

L'italiano e le sue varietà nell'Intelligenza Artificiale



Giuseppe Samo

Professore di Chiara Fama (特聘教授),
Università di Lingua e Cultura di Pechino

Comites & Università di Ginevra,
23.10.2026



SETTIMANA
DELLA LINGUA
ITALIANA
NEL MONDO



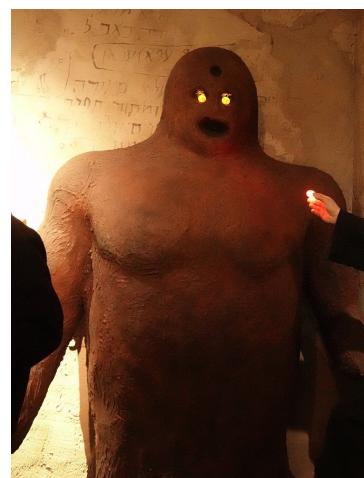
"Macchine" e linguaggio nelle arti



*L'automa descritto
da Liezi*



3CPO



il Golem

“Macchine” parlanti *dell’italiano* nelle arti

fonte: <https://upload.wikimedia.org/wikipedia/commons/5/5d/Diogenes.jpg>



fonte: https://upload.wikimedia.org/wikipedia/it/9/9b/ito_e_Caterina.jpg



Intelligenza Artificiale?

Spesso utilizziamo l'espressione Intelligenza Artificiale (IA) per riferirci, più specificamente, a sistemi basati su modelli linguistici di grandi dimensioni come ad esempio ChatGPT, Gemini, Claude o DeepSeek.

I modelli su cui sono basati questi sistemi sono multilingua e di conseguenza possiamo analizzare al loro interno l'italiano e altre varietà parlate in Italia.

Struttura della presentazione

1. *Cosa sono i modelli linguistici di grandi dimensioni*
2. *L'italiano **sintetico** e le varietà dell'italiano nei modelli linguistici*

Modello linguistico di grandi dimensioni

Che cosa significa e che cosa è?

Scomponiamo il nome e spieghiamo brevemente ciascun elemento.

Modello *linguistico di grandi dimensioni*

Un modello è una rappresentazione di un oggetto di studio attraverso un insieme di attributi e dei rispettivi valori

$$X_1, X_2, X_3, \dots, X_n = Y$$

📖 Merlo, P. & Samo, G. (to appear) "Generative Computational Modeling", in The Cambridge Handbook of Minimalism and Its Interdisciplinary Applications, ed. by Evelina Leivada and Kleanthes K. Grohmann, Cambridge University Press.

Modello *linguistico di grandi dimensioni*

Facciamo un esempio pratico:

$$X_1, X_2, X_3, \dots, X_n = Y$$

Spaghetti + *Guanciale* + *Uovo* + *Pecorino* + *Pepe* = Carbonara

Spaghetti + *Guanciale* + $\emptyset$ + *Pecorino* + *Pepe* = Gricia

Spaghetti + *Guanciale* + *Pomodoro* + *Pecorino* + $\emptyset$ = Amatriciana

Spaghetti + $\emptyset$ + $\emptyset$ + *Pecorino* + *Pepe* = Cacio e pepe



Modello linguistico *di grandi dimensioni*

Modello linguistico è un *termine tecnico* (Hale 2016)

Il concetto deriva dal compito di prevedere una parola in un testo.

I fenicotteri sono _____

belli

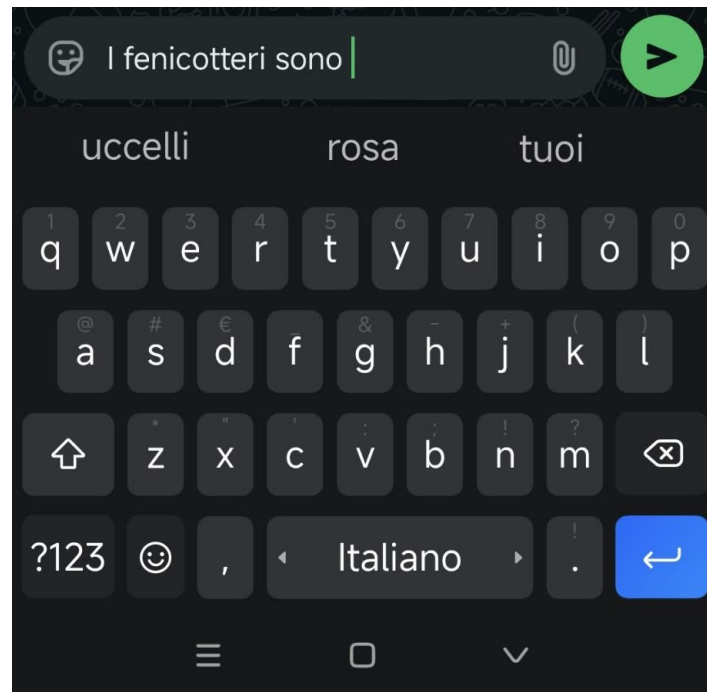
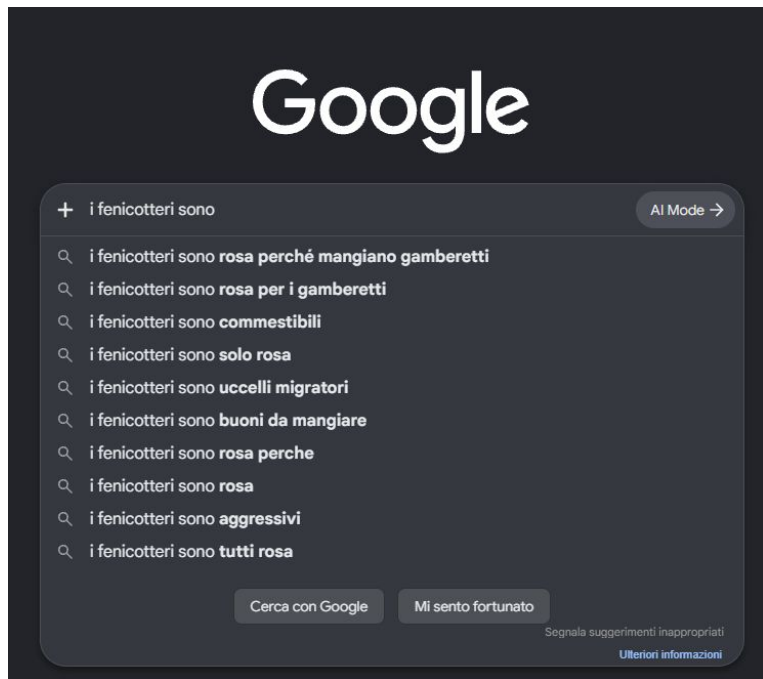
rosa

uccelli

termosifone



Modello linguistico *di grandi dimensioni*



Modello linguistico *di grandi dimensioni*

I fenicotteri sono _____

x_1 x_2

x_3

y

belli

rosa

uccelli

termosifone

$$P(y \mid x^1, x^2, \dots, x^n)$$

dove y può essere una parola appartenente a un vocabolario $V = \{w_1, w_2, \dots, w_{|V|}\}$



Modello linguistico *di grandi dimensioni*

Nei modelli linguistici, questa probabilità può variare in base ai dati di addestramento a cui il modello viene esposto per imparare le associazioni.

		Enciclopedia	Viaggi	Libri infanzia
				
I fenicotteri sono _____	belli	0.02	0.35	0.20
	rosa	0.05	0.15	0.43
	uccelli	0.39	0.10	0.06
	termosifone	0.00	0.00	0.00

Modello linguistico *di grandi dimensioni*

Una rivoluzione arriva con *l'attenzione* e la *bidirezionalità*

(si veda Jurafski & Martin 2024, ch. 9); BERT (Devlin et al. 2019)

I fenicotteri sono _____ che possono vivere nei pressi di una salina.

belli

rosa

uccelli

termosifone



Modello linguistico *di grandi dimensioni*

Una parola, una parte di parola, o più di una parola:
la tokenizzazione e le rappresentazioni interne di un modello

BERT Italiano

Frase: Un bellissimo evento in collaborazione con i comites di Ginevra

Token: ['Un', 'bellissimo', 'evento', 'in', 'collaborazione', 'con', 'i', 'comi', '##tes', 'di', 'Gine', '##vra']

Numero di token: 12

ID dei token: [722, 8929, 6681, 139, 3863, 153, 134, 12703, 4563, 120, 25536, 7035]

BERT Multilingua

Frase: Un bellissimo evento in collaborazione con i comites di Ginevra

Token: ['Un', 'belli', '##ssimo', 'evento', 'in', 'collaborazione', 'con', 'i', 'com', '##ites', 'di', 'Gi', '##nev', '##ra']

Numero di token: 14

ID dei token: [11038, 78183, 42822, 20720, 10106, 41052, 10173, 177, 10212, 28754, 10120, 38356, 83310, 10288]

Modello linguistico *di grandi dimensioni*

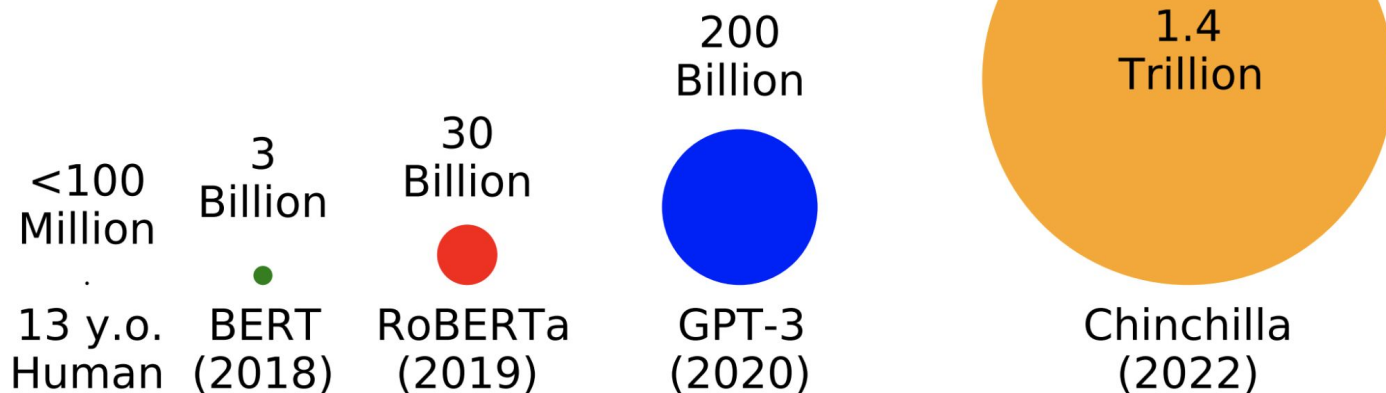
```
Forma della matrice: (14, 768)
File salvato: italian_bert_vectors.csv
```

	token	token_id	0	1	2	3	4
	[CLS]	102	-0.005097	0.042588	-0.024434	0.044958	-0.077487
	Un	722	0.074703	0.017132	-0.358274	0.188549	-0.144034
	bellissimo	8929	0.273304	0.064859	0.083723	-0.491106	0.207631
	evento	6681	-0.133753	0.346777	-0.064417	-0.622220	-0.203355
	in	139	0.200666	0.091594	0.027247	-0.375120	-0.059133
	collaborazione	3863	-0.708700	0.393967	0.135935	-0.680767	0.379089
	con	153	-0.150687	0.083300	0.587386	0.290390	-0.404124
	i	134	-0.020834	-0.283211	0.499007	0.806431	-0.548580
	comi	12703	-0.740156	0.308442	0.559193	-0.124940	-0.076047
	##tes	4563	-0.985555	0.037771	0.791370	-0.051948	0.297620
	di	120	-0.515737	0.346348	-0.021173	0.085814	0.028986
	gine	25536	0.201924	0.375403	-0.217842	-0.591834	0.215834
	##vra	7035	-0.176500	0.216424	-0.321684	-0.938280	1.167957
	[SEP]	103	-0.004101	0.126174	-0.215794	-0.352616	0.328968

Modello linguistico di grandi dimensioni

Un'altra rivoluzione arriva con le dimensioni dei dati di addestramento

(grafico da <https://babylm.github.io/>)



Modello linguistico di grandi dimensioni

Con *modello linguistico di grandi dimensioni* ci riferiamo al modello vero e proprio, addestrato su grandi quantità di dati testuali e basato su un'architettura che permette di elaborare dati linguistici.

Tuttavia, nel dibattito pubblico e anche in quello accademico al di fuori della linguistica e della linguistica computazionale, il termine, insieme al solo iperonimo **IA** o **AI**, viene spesso utilizzato per riferirsi alle piattaforme che consentono di interagire con questi modelli attraverso istruzioni (*prompt*), come ChatGPT, DeepSeek, etc..

Piccola parentesi: Manzoni, Ascoli e l'italiano



L'italiano *sintetico* che dovrebbe emergere richiama, in un certo senso, il sogno ascoliano di una lingua comune che si sviluppa dal basso.

Una lingua sintentica

De Cesare (2026) utilizza l'aggettivo **sintetico**, in contrasto con lingua naturale, riportando una definizione del Nuovo De Mauro:

“prodotto ottenuto artificialmente mediante sintesi chimica, che può essere uguale al suo corrispettivo naturale oppure non esistere affatto in natura” (Nuovo De Mauro, s.v. sintetico)

Per una voce di glossario tecnico, in inglese, si veda Samo (2026, Samo 2026, <https://ling.auf.net/lingbuzz/010447>)

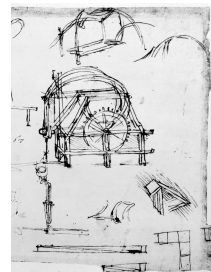


Dati naturali e sintetici

Una parte della comunità scientifica (e.g. Piantadosi 2023) considera i modelli linguistici una possibile rappresentazione del linguaggio umano.

Un'altra parte (Katzir 2023, Moro et al. 2023, Shi et al. 2025) mette in dubbio che i dati sintetici possano fornire evidenze significative sul linguaggio come sistema cognitivo (si veda anche Chesi 2025).

Studiare i dati naturali e sintetici come due **oggetti di indagine complementari** (Rizzi 2025).



L'italiano sintetico

Prima di passare ai casi di studio sull'italiano sintetico è **importante capire su quali dati vengono *addestrati* i modelli.**

L'italiano sintetico

Esistono modelli *Transformer* **monolingua** per l'italiano, ma sono generalmente più piccoli (ad esempio, Electra-italian, **13GB** and 2,050,057,573 tokens).

I modelli linguistici di grandi dimensioni, invece, sono spesso multilingua poiché vengono addestrati su testi in più lingue.

Anche modelli come *Minerva*, sviluppato per l'italiano, includono materiale in inglese nei dati di addestramento (Orlando et al. 2024).



Il ruolo dell'inglese

L'inglese è fortemente presente nei dati di addestramento.

A destra possiamo vedere i dati per Teuken-7b, un modello "europeo".

Per **Apertus**, un modello aperto svizzero, 60% dei dati di addestramento è dato dall'inglese e 40% da altre lingue. (Apertus project 2025, <https://arxiv.org/abs/2509.14233v2>)

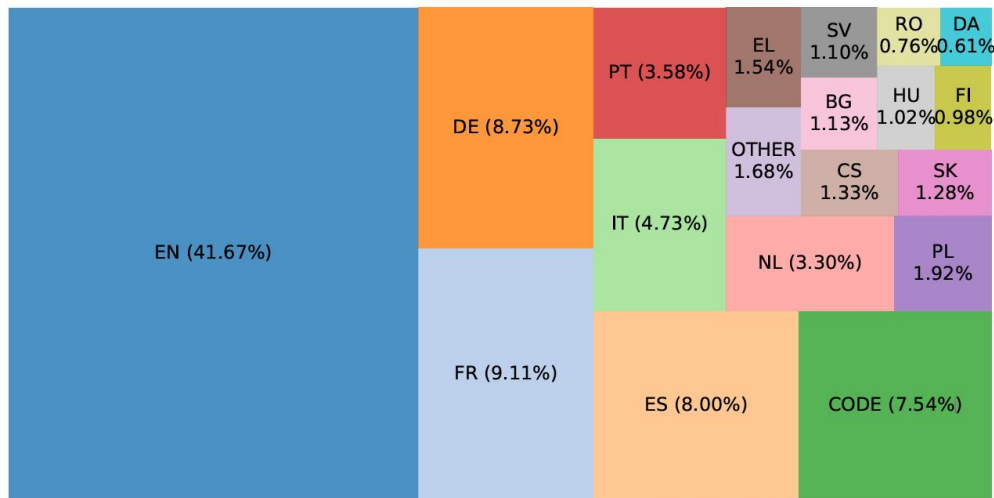


Figure 1: Language distribution of the tokenized dataset, comparing the presence of English and other European languages.

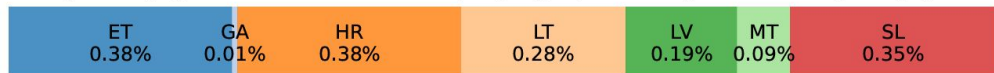


Figure 2: Breakdown of the "OTHER" category.

Da Ali et al. (2025: 6, Figura 1 e 2;
<https://arxiv.org/abs/2410.03730v3>)

Il ruolo dell'inglese

L'inglese quindi vien spesso viene etichettato come lingua *pivot* (Wendler et al. 2025).

Si noti che le istruzioni formulate in inglese possono influenzare le caratteristiche degli output generati in italiano (Cicero 2023; Nastase et al. 2025).

Analizzare l'italiano sintetico

Una possibile metodologia consiste nell'analizzare gli output dei modelli linguistici con gli strumenti della linguistica dei corpora (Mikros 2025).

Questo approccio permette di confrontare dati naturali e sintetici, studiandone le caratteristiche stilistiche ed eventuali i bias grammaticali (De Cesare 2023, 2024).

Questa metodologia può far emergere anche altri *artefatti culturali o sociali*.



Alcune caratteristiche dell'italiano sintetico

Quelle marcate con  sono anche chiamate impronte dell'inglese

- Errori di accordo di genere e numero (*La squadra è noto per il suo calcio*)
- Maiuscole non convenzionali, come nei nomi dei mesi o le lingue (*Giugno o Italiano*)
- Virgola seriale prima della congiunzione (*il pane, i pomodori, e la mozzarella*)
- Neologismi e forme ibride o anomale, come *omittare*
- Anglicismi e prestiti non adattati (*legato* da 'legacy' per eredità)
- Preferenza per la paratassi e la coordinazione e forme grammaticali anomale
- Verbosità, ridondanza e ripetitività
- Cortesia convenzionale ed effetti di *compiacenza* (cf. sycophancy, Cheng et al. 2026)

L'italiano sintetico: il modello come un informante

In un lavoro “preistorico”, Massaro e Samo (2023), abbiamo utilizzato l'output di ChatGPT come quello di un informante, seguendo un approccio discusso anche in Haider (2023).

Ad esempio, ChatGPT giudicava arcaiche o antiquate strutture come *Il libro, Mario lo ha preso*, in cui il complemento oggetto è collocato all'inizio della frase attraverso una costruzione di topicalizzazione.



Si può dire: 'il libro Mario lo ha preso'?



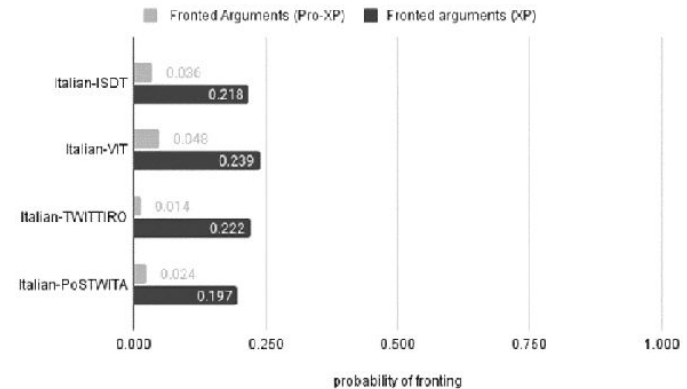
Sì, si può dire "Il libro Mario lo ha preso".

Questa è una forma corretta della costruzione italiana della frase per indicare che Mario ha preso il libro. Tuttavia, va notato che questa è una costruzione molto formale e a volte può risultare un po' arcaica o antiquata. In italiano moderno, è più comune usare la costruzione "Mario ha preso il libro", che è più semplice e diretta.



L'italiano sintetico: dati selvatici sintetici

Tuttavia, se controlliamo i corpora annotati sintatticamente, come quelli delle dipendenze universali, osserviamo che in italiano la topicalizzazione dell'oggetto diretto è tutt'altro che marginale: **si riscontra in quasi una frase su cinque**, attraverso diversi registri e generi testuali.



Il progetto CALAMITA

Con oltre 80 autori e un'ampia varietà di task, l'iniziativa CALAMITA (si veda l'articolo pubblicato su IJCoL) offre un quadro articolato per valutare le capacità dei modelli linguistici in italiano, dalla sintassi alla morfologia, dalla pragmatica agli usi linguistici e alla traduzione.



12-1 | 2026

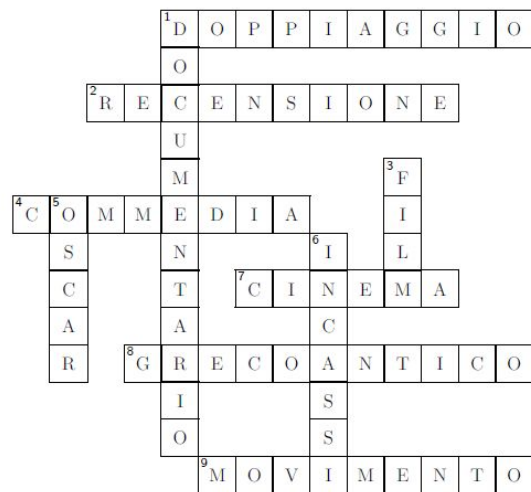
Italian Journal of Computational Linguistics vol. 12, n. 1 June 2026

Challenging the Abilities of Large Language Models in Italian: a Community Initiative

Malvina Nissim, Danilo Croce, Viviana Patti, Pierpaolo Basile, Giuseppe Attanasio, Elio Musacchio, Matteo Rinaldi, Federico Borazio, Maria Francis, Jacopo Gili, Daniel Scalena, Begoña Altuna, Ekhi Azurmendi, Valerio Basile, Luisa Bentivogli, Arianna Bisazza, Marianna Bolognesi, Dominique Brunato, Tommaso Caselli, Silvia Casola, Maria Cassese, Mauro Cettolo, Claudia Collacciani, Leonardo De Cosmo, Maria Pia Di Buono, Andrea Esuli, Julen Etxaniz, Chiara Ferrando, Alessia Fidelangeli, Simona Frenda, Achille Fusco, Marco Gaido, Andrea Galassi, Federico Galli, Luca Giordano, Mattia Goffetti, Itziar Gonzalez-Dios, Lorenzo Gregori, Giulia Grundler, Sandro Iannaccone, Chunyang Jiang, Moreno La Quatra, Francesca Lagioia, Soda Marem Lo, Marco Madeddu, Bernardo Magnini, Raffaele Manna, Fabio Mercorio, Paola Merlo, Arianna Muti, Vivi Nastase, Matteo Negri, Dario Onorati, Elena Palmieri, Sara Papi, Lucia Passaro, Giulia Pensa, Andrea Piergentili, Daniele Poterti, Giovanni Puccetti, Federico Ranaldi, Leonardo Ranaldi, Andrea Amelio Ravelli, Martina Rosola, Elena Sofia Ruzzetti, Giuseppe Samo, Andrea Santilli, Piera Santin, Gabriele Sarti, Giovanni Sartor, Beatrice Savoldi, Antonio Serino, Andrea Seveso, Lucia Siciliani, Paolo Torroni, Rossella Varvara, Andrea Zaninello, Asya Zanollo, Fabio Massimo Zanzotto, Kamyar Zeinalipour and Andrea Zugarini

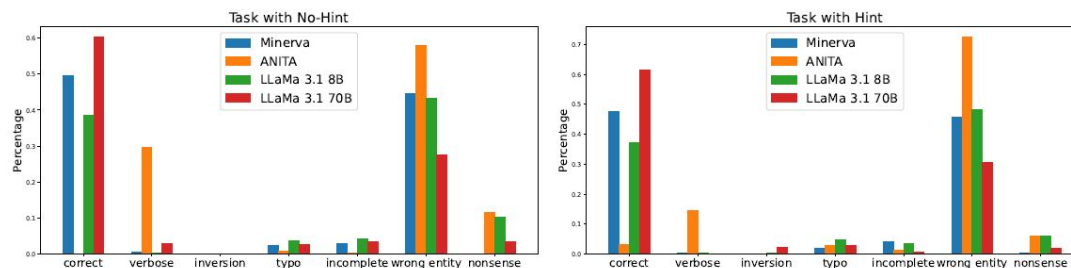


Bravi con compiti linguistici....



Orizzontali 1 * Traduzione Simultanea 2 * Una valutazione critica 4 * Lo è uno spassoso racconto 7 Insieme delle arti, tecniche e attività industriali produttive di un film 8 Lingua originale da cui deriva la parola cinema 9 Uno dei termini estratti dall'antica lingua greca, utilizzato per descrivere il cinema **Verticali** 1 * Un film... come documento 3 Prodotto commerciale finale di un insieme di lavoro comprendente arti, tecniche e attività industriali 5 * Un premio assai ambito 6 * Entrano nelle casse del botteghino

Gli LLM sono molto bravi nei compiti prettamente linguistici, come ad *inter alia*, risolvere o generare cruciverba (Zeinalipour et al. 2025)



da Nissim et al. (2025: 54, figura 4)

Un esempio di errore e di bias (Nissim 2025: 54, tabella 8)

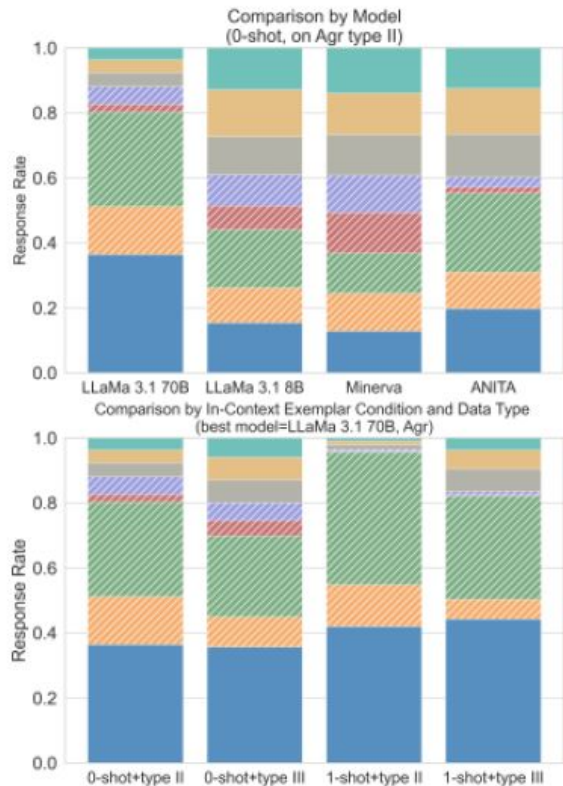
Definizione: Attaccante del Napoli e della nazionale belga

Risposta LLM: Mertens

Risposta "target": Lukaku

da Zeinalipour et al. (2025: 65, figura 6)

meno bravi con compiti più complessi...



Ad esempio puzzle linguistici più complessi dove bisogna generalizzare delle regole linguistiche, come le BLM (Merlo et al. 2026) sviluppato e poi condotto in parte qui all'Università di Ginevra.

Le varietà dell'italiano

L'Italia è caratterizzata da una lunga tradizione di diversità linguistica.

Ma questa variazione è rappresentata nei modelli linguistici?

<https://www.ditals.com/varietà-diatopiche/>



Le varietà d'italiano sintetiche

Già in Massaro & Samo (2023) avevamo utilizzato il modello come informante del napoletano e informante della varietà parlata a San Marco in Lamis (FG), evidenziando difficoltà dei modelli nel cogliere fenomeni specifici di queste varietà, come il raddoppiamento del soggetto.

Chill Giuann è nu buon guaglione

Per esempio, di fronte al prompt commenta questa frase: il modello interpretava letteralmente *chillo* come il dimostrativo *quello*, perdendo la funzione del dimostrativo nella costruzione topicalizzata.

Le varietà d'italiano sintetiche

I dati di San Marco in Lamis (FG)

Quiddu lu canə l'ea spja' ijə
'al cane ci devo badare io'

Nell'articolo del 2023, venivano etichettati come "napoletano".
Qui invece un prompt di ieri (22.10.2026) con ChatGPT

Quiddu lu canə l'ea spja' ijə
mi commenti questa frase?

La frase «*Quiddu lu canə l'ea spja' ijə*» sembra una frase dialettale dell'Italia meridionale, ma dalla grafia non riesco a identificare con sicurezza la varietà né a ricostruire una traduzione affidabile.

Per commentarla dal punto di vista linguistico, mi servirebbe sapere una cosa: di **quale dialetto o zona proviene?** Per esempio, salentino, calabrese, siciliano, napoletano o un'altra varietà?

Le varietà parlate in italia

In Samo & Massaro (2025) invece abbiamo analizzato queste costruzioni chiedendo a ChatGPT e DeepSeek di valutare le frasi su una scala numerica. Questi risultati poi vengono comparati ai risultati in letteratura, ma anche a un gruppo di controllo fatto con transformers.



The screenshot shows the AI Ling journal website. At the top is a pink header with the 'AI Ling' logo and navigation links: 'About', 'Publication', 'Current', and 'Archives'. Below the header, a breadcrumb trail reads '← Archives / Special Issue / Vol. 2 No. 2'. The main article title is 'Syntactic patterns of Italo-Romance CP-layers in Transformers, ChatGPT and Deepseek: two case studies from Romansh and Neapolitan'. The authors are listed as 'Giuseppe Samo, Angelapia Massaro', with a link for '+ Author Details'. The DOI is 'https://doi.org/10.62408/ai-ling.v2i2.25', and the keywords are 'left periphery, cartography, ChatGPT, DeepSeek, Transformers, Italian'. The publication date is 'December 22, 2025'. A dark blue button labeled 'Samo, Massaro_2025_AI-Linguistica' is visible. At the bottom, the 'AI Ling' logo is repeated, along with the text 'Published in Vol. 2 No. 2 (2025) of AI-Linguistica: Linguistic Studies on AI-Generated Texts and Discourses'.

AI Ling

About Publication Current Archives

← Archives / Special Issue / Vol. 2 No. 2

Syntactic patterns of Italo-Romance CP-layers in Transformers, ChatGPT and Deepseek: two case studies from Romansh and Neapolitan

Giuseppe Samo, Angelapia Massaro

+ Author Details

DOI <https://doi.org/10.62408/ai-ling.v2i2.25>

Keywords left periphery, cartography, ChatGPT, DeepSeek, Transformers, Italian

Published December 22, 2025

Published in Vol. 2 No. 2 (2025) of AI-Linguistica: Linguistic Studies on AI-Generated Texts and Discourses

Samo, Massaro_2025_AI-Linguistica

Il raddoppiamento del soggetto in napoletano

Grammaticali

Chill o'rre sunava o'mandulino

Chell a'rreggina sunave o'mandulino

Chill o'rre sunava a chitarra

Chell a'rreggina sunava a chitarra

Non grammaticali

Chell o'rre sunava a' chitarra

Chill a'rreggina sunava a chitarra

Chell o'rre sunava o'mandulino

Chill a'rreggina sunava o'mandulino

Istruzioni: Potresti valutare le frasi con una valutazione da 1 a 3

Accordo di numero nel raddoppiamento del soggetto

BERT multilingua

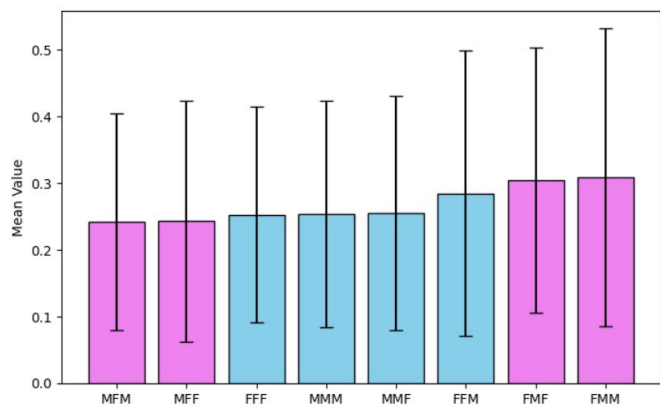


Figure 4. Mean rating of sentences and standard deviation with gender mismatches between subjects and direct objects in Neapolitan.

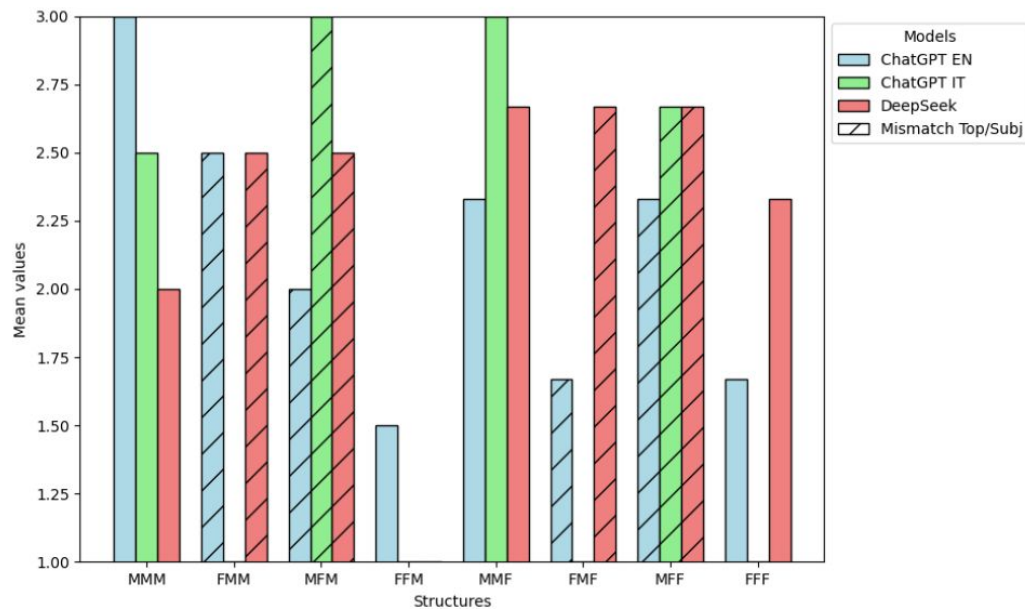


Figure 5: Mean rating (1-3) for each model.

Geonimi nelle varietà dell'italiano

In Samo & Ursini (2023) abbiamo estratto da mappe nomi di geonimi isolati ad alcune aree geografiche italiane e alla loro composizione urbana. E.g. *calle* a Venezia



Geonimi nelle varietà dell'italiano

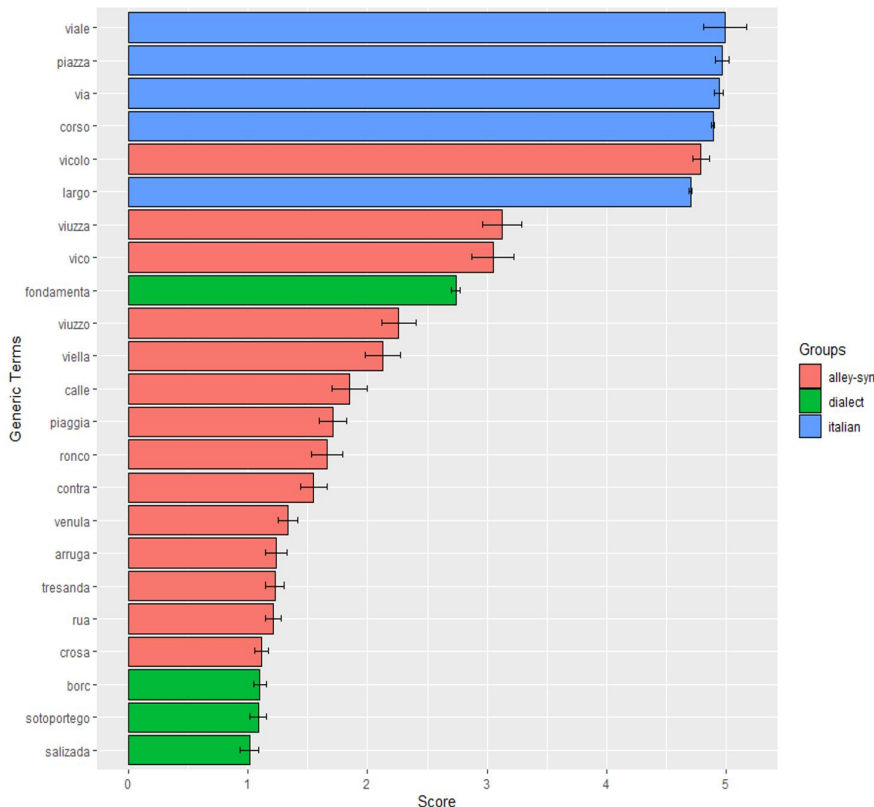
Abbiamo chiesto a 80 partecipanti tramite un questionario online di valutare se un geonimo fosse più o meno dialettale.



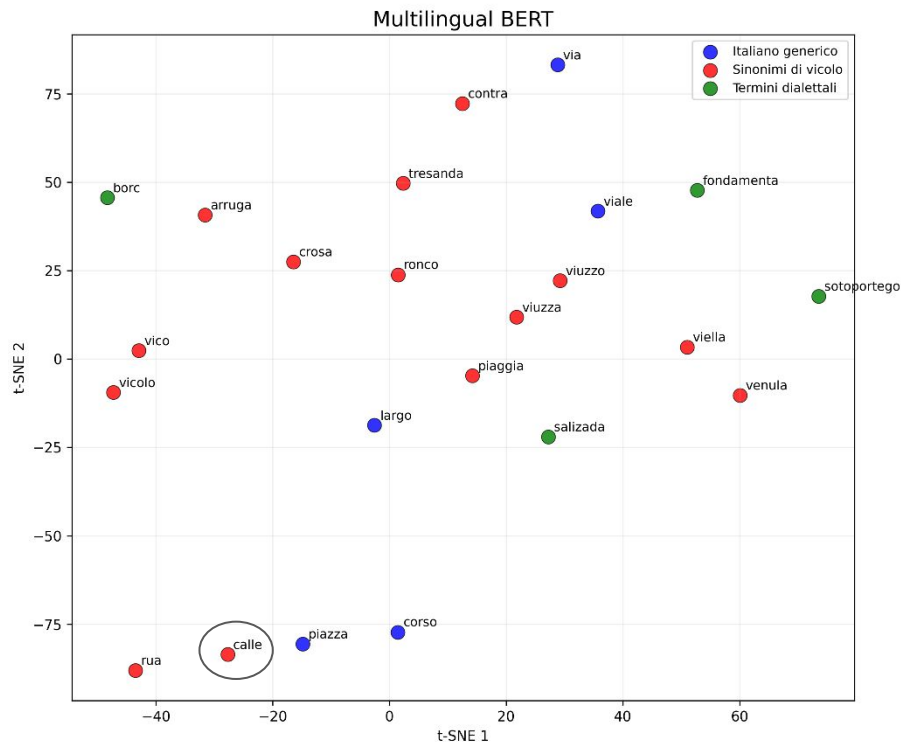
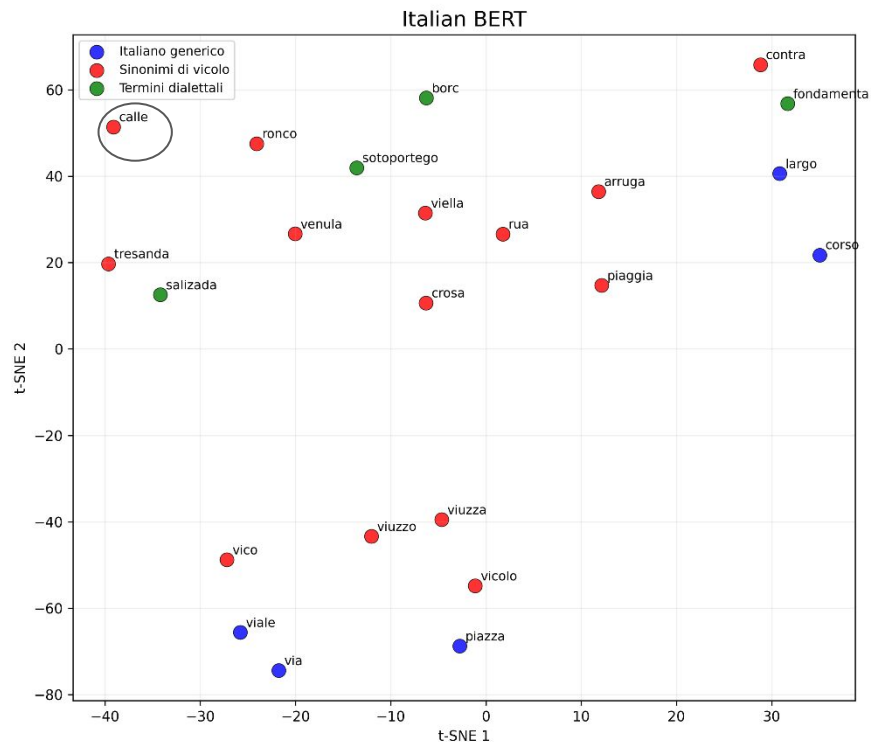
Valuti la parola cerchiata di rosso. ← Please evaluate the word circled in red

Item	1 (lingua o dialetto locale)	2	3	4	5 (italiano standard)
contra'	☐	☐	☐	☐	☐

← target generic term



Cosa succede nei modelli monolingua e multilingua?



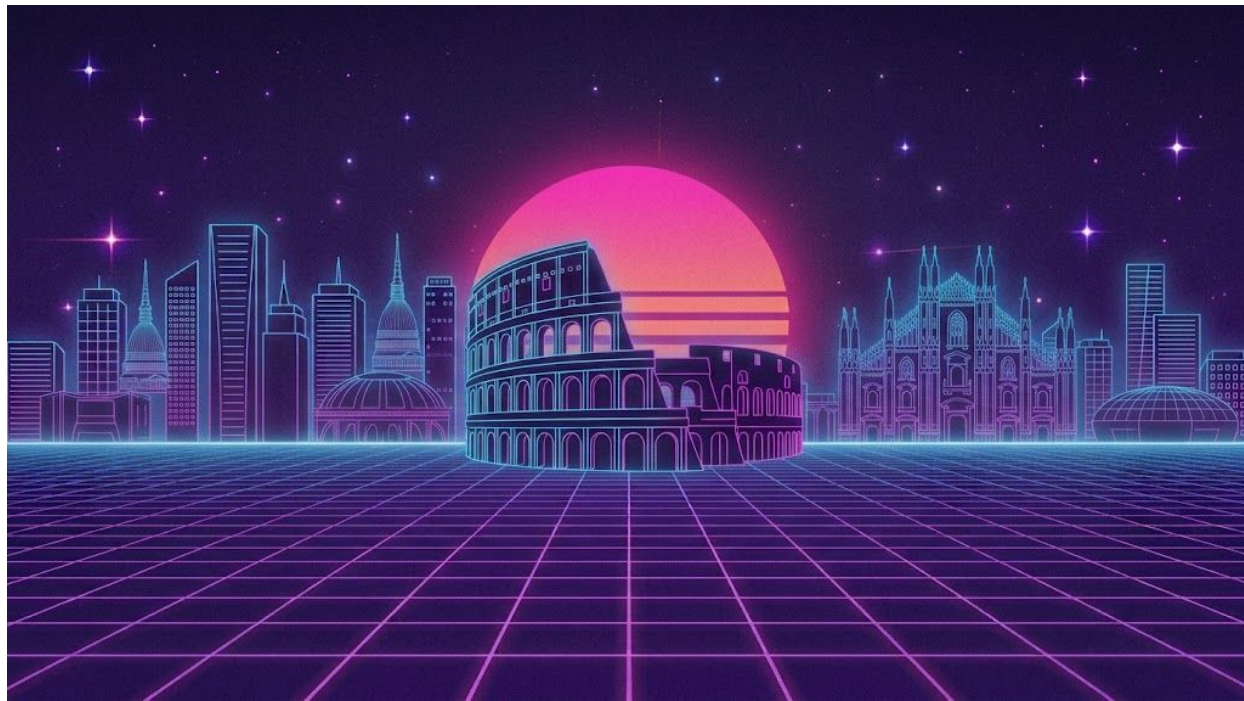
Conclusioni: quale italiano nelle IA?

Risposta breve: dipende dal modello, dai dati di addestramenti e da mille altri fattori.

L'italiano di questi modelli è comunque una lingua sintetica e sintetici, quindi non esistenti in natura, possono essere anche i contenuti.

Conoscere come funzionano questi modelli linguistici può aiutare ad usarli.

Grazie mille!



Una chiacchierata con la professoressa Felici

- Bias di genere nei modelli linguistici
- Modelli generativi e la traduzione
- La sicurezza e la condivisione dei dati

Sicurezza e protezione dei dati



Cg
COMITES
GINEVRA

SERATA INFORMATIVA

Privacy e protezione dati

Conoscere per proteggerci
nell'era digitale

CON IL DOTT. GIACOMO SALEMMA
ESPERTO IN PROTEZIONE DATI, FORMATORE E CONSULENTE

Giovedì 29 ottobre | Ore 18:30
Sede Comites - Quai Ernest-Ansermet 38, Ginevra
Ingresso libero, aperitivo offerto

comites-ginevra.ch/protezione-dati-2026