



Cos'è un token nell'intelligenza artificiale

I mattoncini con cui l'AI legge e scrive, spiegati con esempi pratici e un po' di risate

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



Indice

Come usare questa dispensa.....	3
1. Il problema: l'AI non legge come noi.....	3
2. Che cos'è un token.....	4
2.1 Quanti token è un testo?.....	5
2.2 Dal testo ai numeri.....	5
3. Come l'AI scrive: un token alla volta.....	6
3.1 Perché la stessa domanda dà risposte diverse.....	7
4. Quattro risposte costruite token dopo token	7
4.1 Esempio 1: una domanda facile.....	7
4.2 Esempio 2: un consiglio.....	9
4.3 Esempio 3: la parola rara.....	10
4.4 Esempio 4: la scusa più assurda	10
5. Token da ridere: il museo dei pezzetti	12
6. Perché ti conviene conoscerli	14
6.1 Costi.....	14
6.2 Limiti e memoria.....	14
6.3 Errori buffi.....	14
Esercitazione pratica.....	15
Le domande che arrivano in aula.....	16
Glossario	16
Per approfondire	17

Come usare questa dispensa

Questa dispensa raccoglie, in forma scritta e ordinata, tutto quello che vediamo nella lezione. Non serve studiarla a memoria: serve a ritrovare le cose con calma, dopo, quando la lezione è finita e resta una domanda in sospeso.

OBIETTIVO

Capire che cos'è un token, come l'AI trasforma il testo in token e come costruisce una risposta aggiungendone uno alla volta, per usare meglio i chatbot e capire limiti e costi.

Che cosa saprai fare alla fine

- Spiegare con parole tue che cos'è un token.
- Descrivere come l'AI genera una risposta un token alla volta.
- Stimare a occhio quanti token contiene un testo.
- Capire perché i token contano per costi, limiti e certi errori buffi.
- Usare un contatore di token per controllare i propri testi.

1. Il problema: l'AI non legge come noi

Tu vedi parole e frasi. Il computer, dentro, vede soltanto numeri.

Quando scrivi a un chatbot «Ciao, come stai?», nessuno dentro la macchina «capisce» quelle lettere come le capisci tu. Prima il testo va smontato in pezzetti, e ogni pezzetto va trasformato in un numero. Solo dopo il modello può lavorarci.

Quei pezzetti si chiamano **token**. Immagina i mattoncini delle costruzioni: con pochi tipi di mattoncino costruisci castelli, astronavi e perfino un ornitorinco. Con pochi tipi di token l'AI costruisce qualsiasi frase.



I token sono i mattoncini con cui l'AI costruisce le frasi

L'IDEA IN UNA RIGA

Un token è un pezzetto di testo: l'unità con cui l'AI legge quello che scrivi e scrive quello che ti risponde.

2. Che cos'è un token

Un token **non è sempre una parola**. Può essere una parola intera, un pezzo di parola, un segno di punteggiatura, o uno spazio attaccato alla parola che segue. Le parole comuni, che l'AI ha visto milioni di volte, di solito sono un solo token. Le parole rare vengono spezzate in più pezzi.

Testo	Come potrebbe essere spezzato	Token
gatto	«gatto»	1
Ciao, come stai?	«Ciao» «,» «come» «stai» «?»	5
ornitorinco	«orn» «ito» «rin» «co»	4

Esempi illustrativi: ogni modello ha il suo «tagliatore» di testo e le divisioni reali possono differire. Il puntino · indica lo spazio che si attacca davanti alla parola.

2.1 Quanti token è un testo?

Regola pratica, valida soprattutto per l'inglese: **un token è circa quattro caratteri**, e cento token sono circa settantacinque parole. L'italiano, con le sue parole lunghe e le sue desinenze, di solito consuma un po' più token dello stesso testo in inglese.



Una regola a occhio: un token è circa quattro caratteri

ATTENZIONE

Sono stime. Ogni modello ha il proprio sistema per dividere il testo, quindi lo stesso testo può avere un numero di token diverso da un servizio all'altro. Per un numero preciso usa il contatore del servizio che stai usando.

2.2 Dal testo ai numeri

L'AI ha un elenco chiuso di token che conosce, il **vocabolario**: decine di migliaia di pezzetti, ciascuno con un numero di codice. Il testo che scrivi viene tagliato in pezzetti dell'elenco, e ogni pezzetto sostituito dal suo numero. È questa serie di numeri che entra nel modello.

Dal testo ai numeri

gianlu.it

Quello che succede alla tua frase prima che l'AI la elabori

1 Scrivi

Il gatto dorme sul divano.

2 L'AI la taglia in pezzetti

Il ·gatto ·dorme ·sul ·divano .

3 Ogni pezzetto diventa un numero

532 17844 9021 311 20765 4108

Tagli e numeri di esempio: ogni modello usa i suoi.

Dal testo che scrivi ai numeri che l'AI elabora: pezzetti colorati

La stessa cosa succede al contrario: il modello produce un numero, e il numero viene tradotto di nuovo nel pezzetto di testo che vedi comparire sullo schermo.

3. Come l'AI scrive: un token alla volta

L'AI non scrive la risposta tutta insieme: la costruisce un pezzetto per volta, come quando componi un messaggio sul telefono.

Il meccanismo si ripete sempre uguale. Il modello legge tutto il testo che c'è finora, calcola quanto è probabile ciascun token possibile come prossimo, ne sceglie uno e lo aggiunge. Poi ricomincia, con un token in più da leggere.

Come scrive l'AI: un token alla volta

gianlu.it

Quattro mosse che si ripetono, come il completamento della tastiera



Esempio: «La capitale d'Italia è ...»

Roma		97%
Milano		<1%
Napoli		<1%

Probabilità di esempio.

Il ciclo con cui l'AI scrive: si ripete per ogni token

IL TRUCCO

È come il completamento automatico della tastiera del telefono, ma allenato su una quantità enorme di testi, e che ricomincia dopo ogni pezzetto.

3.1 Perché la stessa domanda dà risposte diverse

Il modello non sceglie sempre il token più probabile: a volte pesca anche tra quelli meno probabili. Quanto osa lo regola un'impostazione chiamata **temperatura**. Bassa: risposte prevedibili e uniformi. Alta: più fantasia e più sorprese. Per questo, chiedendo due volte la stessa cosa, ricevi due testi diversi.

4. Quattro risposte costruite token dopo token

Vediamolo dal vivo. In ogni esempio c'è una domanda e la risposta che nasce **un token alla volta**. Leggi le tabelle dall'alto in basso: ogni riga aggiunge un pezzetto, e l'ultima colonna mostra il testo che esiste a quel punto. I tagli sono illustrativi, ma il meccanismo è quello vero.

4.1 Esempio 1: una domanda facile

DOMANDA

«Qual è la capitale d'Italia?»

Passo	Token aggiunto	Testo costruito finora
1	«La»	La
2	«·capitale»	La capitale
3	«·d'»	La capitale d'
4	«Italia»	La capitale d'Italia
5	«·è»	La capitale d'Italia è
6	«·Roma»	La capitale d'Italia è Roma
7	«.»	La capitale d'Italia è Roma.

Sette token, sette passi. Il testo finale è l'ultima riga.

Esempio 1 · Una domanda facile

gianlu.it

Domanda: Qual è la capitale d'Italia?

▼ L'AI risponde un token alla volta:

1

2

3

4

5

6

7

La

·capitale

·d'

Italia

·è

·Roma

.

7 token in tutto

RISPOSTA FINALE

La capitale d'Italia è Roma.

Esempio 1: «La capitale d'Italia è Roma.» in sette token

Al passo 6 il modello sceglie tra molti candidati, ciascuno con la sua probabilità. Ecco come potrebbe apparire quel momento:

Candidato per il passo 6	Probabilità (illustrativa)
«·Roma»	97%
«·Milano»	1%
«·Napoli»	meno dell'1%
tutti gli altri token	il resto

Quando la risposta è un fatto molto noto, un candidato domina e la scelta è quasi obbligata.

4.2 Esempio 2: un consiglio

RICHIESTA
«Dammi un consiglio per la colazione, in una frase.»

Passo	Token aggiunto	Testo costruito finora
1	«Inizia»	Inizia
2	«·con»	Inizia con
3	«·un»	Inizia con un
4	«·caffè»	Inizia con un caffè
5	«·e»	Inizia con un caffè e
6	«·una»	Inizia con un caffè e una
7	«·fetta»	Inizia con un caffè e una fetta
8	«·di»	Inizia con un caffè e una fetta di
9	«·pane»	Inizia con un caffè e una fetta di pane
10	«·e»	Inizia con un caffè e una fetta di pane e
11	«·marmellata»	Inizia con un caffè e una fetta di pane e marmellata
12	«·»	Inizia con un caffè e una fetta di pane e marmellata.

Dodici token. Nota che il modello non sapeva, al passo 1, come sarebbe finita la frase: l'ha scoperto scrivendola.

Esempio 2 · Un consiglio

gianlu.it

Richiesta: Dammi un consiglio per la colazione, in una frase.

▼ L'AI risponde un token alla volta:

1Inizia

2·con

3·un

4·caffè

5·e

6·una

7·fetta

8·di

9·pane

10·e

11·marmellata

12.

12 token in tutto

RISPOSTA FINALE

Inizia con un caffè e una fetta di pane e marmellata.

Esempio 2: il consiglio per la colazione in dodici token

4.3 Esempio 3: la parola rara

DOMANDA

«Come si chiama l'animale australiano con il becco d'anatra?»

Passo	Token aggiunto	Testo costruito finora
1	«È»	È
2	«l'»	È l'
3	«orn»	È l'orn
4	«ito»	È l'ornito
5	«rin»	È l'ornitorin
6	«co»	È l'ornitorinco
7	«.»	È l'ornitorinco.

Una parola rara si costruisce a pezzi: «ornitorinco» da solo vale quattro token, un sacco per una sola parola.

Esempio 3 · La parola rara

gianlu.it

Domanda: Come si chiama l'animale australiano con il becco d'anatra?

▼ L'AI risponde un token alla volta:



7 token in tutto

RISPOSTA FINALE

È l'ornitorinco.

Esempio 3: «ornitorinco» costruito con quattro token

CHE COSA IMPARI

Le parole comuni costano un token, quelle rare ne costano di più. Il modello le ricostruisce mattoncino su mattoncino.

4.4 Esempio 4: la scusa più assurda

RICHIESTA

«Scrivi la scusa più assurda per essere arrivato in ritardo.»

Passo	Token aggiunto	Testo costruito finora
1	«Un»	Un
2	«gabbiano»	Un gabbiano
3	«mi»	Un gabbiano mi
4	«ha»	Un gabbiano mi ha
5	«rubato»	Un gabbiano mi ha rubato
6	«la»	Un gabbiano mi ha rubato la
7	«colazione»	Un gabbiano mi ha rubato la colazione
8	« , »	Un gabbiano mi ha rubato la colazione,
9	«e»	Un gabbiano mi ha rubato la colazione, e
10	«l'»	Un gabbiano mi ha rubato la colazione, e l'
11	«ho»	Un gabbiano mi ha rubato la colazione, e l'ho
12	«inseguito»	Un gabbiano mi ha rubato la colazione, e l'ho inseguito
13	«fino»	Un gabbiano mi ha rubato la colazione, e l'ho inseguito fino
14	«al»	Un gabbiano mi ha rubato la colazione, e l'ho inseguito fino al
15	«mare»	Un gabbiano mi ha rubato la colazione, e l'ho inseguito fino al mare
16	« . »	Un gabbiano mi ha rubato la colazione, e l'ho inseguito fino al mare.

Esempio 4 · La scusa più assurda

gianlu.it

Richiesta:

Scrivi la scusa più assurda per essere arrivato in ritardo.

▼

L'AI risponde un token alla volta:

1

Un

2

gabbiano

3

mi

4

ha

5

rubato

6

la

7

colazione

8

,

9

e

10

l'

11

ho

12

inseguito

13

fino

14

al

15

mare

16

.

16 token in tutto

RISPOSTA FINALE

Un gabbiano mi ha rubato la colazione, e l'ho inseguito fino al mare.

Esempio 4: la scusa del gabbiano in sedici token

Qui la scelta più divertente è al passo 2. Quando la richiesta è creativa, nessun candidato domina e la temperatura fa la differenza:

Candidato per il passo 2	Probabilità (illustrativa)
«gabbiano»	22%
«piccione»	19%
«pinguino»	8%
«alieno»	4%
tutti gli altri token	il resto

Rilancia la domanda e potresti ottenere un piccione, un pinguino o un alieno: tutta un'altra scusa, tutta un'altra storia.



A ogni passo l'AI deve scegliere il prossimo token

5. Token da ridere: il museo dei pezzetti

Se pensi che i token siano noiosi, non hai ancora visto che cosa succede quando l'AI deve tagliare certe parole.

Più una parola è lunga, rara o scritta in modo strano, più pezzetti servono per ricostruirla. Ecco il nostro piccolo museo: i tagli sono di esempio, ma il principio è reale.



Il museo dei token buffi: parole che l'AI taglia in pezzetti

- **Parole chilometriche.** «Supercalifragilistichespiralidoso» è un puzzle da nove pezzi: l'AI lo monta mattoncino dopo mattoncino.
- **Maiuscole.** «CIAO» e «ciao» sono per l'AI due cose diverse: la versione urlata può costare più token.
- **Lettere allungate.** Più «o» aggiungi a «Ciaoooooooo», più pezzetti servono: allungare le lettere allunga il conto.
- **Numeri.** Anche un anno come 2026 può essere spezzato in due, per esempio «20» e «26».
- **Il tuo nome.** «Gianluca» potrebbe diventare «G», «ian», «luca»: provalo sul contatore e scopri in quanti pezzi si taglia il tuo.

MINI SFIDA

Quale costa di più: «ciao» o «CIAAAAAAO»? Indovina, poi controlla con il contatore.

ATTENZIONE

I tagli mostrati sono illustrativi: ogni modello ha il suo tagliatore, e il risultato reale può cambiare.

6. Perché ti conviene conoscerli

I token non sono una curiosità da ingegneri: spiegano tre cose che incontri ogni giorno quando usi un chatbot.

6.1 Costi

Molti servizi di AI, soprattutto quelli per aziende e sviluppatori, si pagano **a token**: si contano sia quelli che invii sia quelli che ricevi. Prompt corti e risposte mirate costano meno. I prezzi cambiano spesso: guarda sempre il listino aggiornato del servizio.

6.2 Limiti e memoria

Il modello può tenere in considerazione solo una quantità limitata di testo, la **finestra di contesto**, misurata in token. Dentro ci stanno la tua domanda, la conversazione fin lì e la risposta. Quando la conversazione diventa troppo lunga, le parti più vecchie escono e il modello le «dimentica».

6.3 Errori buffi

Il modello vede pezzetti, non lettere. Per questo può inciampare in compiti che per noi sono banali, come contare quante volte compare una lettera in una parola o scrivere una parola al contrario. Non è stupido: sta guardando i mattoncini, non i singoli granelli di plastica.

Situazione	Che cosa succede	Che cosa fare
Testo molto lungo	Si superano i limiti o si paga di più	Riassumi, dividi in parti, toglì il superfluo
Conversazione infinita	Il modello perde i dettagli iniziali	Ricomincia e incolla un breve riassunto
Domanda sulle singole lettere	Il modello vede pezzetti, non lettere	Verifica sempre tu il risultato
Testo in italiano	Di solito costa qualche token in più	Nessuna paura: conta solo se i volumi sono grandi

Problemi comuni e rimedi.

Un token è un mattoncino. L'AI non ha mai scritto una frase intera in un colpo solo: ne ha sempre messo uno sopra l'altro.

Gianluca Bonomo

Esercitazione pratica

Quindici minuti, un browser e il tuo chatbot preferito. Non serve installare nulla.

1. Scrivi una frase di tuo gusto, per esempio la tua prima frase di una mail di lavoro.
2. Cerca online il contatore di token («tokenizer») messo a disposizione dal fornitore del chatbot che usi, e incolla la frase.
3. Annota quanti token e quanti caratteri ci sono, e guarda come è stata spezzata.
4. Prova con tre varianti: la stessa frase in inglese, una parola lunga e rara, e una riga di soli numeri. Confronta i conteggi.
5. Gara dei token buffi: confronta «ciao», «CIAO», «ciaooooo» e il tuo nome. Quale costa di più?
6. Gioco finale: sulla tastiera del telefono scrivi l'inizio di una frase e premi sempre il primo suggerimento. Dopo dieci parole leggi che cosa è uscito: hai appena fatto, a mano, il lavoro dell'AI.

ATTENZIONE

Sai dire quanti token ha la tua frase?

Hai notato quali parole sono state spezzate in più pezzi?

Sai spiegare a un collega perché la tastiera del telefono assomiglia a un'AI?

Le domande che arrivano in aula

Un token è la stessa cosa di una parola?

No. A volte coincide con una parola, a volte è un pezzo di parola, un segno di punteggiatura o uno spazio. Le parole comuni di solito sono un solo token, quelle rare ne usano di più.

Quanti token ci sono in una pagina di testo?

Dipende dal modello e dalla lingua. Come ordine di grandezza, per l'inglese cento token sono circa settantacinque parole; per l'italiano il conto è di solito un po' più alto. Per un numero preciso usa il contatore del servizio.

Perché la stessa domanda dà risposte diverse?

Perché il modello a ogni passo sceglie tra più token probabili e non sempre prende il primo. La temperatura regola quanto osa: più è alta, più le risposte variano.

Pago davvero per ogni token?

Con molti servizi per aziende e sviluppatori, sì: si pagano i token inviati e quelli ricevuti, con tariffe che cambiano. Gli abbonamenti per il pubblico di solito applicano limiti d'uso e non un conto a token. Controlla le condizioni del tuo servizio.

Perché l'AI sbaglia a contare le lettere di una parola?

Perché non vede le lettere, ma pezzetti di parola. Contare le lettere gli richiede un ragionamento in più, e a volte sbaglia. Controlla sempre tu.

Glossario

Termine	Che cosa significa
Finestra di contesto	Quantità massima di token che il modello può considerare in una volta: domanda, conversazione e risposta.
Modello	Il sistema addestrato che calcola, a ogni passo, quale token è il più probabile.
Probabilità	Il punteggio che il modello assegna a ogni token possibile come prossimo.
Temperatura	Impostazione che regola quanto il modello osa scegliere token meno probabili.
Token	Pezzo di testo (parola, parte di parola, punteggiatura) che l'AI legge e scrive.

Termine	Che cosa significa
Tokenizzazione	L'operazione che taglia il testo in token e li trasforma in numeri.
Vocabolario	L'elenco chiuso dei token che un modello conosce, ciascuno con il suo codice numerico.

Per approfondire

Documentazione dei fornitori di AI sui token e sul conteggio dei token – pagine ufficiali dei servizi: regole pratiche (circa quattro caratteri per token) e contatori online

Sennrich, Haddow, Birch – Neural Machine Translation of Rare Words with Subword Units (2016) – l'articolo che ha diffuso lo spezzettamento in pezzi di parola

Vaswani e altri – Attention Is All You Need (2017) – l'architettura alla base dei moderni modelli linguistici

Gianluca Bonomo, Niente paura, è solo IA – il libro di riferimento

ATTENZIONE

Le suddivisioni in token e le probabilità degli esempi sono illustrative, scelte per spiegare il meccanismo; i numeri reali dipendono dal modello. Le regole pratiche sono ordini di grandezza. Prezzi e limiti dei servizi cambiano spesso: verifica sempre la documentazione aggiornata del fornitore.