

"Apriamo la scatola nera dell'IA... per quanto possibile!"

Franco Cutugno – Università Federico II di Napoli



Centro Interdipartimentale di ricerca
dell'Università Federico II di Napoli -
Urban/Eco



Palermo 4 luglio 2025

Studies in
NATURAL
LANGUAGE
PROCESSING

DISTRIBUTIONAL SEMANTICS



Alessandro Lenci and Magnus Sahlgren

1. every language has a structure of its own, and there are no universal linguistic categories;
2. the study of language must be primarily synchronic;
3. linguistics must be autonomous with respect to other disciplines, especially psychology.

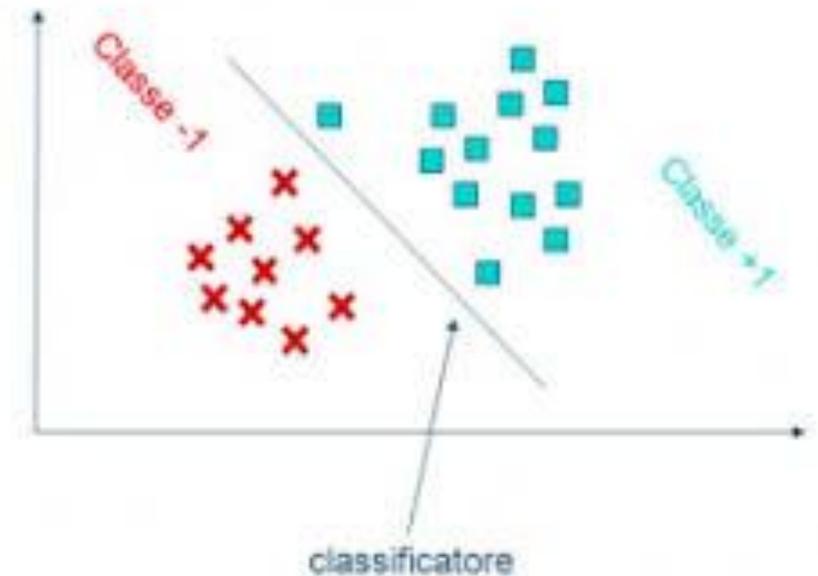
Distributional Hypothesis

The semantic similarity between two lexemes is a function of the similarity of their linguistic contexts.

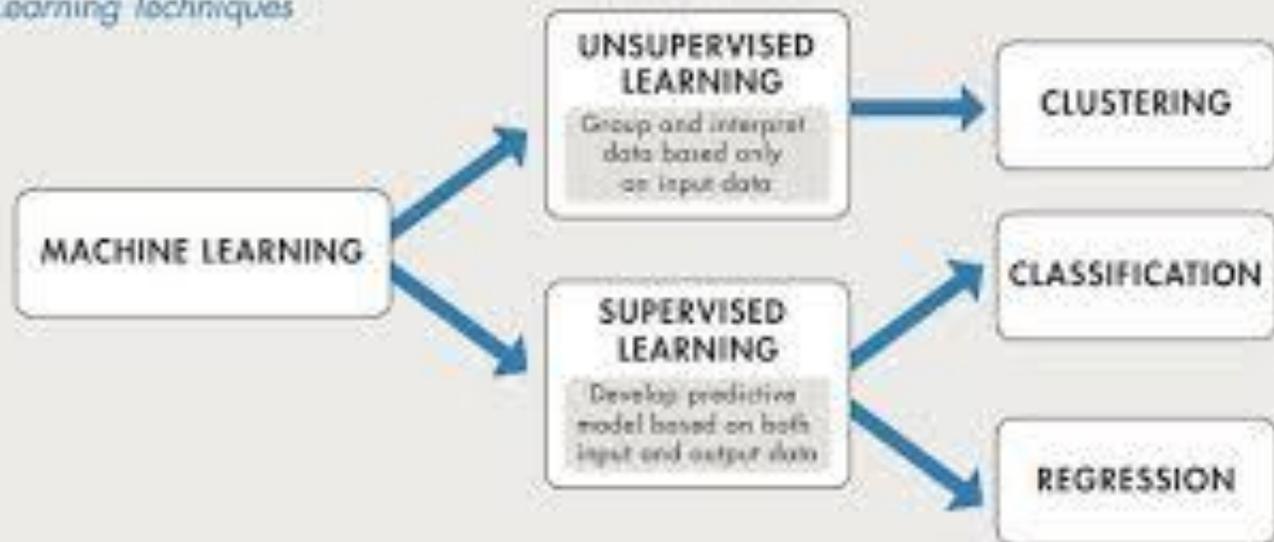
Linguistic elements are identified by their **distributions**, defined as the sum of the contexts in which they occur.

Età	Genere	Pressione sanguigna	Livello di glucosio	Colesterolo	Fumo	Familiarità con la malattia	Malattia cardiovascolare
45	Maschio	Alta	Elevato	Alto	Sì	No	Sì
55	Femmina	Normale	Elevato	Normale	No	Sì	No
60	Maschio	Alta	Normale	Alto	Sì	No	Sì
40	Femmina	Normale	Normale	Normale	No	Sì	No
65	Maschio	Normale	Elevato	Alto	No	No	Sì
50	Femmina	Alta	Elevato	Alto	Sì	Sì	Sì
70	Maschio	Normale	Normale	Normale	No	No	No
48	Femmina	Alta	Normale	Normale	No	Sì	No
42	Maschio	Alta	Elevato	Alto	No	No	
35	Femmina	Normale	Normale	Normale	No	No	

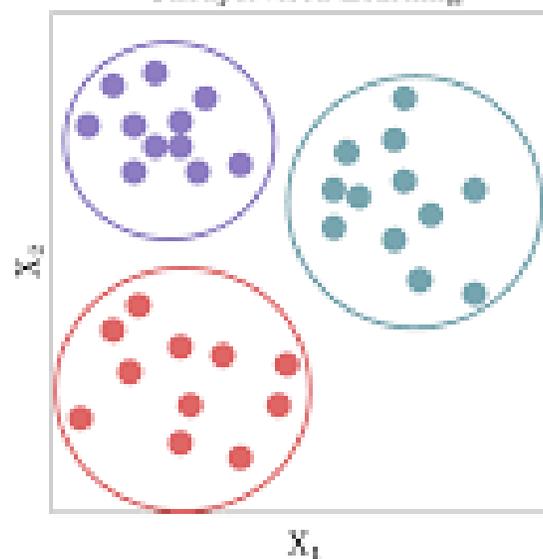
Le tabelle sono poi convertite in forma numerica



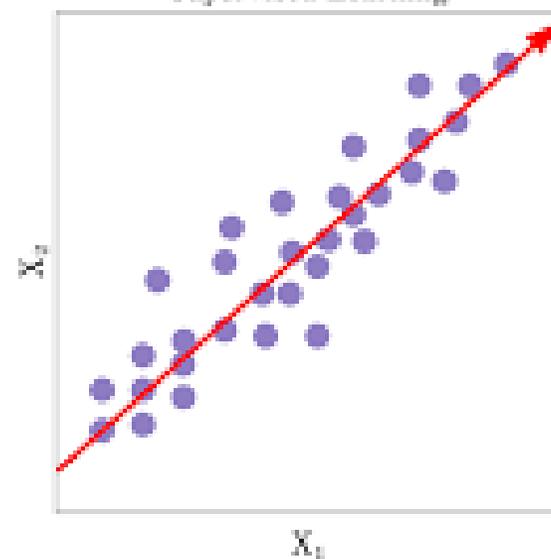
Machine Learning Techniques



Unsupervised Learning



Supervised Learning



	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.3 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.5 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each word is represented as a row vector of length four.



$$Y = W \% X \text{ train}$$

$$Y = W \% X \text{ test}$$

... lemon, a [tablespoon of apricot jam, a] pinch ...
 c1 c2 w c3 c4

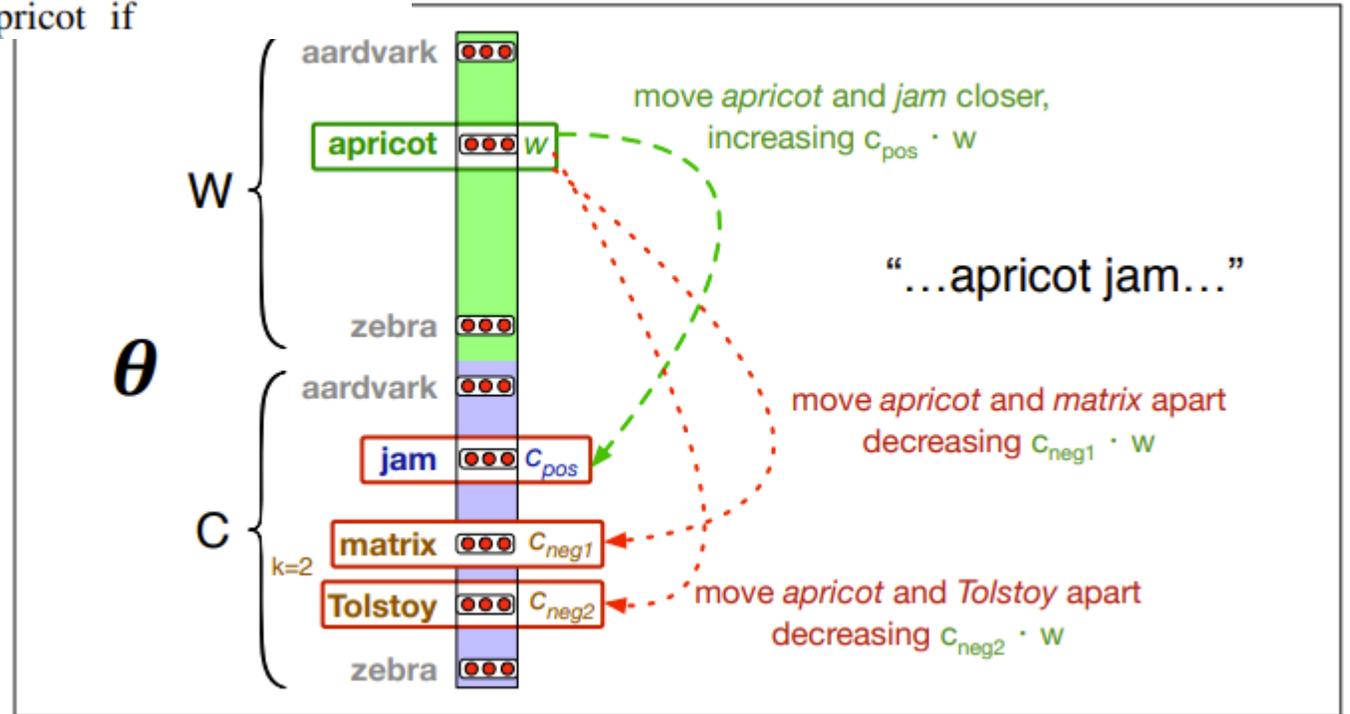
This example has a target word w (apricot), and 4 context words in the $L = \pm 2$ window, resulting in 4 positive training instances (on the left below):

positive examples +

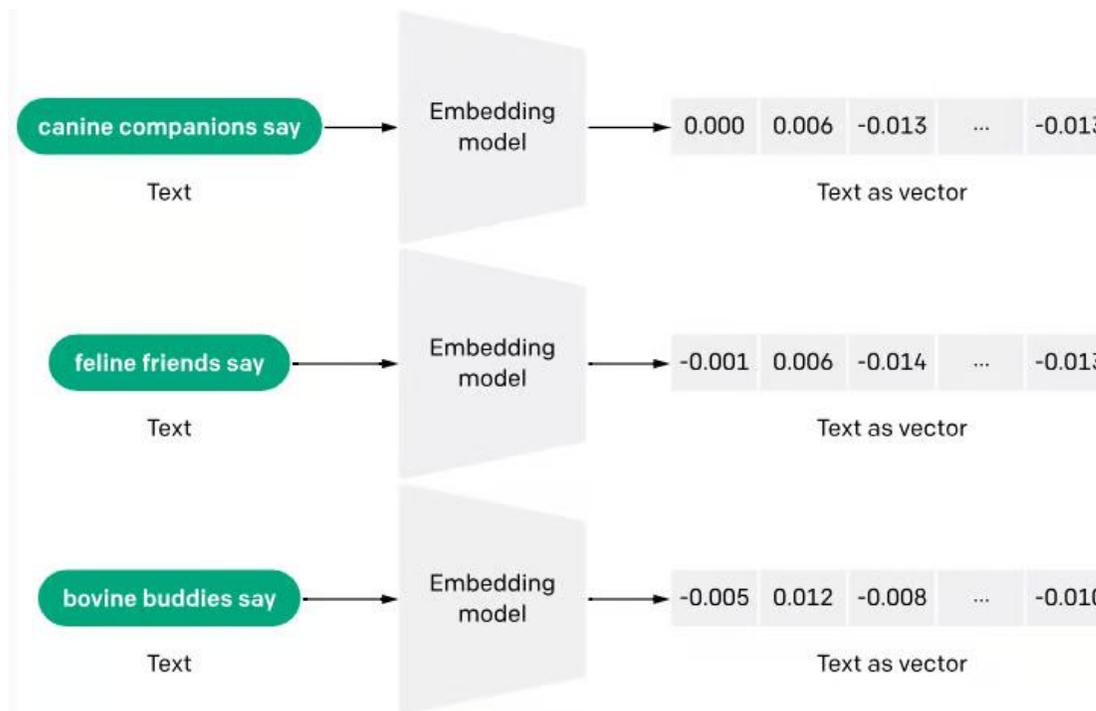
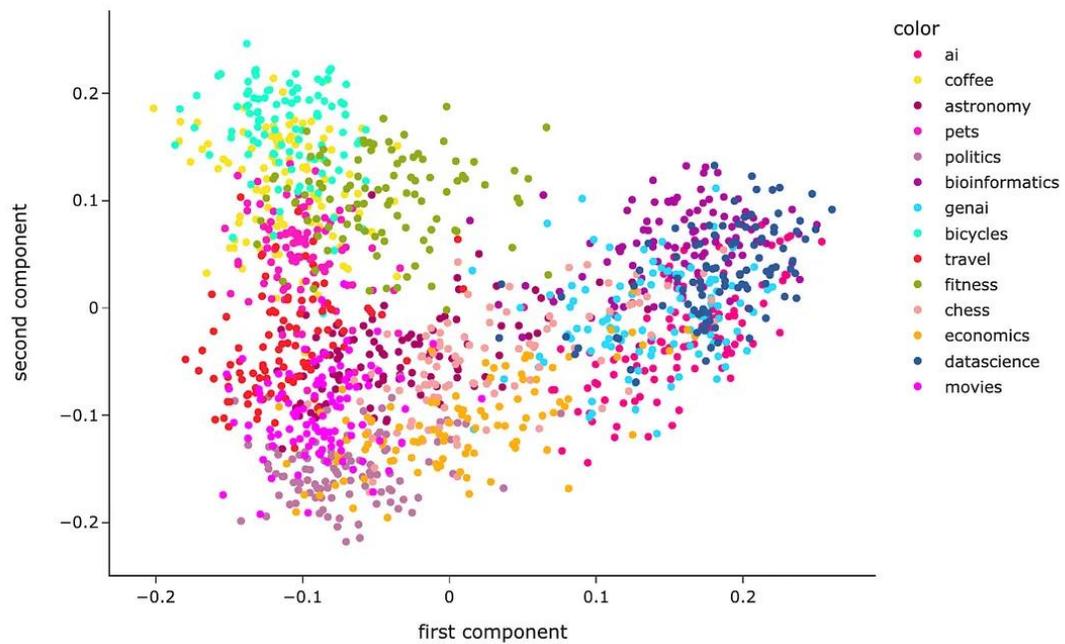
w	c_{pos}
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

negative examples -

w	c_{neg}	w	c_{neg}
apricot	aardvark	apricot	seven
apricot	my	apricot	forever
apricot	where	apricot	dear
apricot	coaxial	apricot	if



PCA embeddings

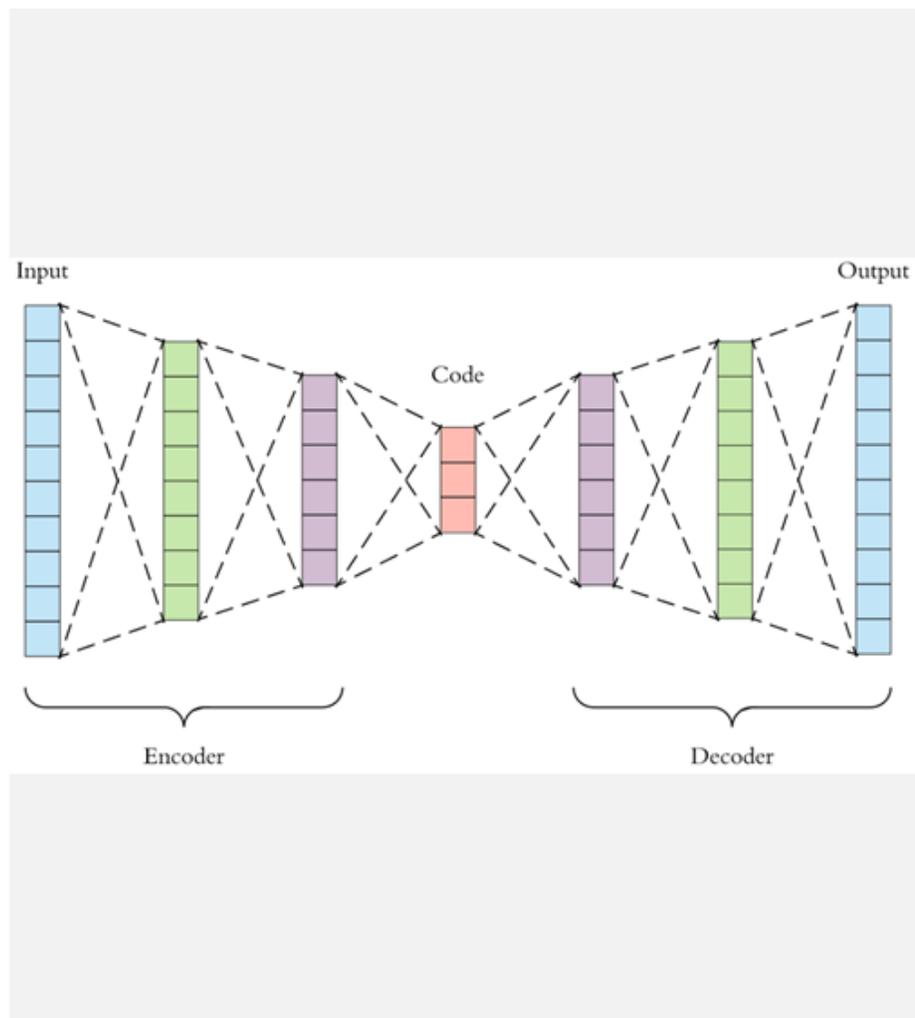


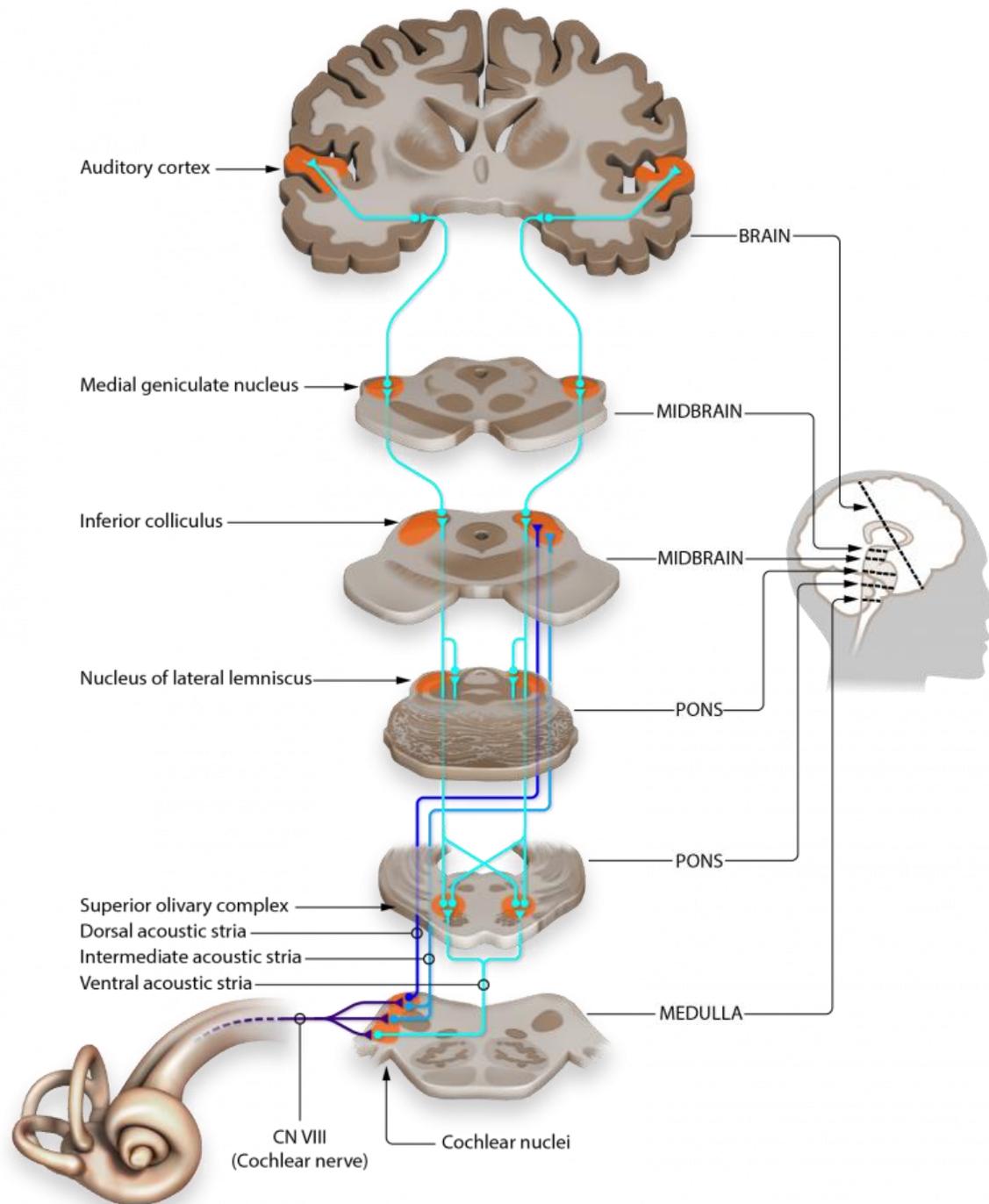
AUTOENCODING

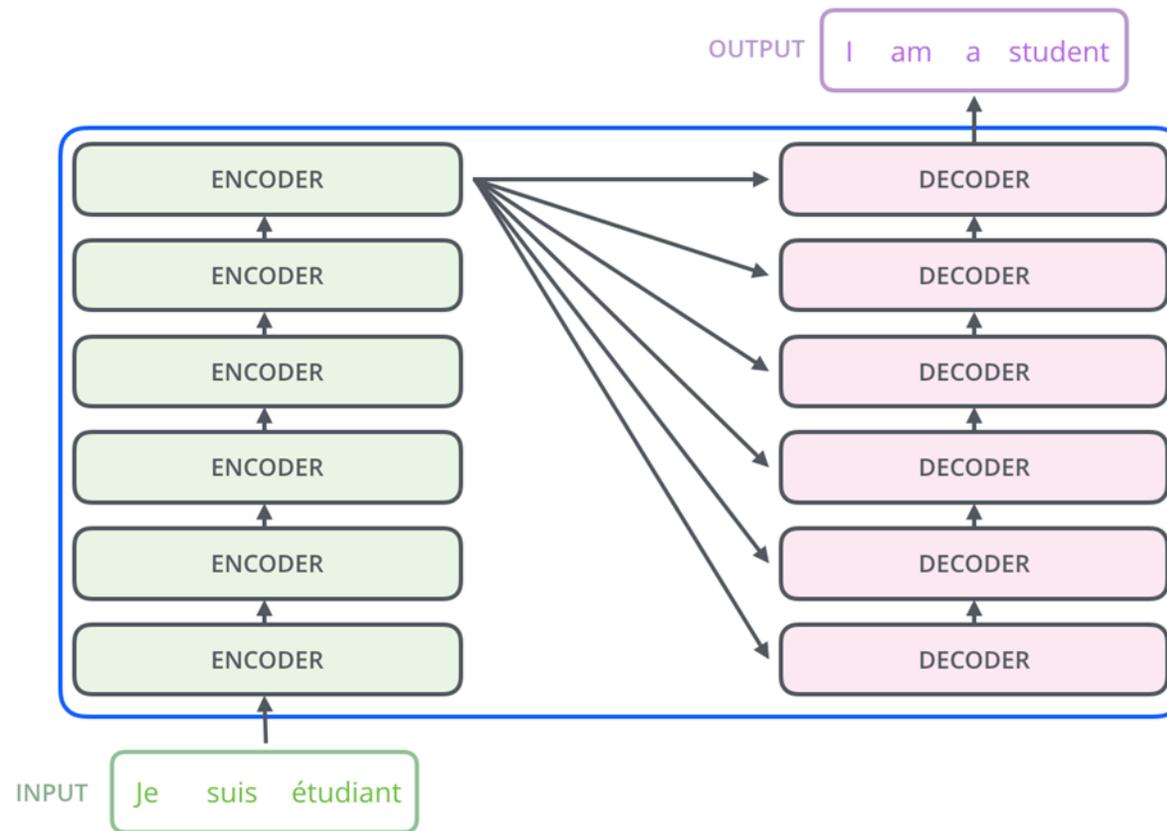
Supervised approach:
Use features to predict labels

Unsupervised approach:
Use features to group items

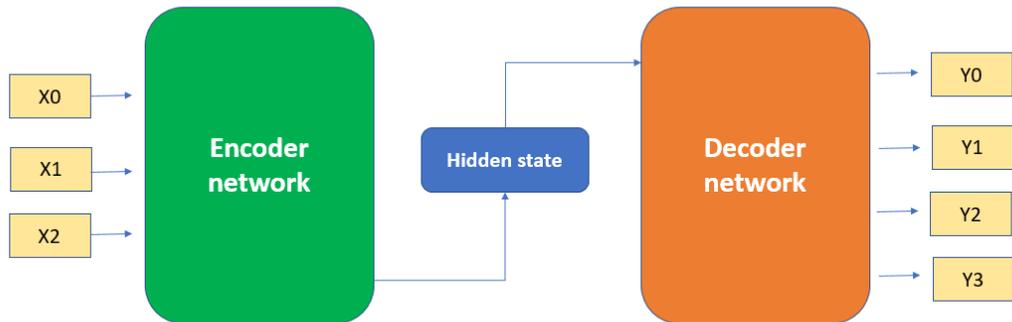
Self-supervised approach:
Use features to predict... features???



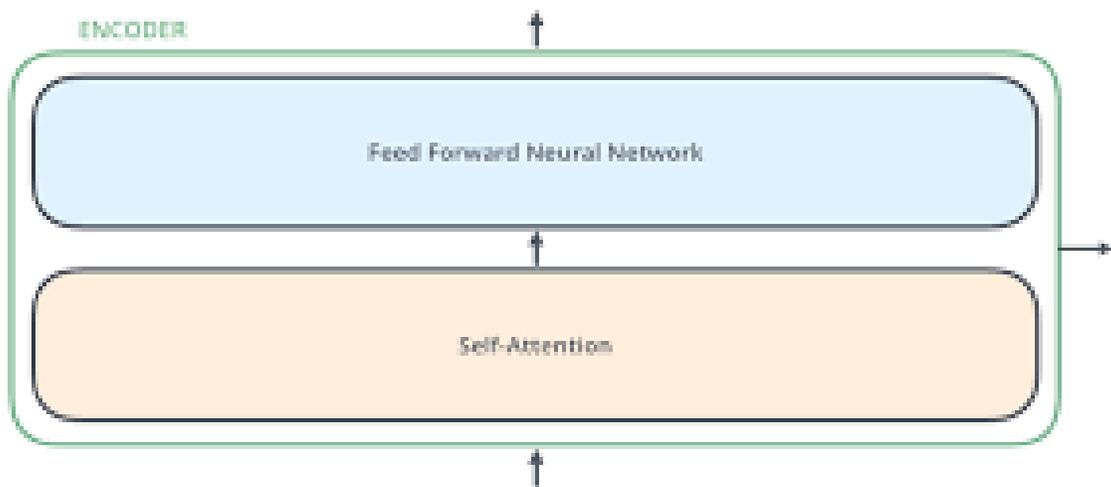




Encoder-Decoder



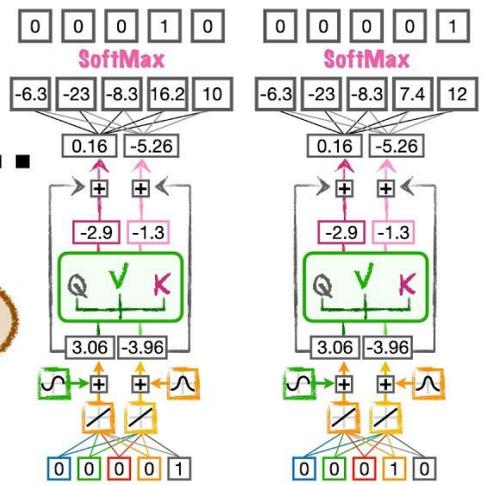
Encoder-only -> BERT



Decoder-Only Transformers...



Decoding is all you need!



...Clearly Explained!!!

I Language Models – Gli N-Grammi

Text: “In consciousness studies the consciousness studies are fundamental”

Unigram

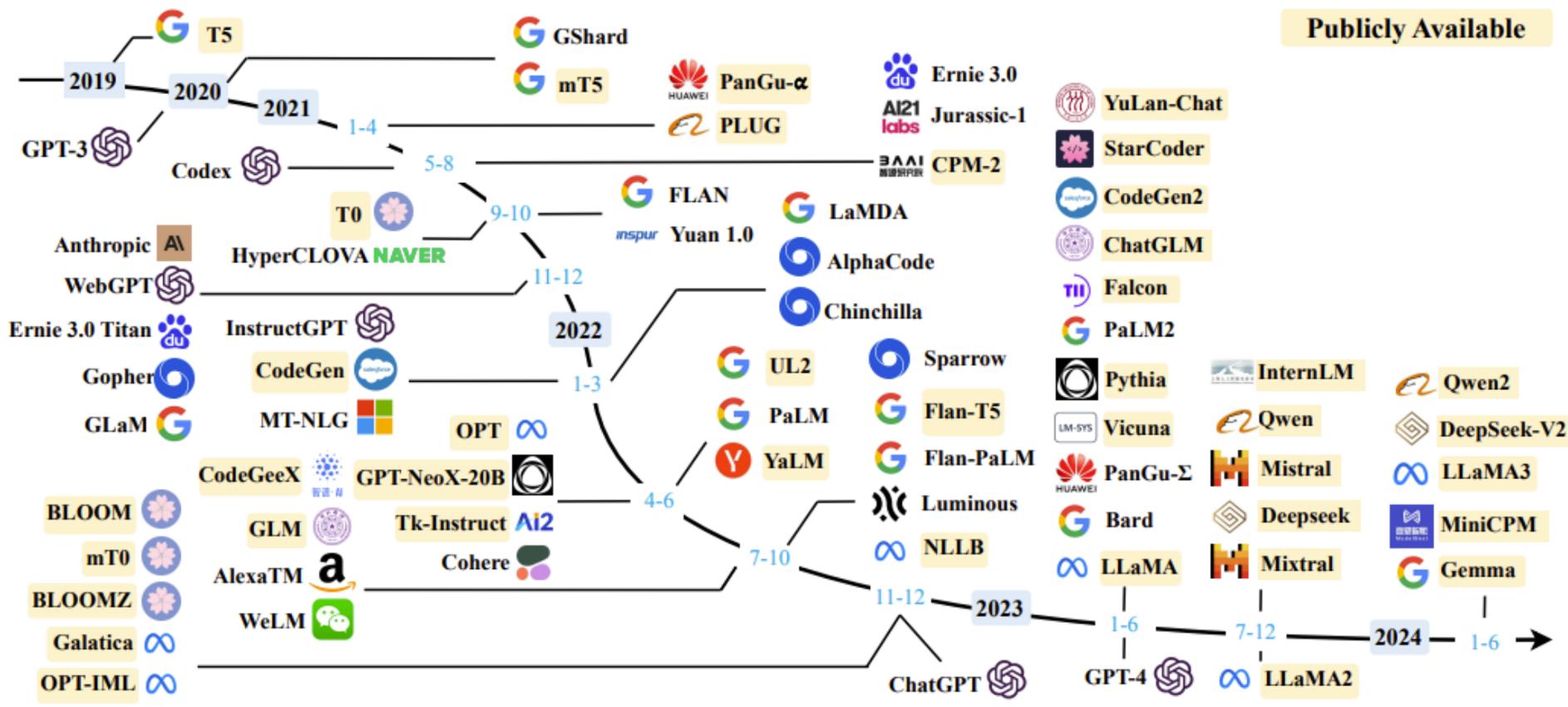
In	consciousness	studies	the	are	fundamental
----	---------------	---------	-----	-----	-------------

Bigram

In consciousness	consciousness studies	studies the	the consciousness	studies are	are fundamental
------------------	-----------------------	-------------	-------------------	-------------	-----------------

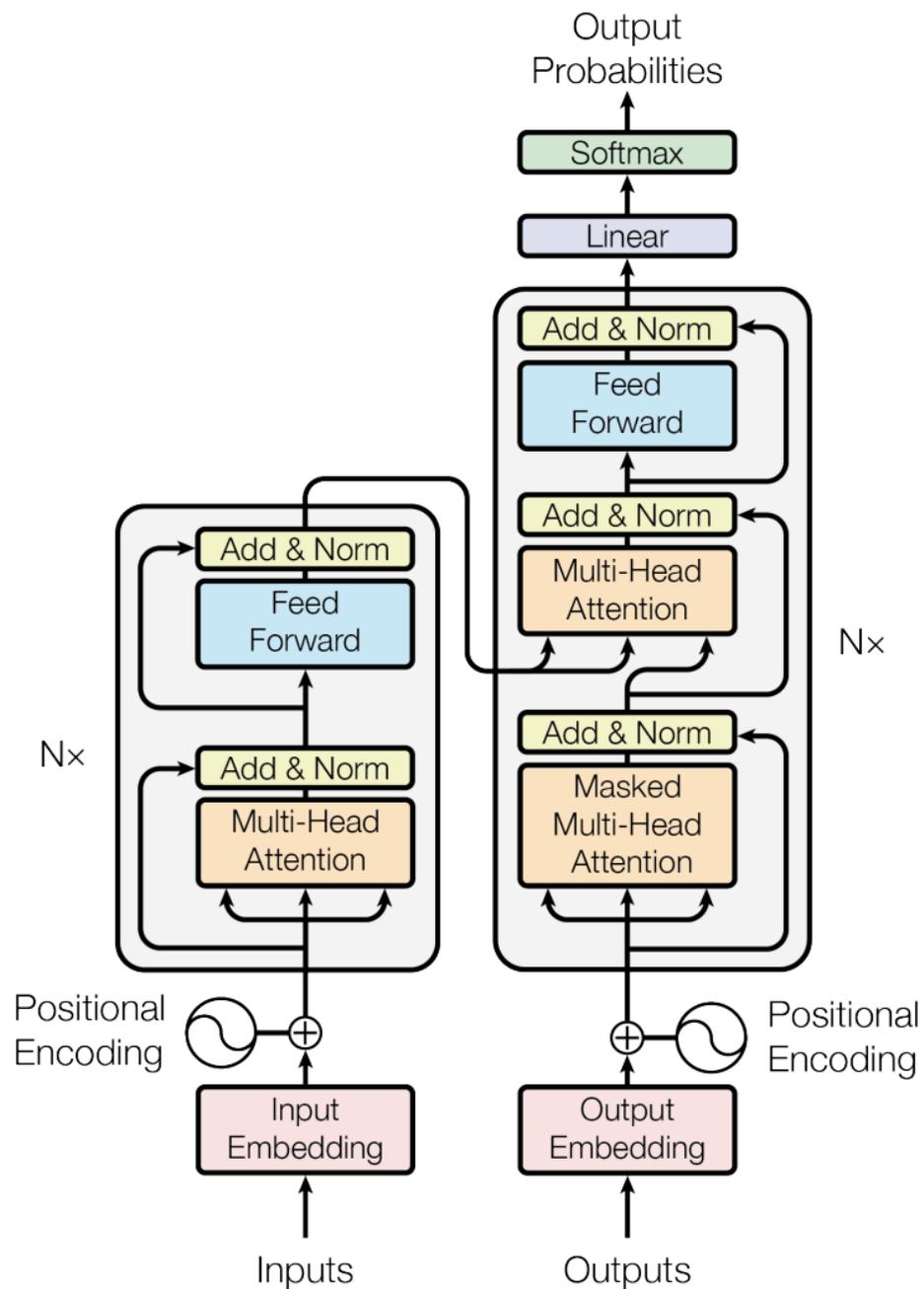
Trigram

In consciousness studies	consciousness studies the	studies the consciousness	the consciousness studies	consciousness studies are	studies are fundamental
--------------------------	---------------------------	---------------------------	---------------------------	---------------------------	-------------------------



BERT

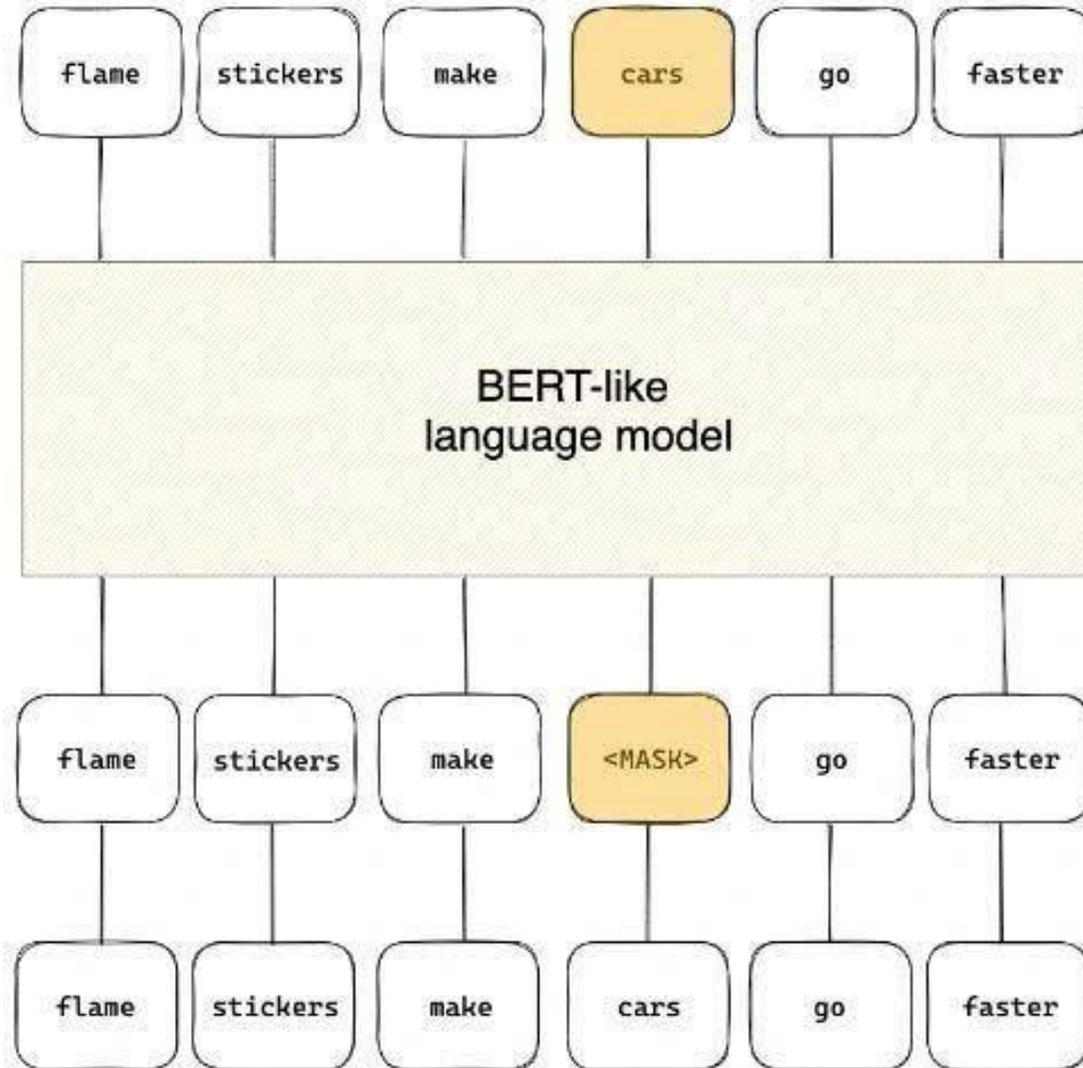
Encoder

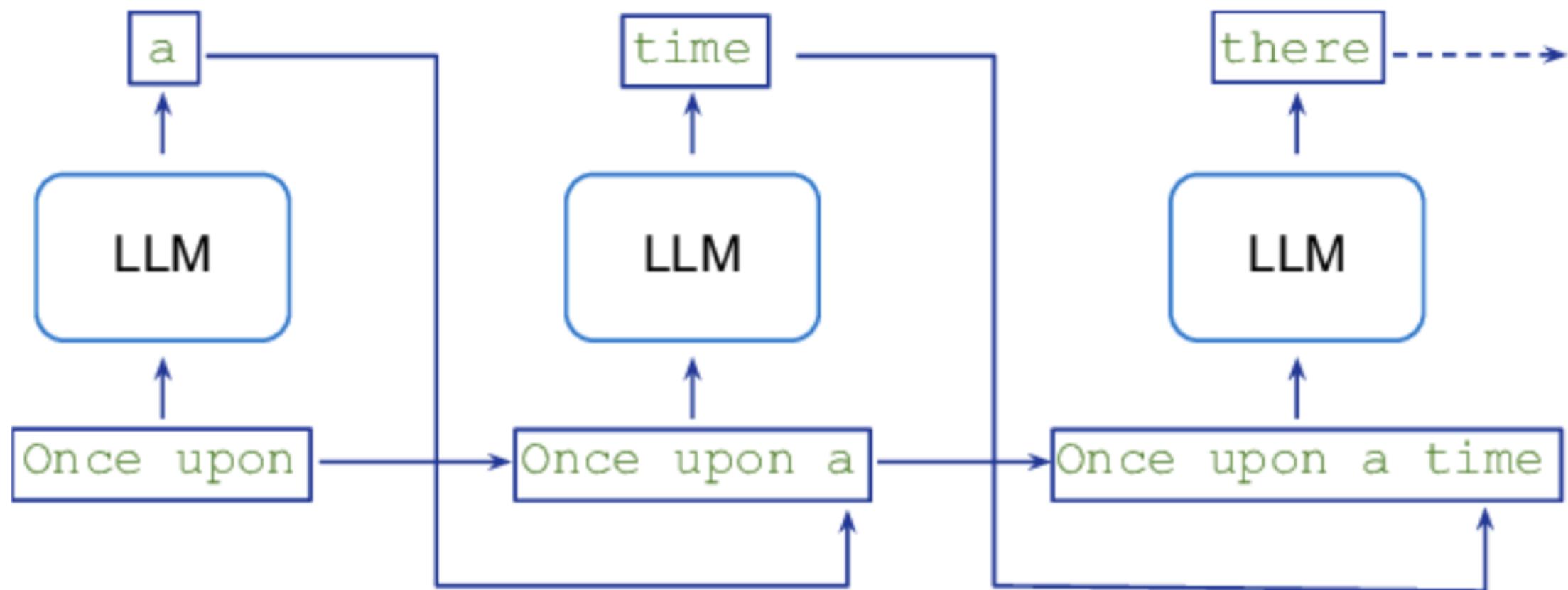


GPT

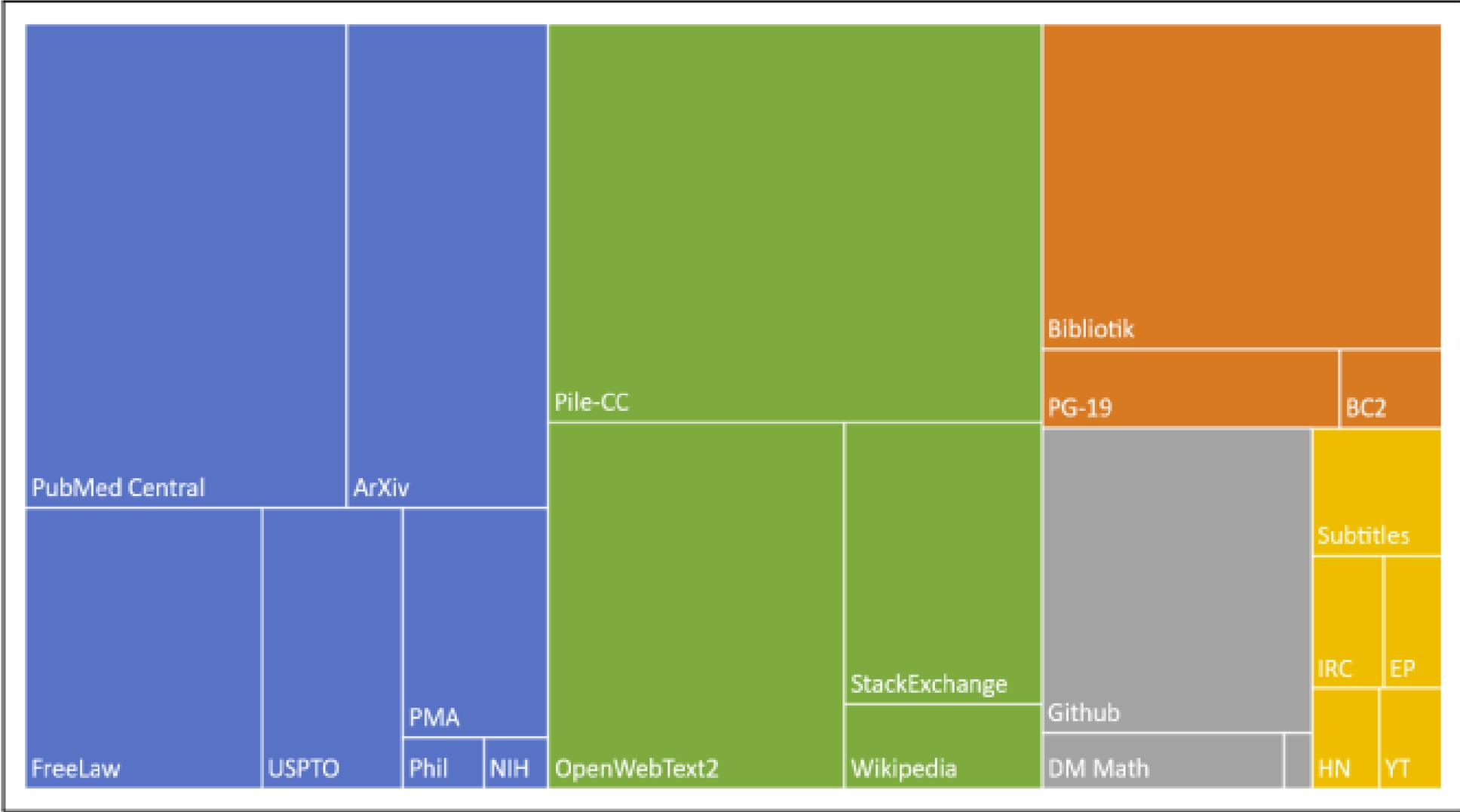
Decoder

Masked Language Modeling MLM





The Pile (825GB)



1. Tempo di lettura per 300 miliardi di parole:

- $300.000.000.000 \text{ parole} \div 150 \text{ parole/minuto} \approx 2.000.000.000 \text{ minuti}$

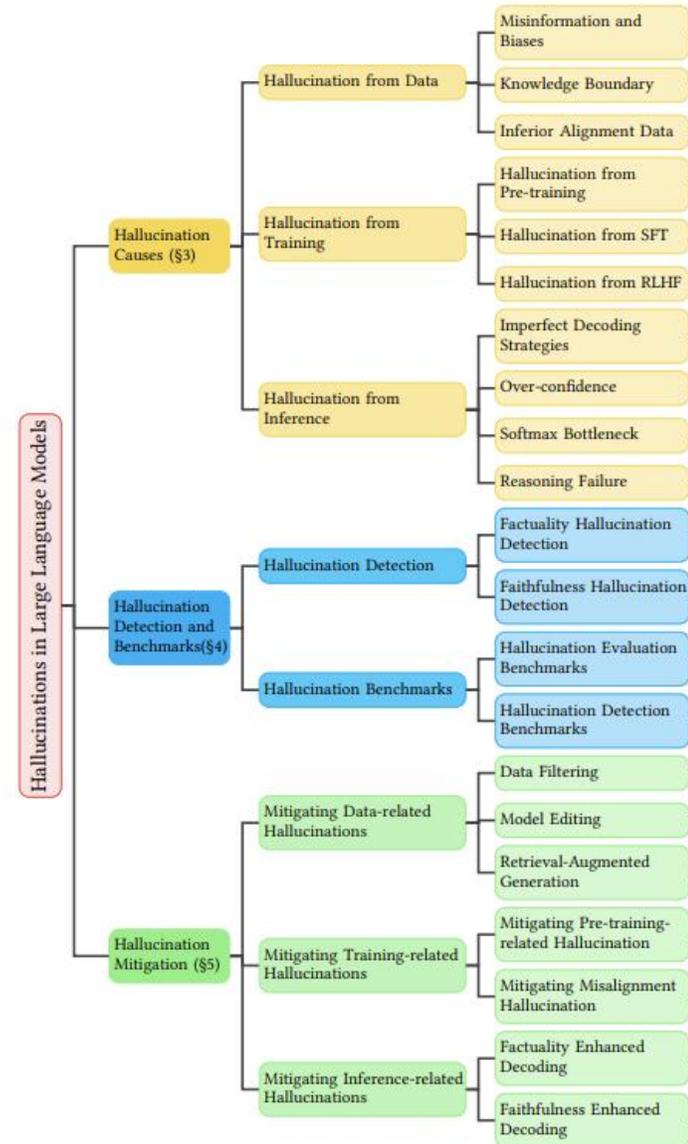
2. Conversione in anni:

- $2.000.000.000 \text{ minuti} \div (60 \text{ minuti} \times 24 \text{ ore} \times 365 \text{ giorni}) \approx \text{circa } 3.800 \text{ anni}$

Risultato:

Ci vorrebbero circa 3.800 anni di lettura continua, ad alta voce, senza pause, per leggere tutto il testo che ha contribuito ad addestrare GPT-3.

Nota: Questa è una stima molto approssimativa e serve solo a dare un'idea della vastità del corpus. La quantità di dati è enormemente grande e rappresenta un insieme di testi molto variegato e complesso.



- Gli LLM manifestano capacità linguistiche superiori alle generazioni precedenti, **senza essere stati specificatamente addestrati a riconoscere e a manipolare categorie e strutture linguistiche**
- Le conoscenze non sono fornite ai LLM né sotto forma di regole, né di dati annotati, ma sono **‘abilità emergenti’** (Wei et al. 2022), poiché nascono come risultato del compito di predizione con cui le reti vengono addestrate
 - ma anche con il contributo **determinante** di raffinamenti a posteriori come instruction tuning, reinforcement learning, ecc.

Scienze del linguaggio e IA *post LLM*

- (2) a. *prompt*: Ieri sera Gianni ha chiesto a Paolo se volesse andare al cinema con lui. Paolo gli ha risposto: “Domani devo alzarmi presto”. Cosa intendeva dire Paolo a Gianni?
b. *ChatGPT*: Paolo ha risposto a Gianni dicendo “Domani devo alzarmi presto”, il che implica che Paolo non può andare al cinema con Gianni la sera stessa. Questo perché deve svegliarsi presto il giorno seguente e probabilmente vuole riposare e andare a letto presto per essere riposato
- (3) a. *prompt*: Qual è il soggetto di “volesse andare” nella frase “Ieri sera Gianni ha chiesto a Paolo se volesse andare al cinema con lui.”?
b. *ChatGPT*: Il soggetto di “volesse andare” nella frase “Ieri sera Gianni ha chiesto a Paolo se volesse andare al cinema con lui.” è “Paolo”. La frase può essere parafrasata come “Ieri sera Gianni ha chiesto a Paolo se Paolo volesse andare al cinema con lui.”

IA linguistica senza scienze del linguaggio?

Gli LLM sono le prime “macchine parlanti” costruite **senza un contributo diretto delle teorie della linguistica sull’organizzazione e il funzionamento del linguaggio umano**

“Quand’è che abbiamo smesso di capire il mondo?”

B. Labatut

- Abbiamo bisogno della linguistica per capire la nuova generazione di “macchine parlanti”, ma questo sarà possibile solo con **nuove teorie del linguaggio umano**
 - i successi e gli errori degli LLM possono portarci ad abbandonare molti degli attuali presupposti sul linguaggio
- Abbiamo bisogno di una teoria del linguaggio umano che sappia essa stessa “parlare” il linguaggio delle macchine: **vettori, matrici e probabilità**
 - è lo stesso linguaggio del sostrato neuronale che sostiene la capacità del linguaggio
- Abbiamo bisogno di una teoria del linguaggio umano che sia più vicina alla **meccanica quantistica** che al cartesianesimo della fisica classica

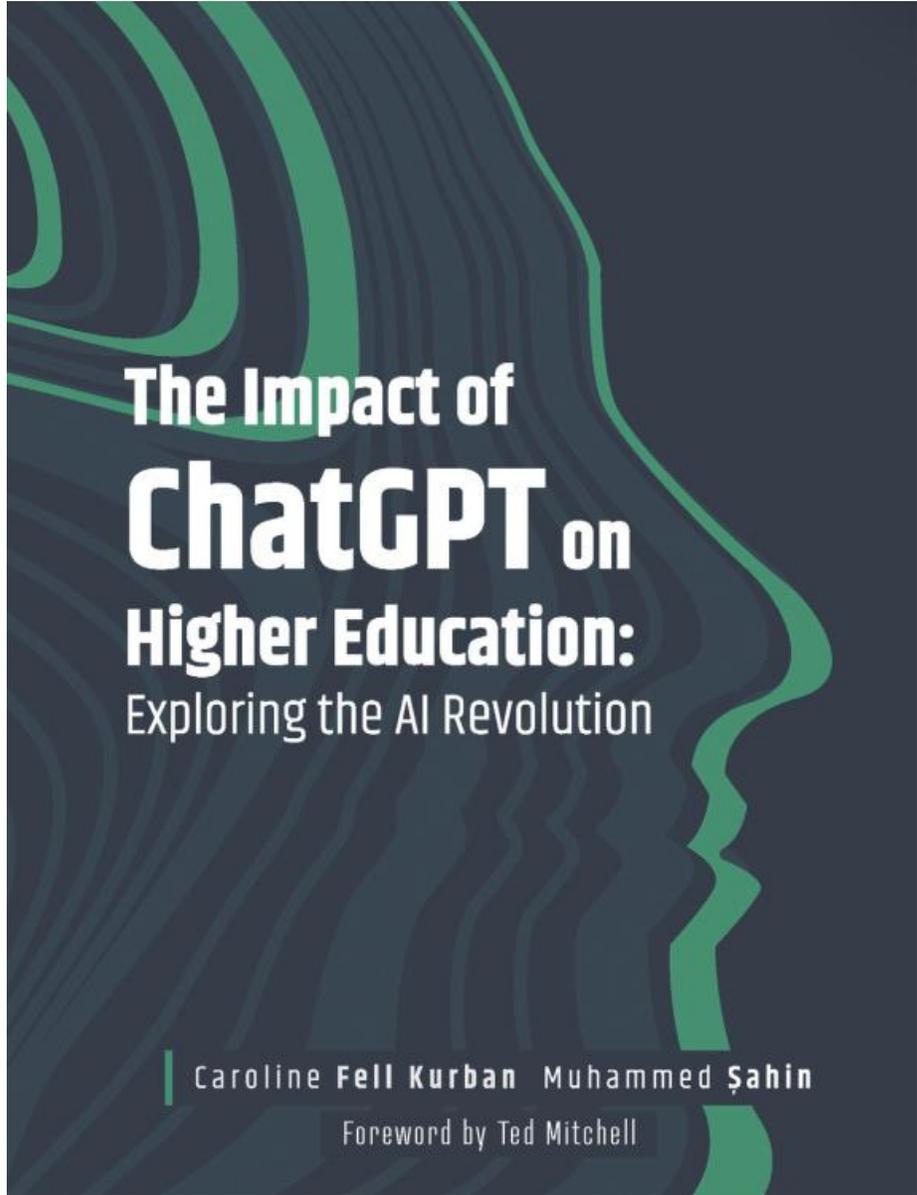
"Come (non) usare l'IA per la didattica accademica"

Franco Cutugno – Università Federico II di Napoli



Centro Interdipartimentale di ricerca
dell'Università Federico II di Napoli -
Urban/Eco





The Impact of ChatGPT on Higher Education: Exploring the AI Revolution

BY

CAROLINE FELL KURBAN

MEF University, Turkey

AND

MUHAMMED ŞAHİN

MEF University, Turkey



United Kingdom – North America – Japan – India – Malaysia – China

2024

The background features a dark blue grid with various colored circles (red, orange, white) scattered across it. On the right side, there are vertical, blurred streaks of light in shades of teal, purple, and orange, suggesting a digital or data-driven environment.

L'impatto delle nuove tecnologie e dei nuovi media

Incalzato ad ogni decennale dalla domanda sull'attualità delle *Tesi*, Tullio De Mauro propone l'apertura ad altre due tesi – una XI e XII *Tesi* – esclusivamente per meglio esplicitare a chi non ha la capacità di adeguare ai tempi e proiettare sull'oggi e sul domani la portata del dettato scientifico e teorico di quanto c'è già scritto.

Le sue osservazioni si soffermano sui flussi informativi che attraverso la televisione prima e la rete poi raggiungono e sommergono chi non è abbastanza educato linguisticamente a farvi fronte. Inoltre, le condizioni di vita portano bambine e bambini a staccarsi dai luoghi di confronto più usuali: il cortile di casa, la strada, il giardino e dal fare materialmente i conti con l'operare manuale. Non che ciò non si possa ritrovare, a ben leggere, nella formulazione originaria delle *Tesi*, ma De Mauro prova a esplicitare:

Si può ragionevolmente delineare una undicesima tesi. Le condizioni dell'informazione che qui ho già evocato mettono a dura prova anche il nostro linguaggio e chiedono che la scuola sappia educarsi e sappia educare a un controllo critico del linguaggio e dei contenuti dell'informazione scritta e variamente trasmessa.

A questa occorre a mio avviso aggiungere una dodicesima tesi [...]. Per una parte notevole del lessico di base c'è un ingresso privilegiato attraverso cui lo apprendiamo ed è un fare, un sentire e operare che ci ha portato a cominciare a costruire il significato di quella parte del lessico. È ciò che ... ho provato a chiamare (rubando il titolo a un tardo saggio di Benedetto Croce) «il primato del fare». Non è che ciò fosse assente nelle Dieci tesi. Ma per vie diverse le società consumistiche in cui viviamo rischiano di far perdere a bambine e bambini l'esperienza del rapporto corporeo con le cose, le realtà della natura, i giochi. Dobbiamo recuperare in pieno il nostro diritto umano a costruirci in questa esperienza. Il senso di troppe parole ci è precluso senza ciò (De Mauro 2018: 39-40).

Da Silvana Ferreri: A 50 anni...

- Dal parlato parlato allo scritto scritto su asse continuo (Nencioni 1976)
- Voghera (2022) multimodalità

Caratteristiche	Parlato	Scritto	Scritto digitale non continuo	Scritto dialogico
Canale	Fonico-uditivo Gestuale-visivo	Grafico-visivo	Grafico-visivo Fonico-uditivo Audiovisivo	Grafico-visivo Fonico-uditivo Audiovisivo
Produzione e ricezione	In presenza	In assenza	In assenza	In assenza
	In tempo reale	Differita	Differita	Differita
	Continua	Discontinua	Discontinua	Discontinua
	Non pianificata	Pianificata	Pianificata	± Pianificata
Rapporto tra produttore e ricevente	Dialogo	Monologo	Monologo	Dialogo/Monologo
Codici	Lingua, gesti	Lingua	Lingua, calcoli, foto, filmati, musica...	Lingua, foto, filmati, musica...
Testo verbale	Discontinuo in turni	Continuo	Discontinuo senza turnazione	Discontinuo in turni



La lingua dei social

La lingua dei chatbot

I messaggi vocali

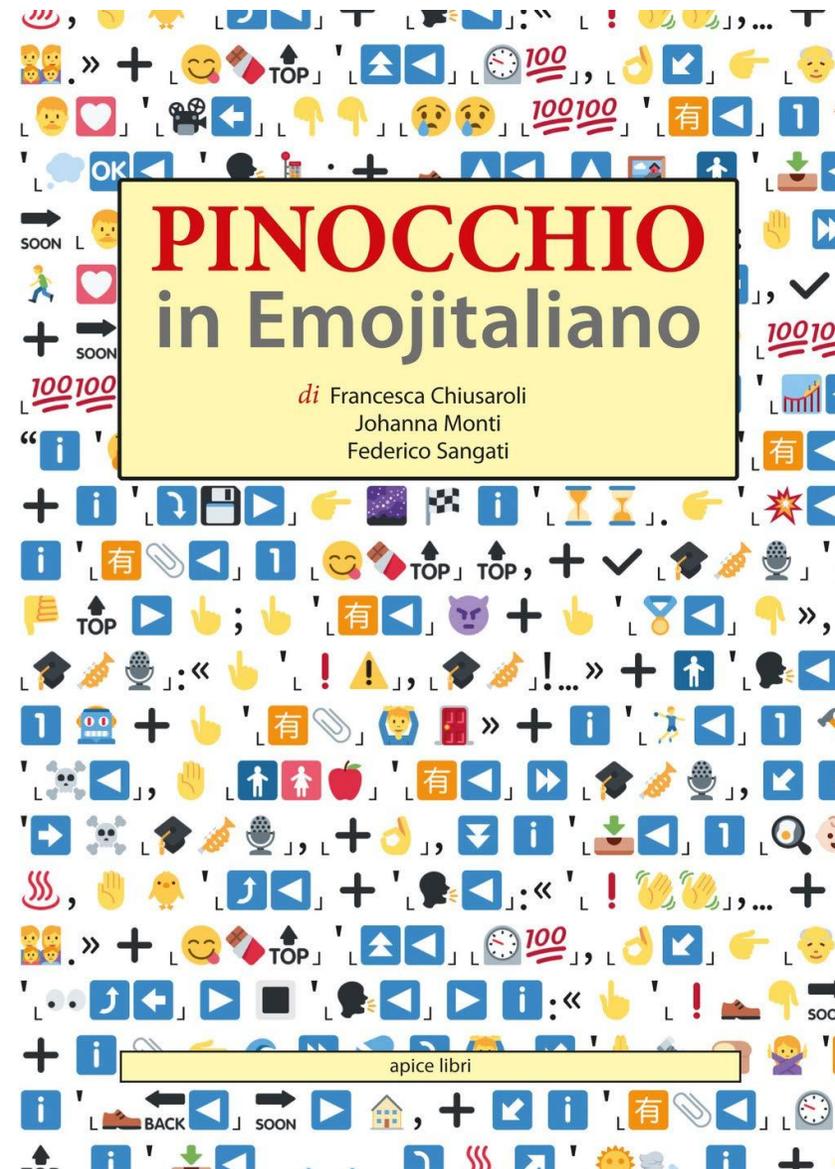
La musica (trap, rap, pop contemporaneo)

Dialettismi e stranierismi estesi (meme, termini resi famosi da serie TV, tormentoni da stand-up, ma anche tanti esempi dalla rete,)

EMOTICONS:

fino alle più
estreme
conseguenze

...





Incel = involuntary celibate

Np4np = naked picture 4 naked picture

Gnoc = get naked on camera

Il vento normativo della IA

- LLM demo



Vantaggi dell'uso degli strumenti IA per l'insegnamento linguistico (e non solo):

Apprendimento personalizzato, feedback immediato



Generazione di contenuti



Autovalutazione



Correzione della pronuncia



Accessibilità, flessibilità coinvolgimento

Svantaggi dell'uso degli strumenti IA ::

Isolamento e mancanza di interazione e socialità



Scarsa attenzione alla creatività ed agli approcci diversificati



Affidamento eccessivo



Bias etici e produzione di stereotipi (genere, posizionamento politico, discriminazione)



Rischio di trasferimento di dati sensibili

Tools

twee

Create, Share and Grade
Language Lessons in Minutes

Diffit
For Teachers

Get student-ready resources for...

[Literally Anything](#) [Books](#) [Article or YouTube Link](#) [Text or PDF](#) [Vocabulary List](#)

 **Wordwall**

Crea lezioni stimolanti e divertenti in poco tempo

 **Pronounce**

PRONOUNCE[®] | Sounds good

1. Le trappole da evitare:

- Sostituzione dell'interazione critica: l'abuso di chatbot come sostituti del dialogo docente-studente;
- Automazione della creatività: l'uso passivo di generatori di testi/analisi che svuotano il processo interpretativo;
- Valutazione algoritmica: sistemi di grading automatizzato inadeguati per saggi argomentativi o pensiero complesso.

2. Criticità epistemologiche:

- - L'"illusione dell'oggettività": quando l'apparente autorevolezza degli output IA maschera bias culturali o superficialità analitica;
- - La crisi della provenienza: difficoltà nel tracciare fonti e processi logici dietro contenuti generati automaticamente;
- - La banalizzazione degli errori: sistemi che producono "mediocrità convincente" (hallucinations) senza segnali d'allarme.

3. Linee guida per un uso consapevole:

- Principio della trasparenza: dichiarare esplicitamente quando e come si usa l'IA nel percorso didattico;
- Modello "critico-collaborativo": impiegare strumenti IA come contrappunto da smontare (es.: confrontare analisi letteraria umana e artificiale);
- Focus sul processo: valutare non il prodotto finale bensì l'iter critico dello studente (bozze, scelte metodologiche, revisione degli output IA).

Grazie per l'attenzione!!!!

**When I realize
ChatGPT can do
my job for me**



**When I realize
ChatGPT can do
my job for me**



When I realize how ChatGPT made my job