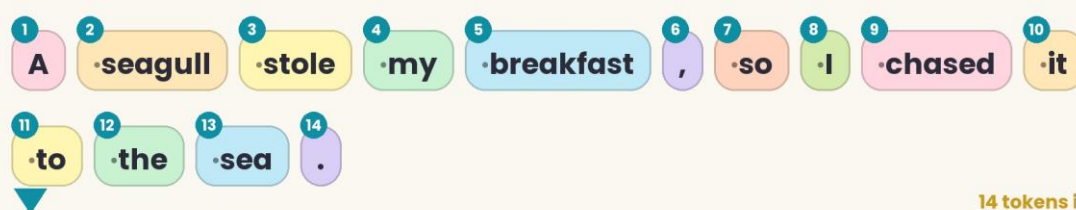
Step	Token added	Text built so far
1	«A»	A
2	«seagull»	A seagull
3	«stole»	A seagull stole
4	«my»	A seagull stole my
5	«breakfast»	A seagull stole my breakfast
6	«,»	A seagull stole my breakfast,
7	«so»	A seagull stole my breakfast, so
8	«I»	A seagull stole my breakfast, so I
9	«chased»	A seagull stole my breakfast, so I chased
10	«it»	A seagull stole my breakfast, so I chased it
11	«to»	A seagull stole my breakfast, so I chased it to
12	«the»	A seagull stole my breakfast, so I chased it to the
13	«sea»	A seagull stole my breakfast, so I chased it to the sea
14	«,»	A seagull stole my breakfast, so I chased it to the sea.

Example 4 · The silliest excuse

gianlu.it

Request: Write the silliest excuse for being late.

AI answers one token at a time:



14 tokens in total

FINAL ANSWER

A seagull stole my breakfast, so I chased it to the sea.

Example 4: the seagull excuse in fourteen tokens

The funniest choice is at step 2. When the request is creative, no candidate dominates and temperature makes the difference:

Candidate for step 2	Probability (illustrative)
«seagull»	22%
«pigeon»	19%
«penguin»	8%
«alien»	4%
all other tokens	the rest

Ask again and you might get a pigeon, a penguin or an alien: a whole different excuse, a whole different story.

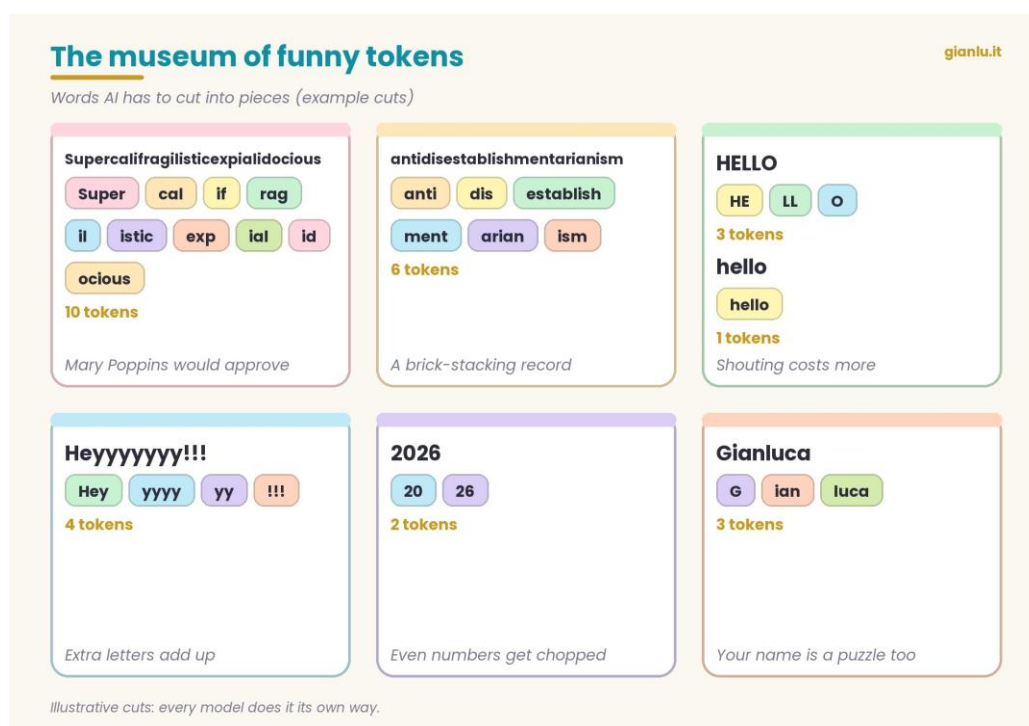


At every step AI must choose the next token

5. Funny tokens: the museum of pieces

If you think tokens are boring, you have not yet seen what happens when AI has to cut up certain words.

The longer, rarer or weirder a word is, the more pieces are needed to rebuild it. Here is our little museum: the cuts are examples, but the principle is real.



The museum of funny tokens: words AI cuts into pieces

- **Mile-long words.** «Supercalifragilisticexpialidocious» is a ten-piece puzzle: AI assembles it brick by brick.
- **Capital letters.** «HELLO» and «hello» are two different things to AI: the shouted version can cost more tokens.
- **Stretched letters.** The more «y» you add to «Heyyyyyy», the more pieces it needs: stretching letters stretches the bill.
- **Numbers.** Even a year like 2026 can be split in two, for example «20» and «26».
- **Your name.** «Gianluca» might become «G», «ian», «luca»: try it in the counter and see how many pieces yours is cut into.

MINI CHALLENGE

Which costs more: «hello» or «HELLOOOOO»? Guess, then check with the counter.

CAREFUL

The cuts shown are illustrative: every model has its own cutter, and real results can differ.

6. Why it is worth knowing

Tokens are not an engineer's curiosity: they explain three things you meet every day when you use a chatbot.

6.1 Cost

Many AI services, especially those for businesses and developers, are paid **by the token**: both the ones you send and the ones you receive are counted. Short prompts and focused answers cost less. Prices change often: always check the service's current price list.

6.2 Limits and memory

The model can only take into account a limited amount of text, the **context window**, measured in tokens. Your question, the conversation so far and the answer all fit inside it. When the conversation gets too long, the oldest parts fall out and the model «forgets» them.

6.3 Funny mistakes

The model sees pieces, not letters. That is why it can stumble on tasks that are trivial for us, like counting how many times a letter appears in a word or spelling a word backwards. It is not stupid: it is looking at the bricks, not at the individual grains of plastic.

Situation	What happens	What to do
Very long text	You hit the limits or pay more	Summarize, split into parts, cut the fluff
Never-ending conversation	The model loses early details	Start over and paste a short summary
Question about single letters	The model sees pieces, not letters	Always check the result yourself
Text in Italian	Usually costs a few more tokens	Don't worry: it only matters at large volumes

Common problems and fixes.

A token is a brick. AI has never written a whole sentence in one go: it has always stacked one piece on top of another.

Gianluca Bonomo

Hands-on exercise

Fifteen minutes, a browser and your favourite chatbot. Nothing to install.

1. Write a sentence you like, for example the first sentence of a work email.
2. Search online for the token counter («tokenizer») offered by the provider of the chatbot you use, and paste your sentence in.
3. Note how many tokens and how many characters there are, and look at how it was split.
4. Try three variants: the same sentence in another language, a long rare word, and a line of numbers only. Compare the counts.
5. Funny token contest: compare «hello», «HELLO», «helloooooo» and your name. Which costs more?
6. Final game: on your phone keyboard type the start of a sentence and always tap the first suggestion. After ten words read what came out: you have just done, by hand, what AI does.

WATCH OUT

Can you say how many tokens your sentence has?

Did you notice which words were split into several pieces?

Can you explain to a colleague why a phone keyboard looks like an AI?

Questions that come up in class

Is a token the same as a word?

No. Sometimes it matches a word, sometimes it is a piece of a word, a punctuation mark or a space. Common words are usually one token, rare ones use more.

How many tokens are in a page of text?

It depends on the model and the language. As an order of magnitude, for English one hundred tokens are about seventy-five words; for other languages the count is usually higher. For an exact number use the service's counter.

Why does the same question give different answers?

Because at each step the model chooses among several likely tokens and does not always take the first one. Temperature controls how adventurous it is: the higher it is, the more the answers vary.

Do I really pay for every token?

With many services for businesses and developers, yes: tokens sent and received are paid for, at rates that change. Consumer subscriptions usually apply usage limits rather than a per-token bill. Check your service's terms.

Why does AI get letter counts wrong?

Because it does not see letters, but pieces of words. Counting letters takes an extra reasoning step, and it sometimes slips. Always check it yourself.

Glossary

Term	What it means
Context window	The maximum number of tokens the model can consider at once: question, conversation and answer.
Model	The trained system that calculates, at each step, which token is most likely.
Probability	The score the model gives to each possible next token.
Temperature	A setting that controls how willing the model is to pick less likely tokens.
Token	A small piece of text (word, part of a word, punctuation) that AI reads and writes.
Tokenization	The step that cuts text into tokens and turns them into numbers.

Term	What it means
Vocabulary	The closed list of tokens a model knows, each with its own numeric code.

Further reading

AI providers' documentation on tokens and token counting – official service pages: rules of thumb (about four characters per token) and online counters

Sennrich, Haddow, Birch – Neural Machine Translation of Rare Words with Subword Units (2016) – the paper that popularised splitting text into subword pieces

Vaswani et al. – Attention Is All You Need (2017) – the architecture behind modern language models

Gianluca Bonomo, Niente paura, è solo IA – the reference book (in Italian)

WATCH OUT

The token splits and probabilities in the examples are illustrative, chosen to explain the mechanism; real numbers depend on the model. The rules of thumb are orders of magnitude. Service prices and limits change often: always check the provider's current documentation. The course is adapted from the Italian original: the Italian examples were replaced with English ones.