Istituto di Linguistica Computazionale
"Antonio Zampolli"
Consiglio Nazionale delle Ricerche

Linguistic Profiling in NLP: From LLM Evaluation to Digital Social Reading

LM4DH @ RANLP 2025, September 11 2025

Alessio Miaschi
ItaliaNLP Lab, Istituto di Linguistica Computazionale (CNR-ILC), Pisa
alessio.miaschi@ilc.cnr.it
https://alemiaschi.github.io/
http://www.italianlp.it/alessio-miaschi/

About me and...

I am a full-time researcher (RTD) at the ItaliaNLP Lab, Institute for Computational Linguistics “A. Zampolli” (CNR-ILC, Pisa). In 2022, I received my PhD in Computer Science at the University of Pisa.

My research interests lie primarily in the context of Natural Language Processing (NLP) and in the study of Language Models (LM). I am particularly interested in the interpretability of large-scale LMs and in the evaluation of their internal representations, with a specific emphasis on understanding their inner linguistic abilities.

CLARIN Trainers' Network

About me and… the team!

I am a full-time researcher (RTD) at the ItaliaNLP Lab, Institute for Computational Linguistics "A. Zampolli" (CNR-ILC, Pisa). In 2022, I received my PhD in Computer Science at the University of Pisa.

My research interests lie primarily in the context of Natural Language Processing (NLP) and in the study of Language Models (LM). I am particularly interested in the interpretability of large-scale LMs and in the evaluation of their internal representations, with a specific emphasis on understanding their inner linguistic abilities.

CLARIN Trainers' Network

The **ItaliaNLP Lab** (**CNR-ILC**) gathers researchers, postdocs and students from computational linguistics, computer science and linguistics who work on developing resources and algorithms for processing and understanding human languages.

Permanent Researchers:
- Felice Dell'Orletta
- Simonetta Montemagni
- Dominique Brunato
- Franco Alberto Cardillo
- Giulia Venturi
- Giulia Benotto

Temporary Researchers:
- Chiara Alzetta
- Alessio Miaschi

Research Fellows:
- Agnese Bonfigli
- Chiara Fazzone
- Ruben Piperno

PhD Students:
- Cristiano Ciaccio
- Luca Dini
- Lucia Domenichelli
- Michele Papucci
- Marta Sartor

+ Master/Undergraduate/Visiting Students

Link to website: http://www.italianlp.it/

Outline

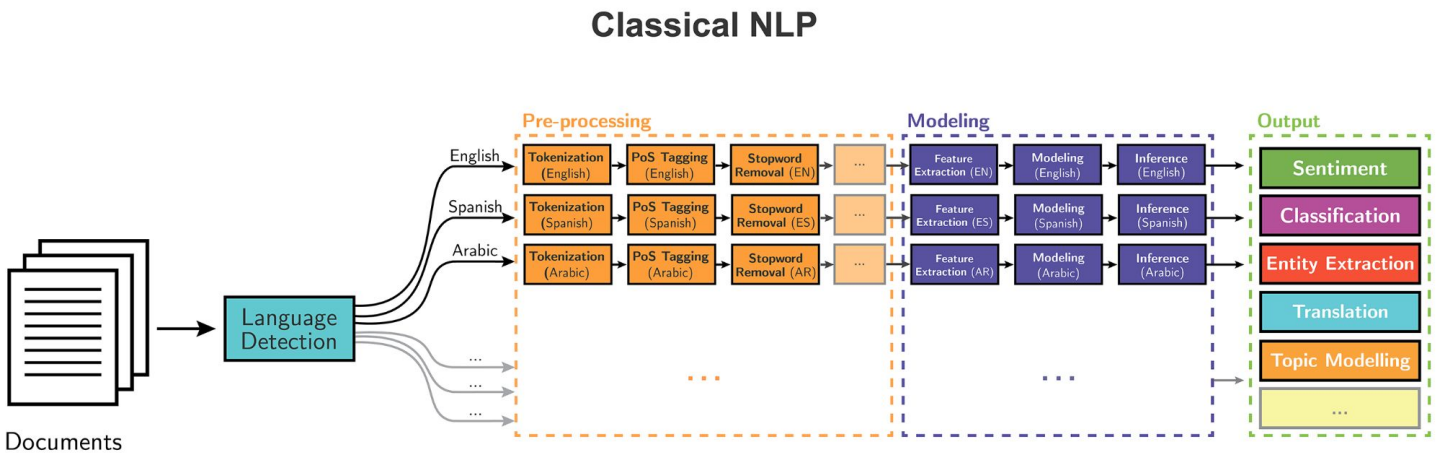
1. Introduction
2. Interpreting and Evaluating NLMs
3. NLP for Digital Social Reading

Introduction

- The field of NLP has seen an unprecedented progress in the last years
- Much of this progress is due to the replacement of traditional systems with newer and more powerful Deep Learning (DL) models

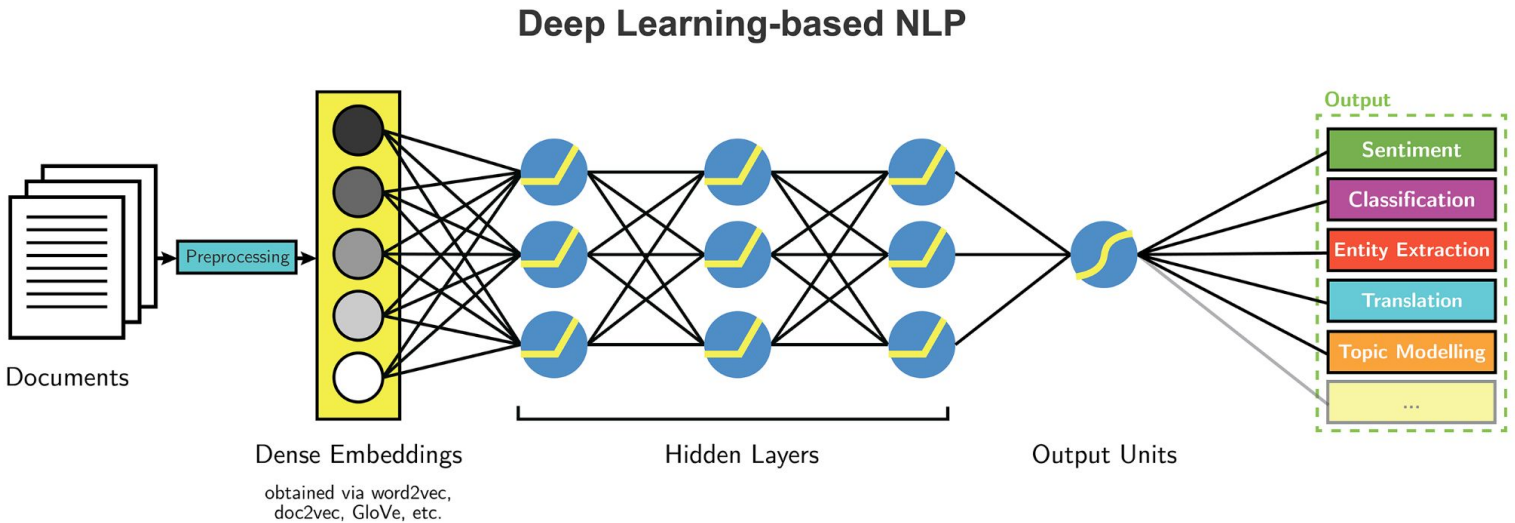
Introduction

- The field of NLP has seen an unprecedented progress in the last years
- Much of this progress is due to the replacement of traditional systems with newer and more powerful Deep Learning (DL) models



Introduction

- The field of NLP has seen an unprecedented progress in the last years
- Much of this progress is due to the replacement of traditional systems with newer and more powerful Deep Learning (DL) models



Neural Language Model (NLM)

- Neural Language Model → Neural Network trained to approximate the **language modeling** function

Neural Language Model (NLM)

- Neural Language Model → Neural Network trained to approximate the **language modeling** function

- **Language Modeling** → probability of a sentence s = *[w_1, w_2, …, w_n]* as:

$$P(s) = \prod_{i=1}^{N} P(w_i | w_1, w_2, ..., w_{i-1})$$

Neural Language Model (NLM)

- Neural Language Model → Neural Network trained to approximate the **language modeling** function

- **Language Modeling** → probability of a sentence s = *[w_1, w_2, …, w_n]* as:

$$P(s) = \prod_{i=1}^{N} P(w_i|w_1, w_2, ..., w_{i-1})$$

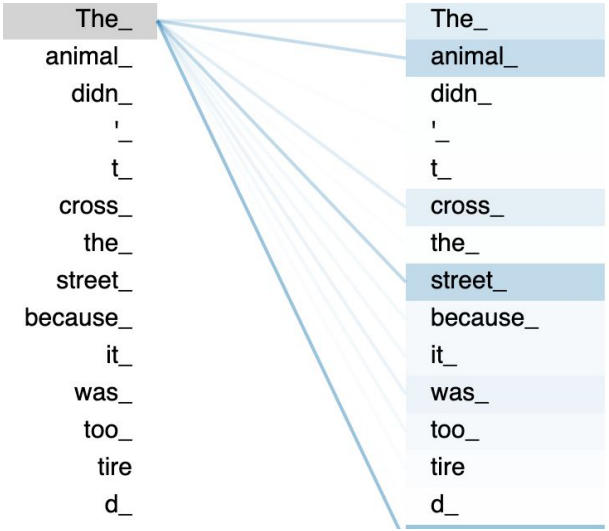
- Bengio et al. (2003) proposed a model to learn this function relying on the architecture of a neural network → **Neural Probabilistic Language Model**

Transformer Models

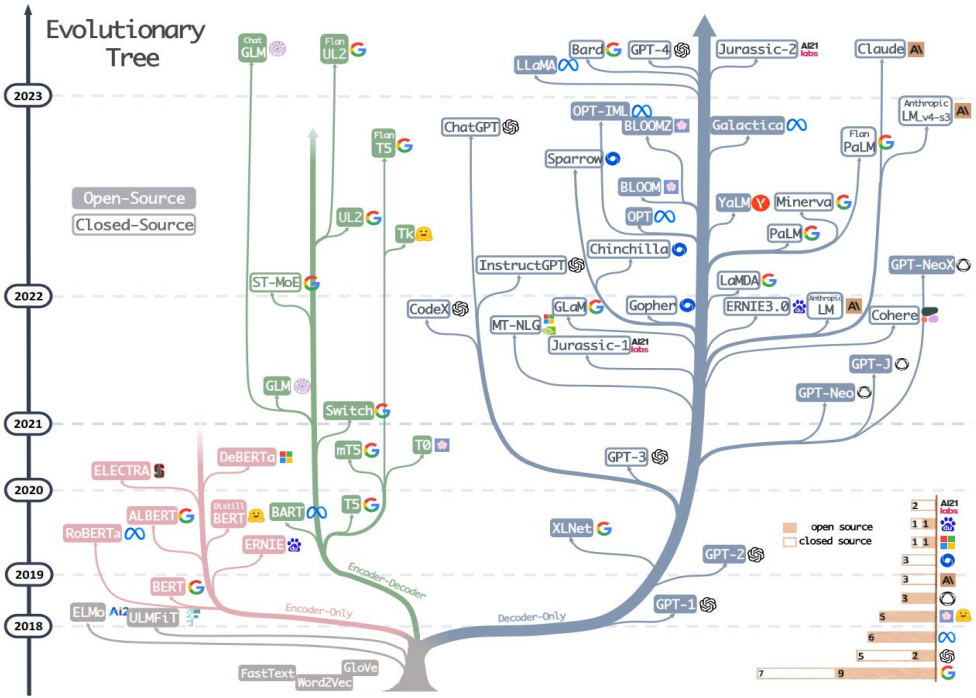
- Nowadays, the Transformer is the most commonly used architecture for the development of NLMs

- The Transformer (Vaswani et al., 2017) exploits the **attention mechanism** to create contextual representations of words and learn the relations among them

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$



“Evolutionary Tree”

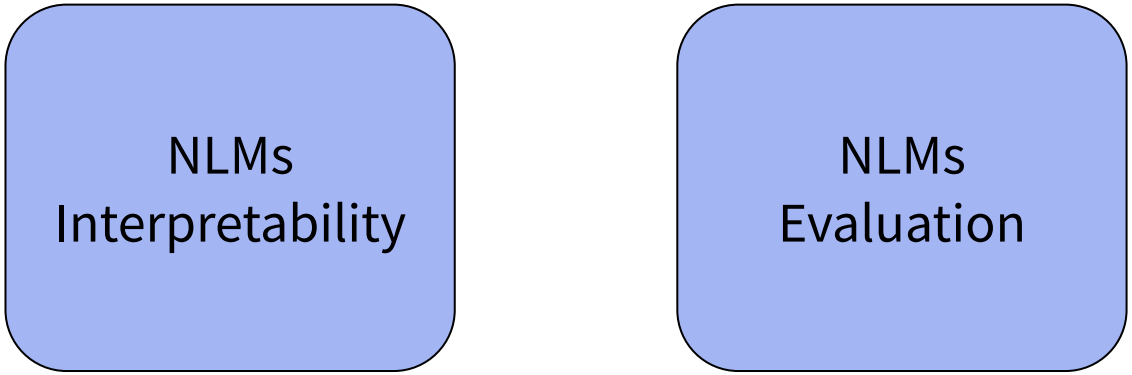


Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond (Yang et al., 2024), https://dl.acm.org/doi/10.1145/3649506

Interpreting and Evaluating NLMs

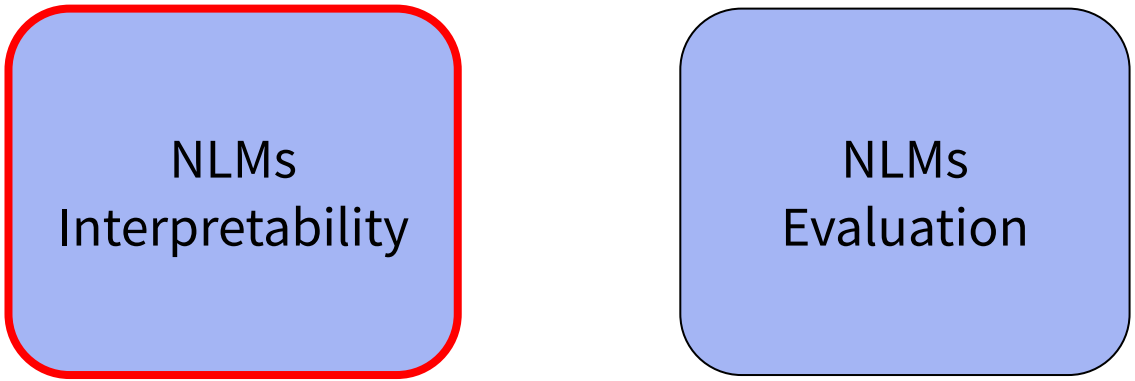
Interpreting and Evaluating NLMs

- The rapid development and widespread adoption of state-of-the-art Neural Language Models (NLMs) have increased the need for studies focused on their **interpretability** and the **evaluation** of their abilities



Interpreting and Evaluating NLMs

- The rapid development and widespread adoption of state-of-the-art Neural Language Models (NLMs) have increased the need for studies focused on their **interpretability** and the **evaluation** of their abilities



Interpretability in NLP

“In the context of NLP, this question needs to be understood in light of earlier NLP work. [...] In some of these systems, features are more easily understood by humans. [...] In contrast, it is more difficult to understand what happens in an end-to-end neural network model that takes input (say, word embeddings) and generates an output.”

Belinkov and Glass, Analysis Methods in Neural Language Processing: A Survey (2019). In Transactions of ACL, Volume 7, pages 49-72.

Interpretability in NLP

"In the context of NLP, this question needs to be understood in light of earlier NLP work. [...] In some of these systems, features are more easily understood by humans. [...] In contrast, it is more difficult to understand what happens in an end-to-end neural network model that takes input (say, word embeddings) and generates an output."

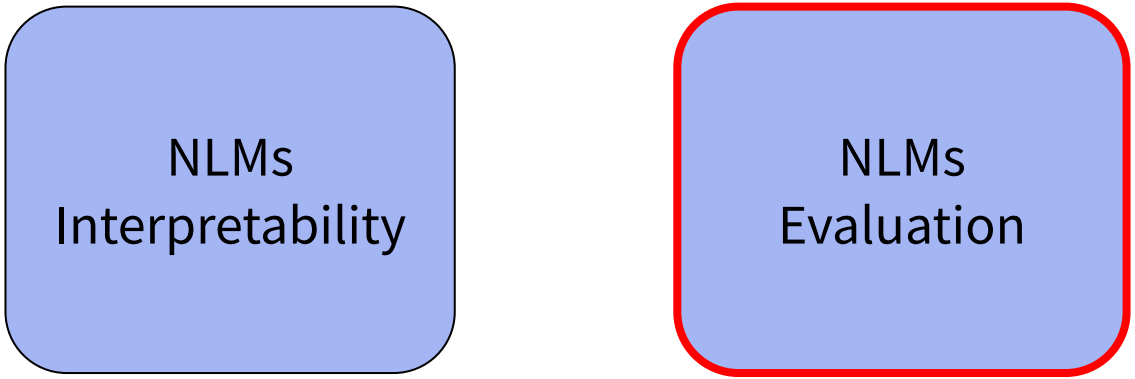
Belinkov and Glass, Analysis Methods in Neural Language Processing: A Survey (2019). In Transactions of ACL, Volume 7, pages 49-72.

Research questions:

- What happens in an end-to-end neural network model when trained on a language modeling task?
- What kind of linguistic knowledge (i.e. features) is encoded within their representations?
- Is there a relationship between the linguistic knowledge implicitly encoded and the ability to solve a specific task?

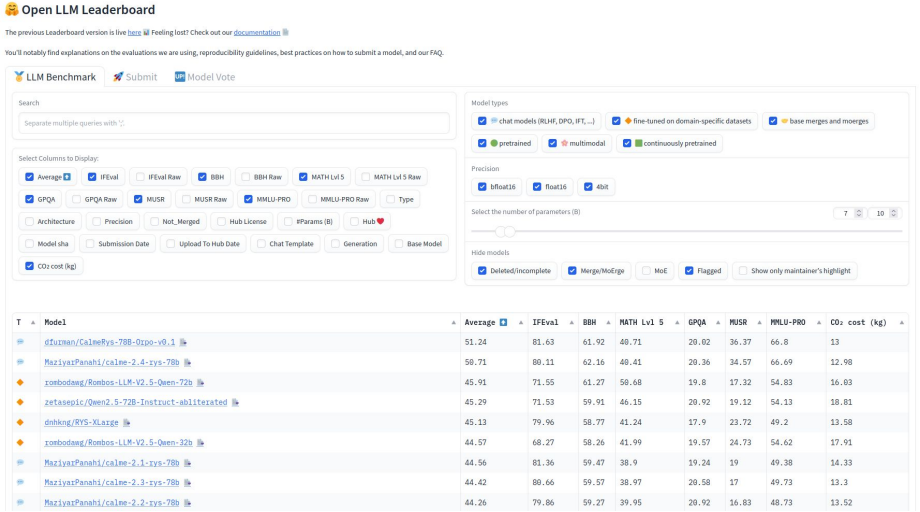
Interpreting and Evaluating NLMs

- The rapid development and widespread adoption of state-of-the-art Neural Language Models (NLMs) have increased the need for studies focused on their **interpretability** and the **evaluation** of their abilities



Evaluation of Neural Language Models

- The evaluation of NLMs has seen significant advancements in the past few years, with the development of dedicated benchmarks and evaluation frameworks
- These benchmarks are designed to assess models' performance on specific tasks and reasoning abilities:
  - OpenLLM Leaderboard
  - BigBench (Srivastava et al., 2023)
  - Holmes (Waldis et al., 2024)



Link: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

Competence vs. Performance in NLMs

- Within the broader context of interpretability and evaluation, one line of research focuses on studying and assessing the linguistic abilities of (Large) Language Models

- Such studies aim to uncover the implicit linguistic competencence encoded within these models and evaluate their generalization abilities

- **Competence vs. Performance**: investigation of the linguistic abilities of NLMs from a competence/performance perspective:
  - Distinction between the information encoded in a model internal representation vs. the model's behavioral responses to prompt during generation (Hu and Levy, 2023)

Profiling Neural Language Models

- The "*linguistic profiling*" methodology (van Halteren, 2004) assumes that wide counts of linguistic features are particularly helpful in the resolution of several NLP tasks, e.g.:
  - Text Profiling (e.g. text readability, textual genres)
  - Author Profiling (e.g. author's age and native language)

Miaschi A., Brunato D., Dell'Orletta F., Venturi G. (2020). Linguistic Profiling of a Neural Language Models. In *Proceedings of the 28th International Conference on Computational Linguistics* (COLING 2020, Barcelona) **[Outstanding paper for COLING 2020]**

Profiling Neural Language Models

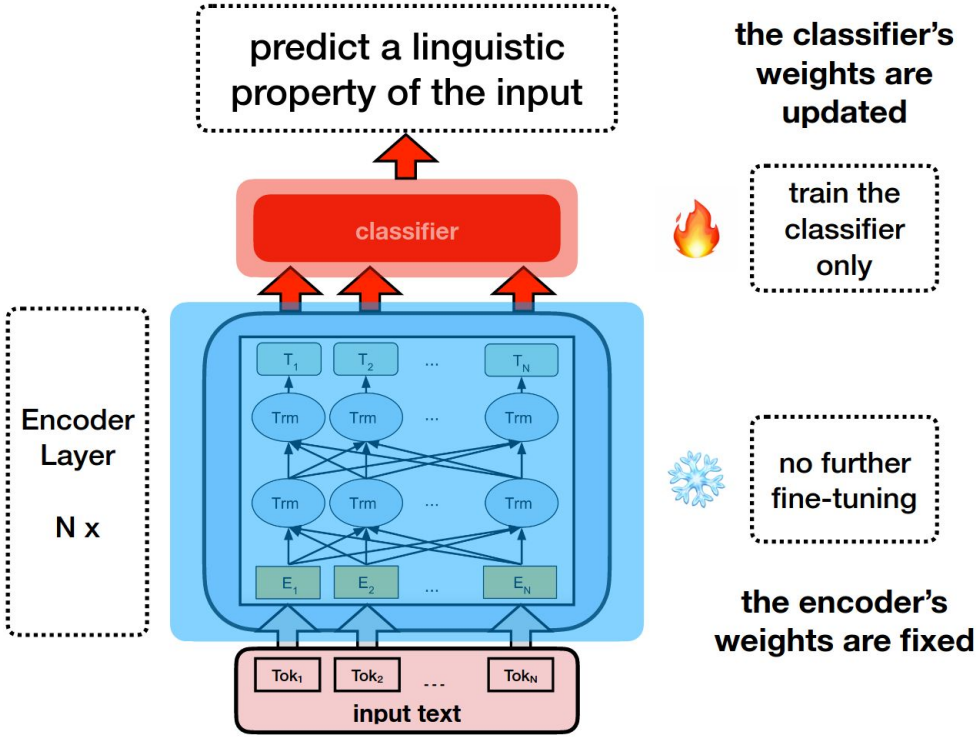
- The "*linguistic profiling*" methodology (van Halteren, 2004) assumes that wide counts of linguistic features are particularly helpful in the resolution of several NLP tasks, e.g.:
  - Text Profiling (e.g. text readability, textual genres)
  - Author Profiling (e.g. author's age and native language)

Research Question:

Could the informative power of these features also be helpful to understand the behaviour of state-of-the-art NLMs?

Miaschi A., Brunato D., Dell'Orletta F., Venturi G. (2020). Linguistic Profiling of a Neural Language Models. In *Proceedings of the 28th International Conference on Computational Linguistics* (COLING 2020, Barcelona) **[Outstanding paper for COLING 2020]**

Probing Task Approach



Slide from: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/19-probes.pdf

Profiling-UD: a tool for Linguistic Profiling of Texts

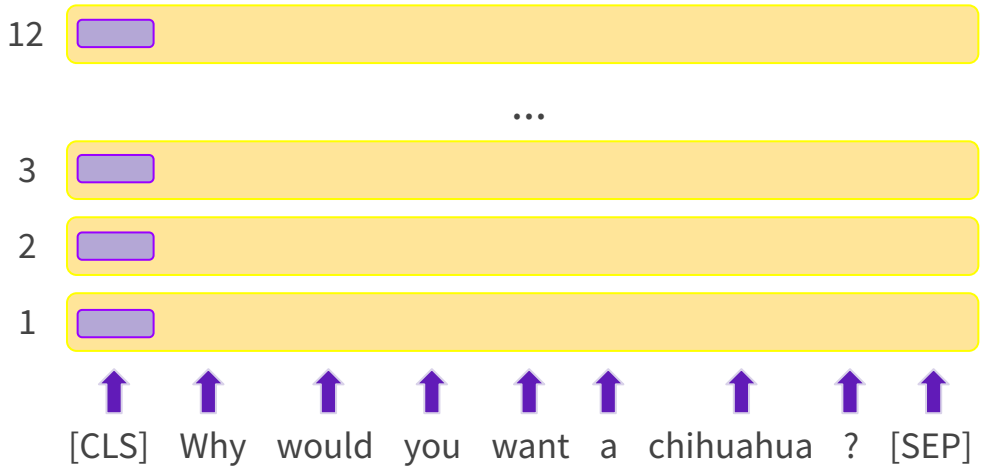
- ProfilingUD (Brunato et al., 2020) is a web–based application that performs linguistic profiling of a text, or a large collection of texts, for multiple languages

- It allows the extraction of more than 130 features, spanning across different levels of linguistic description

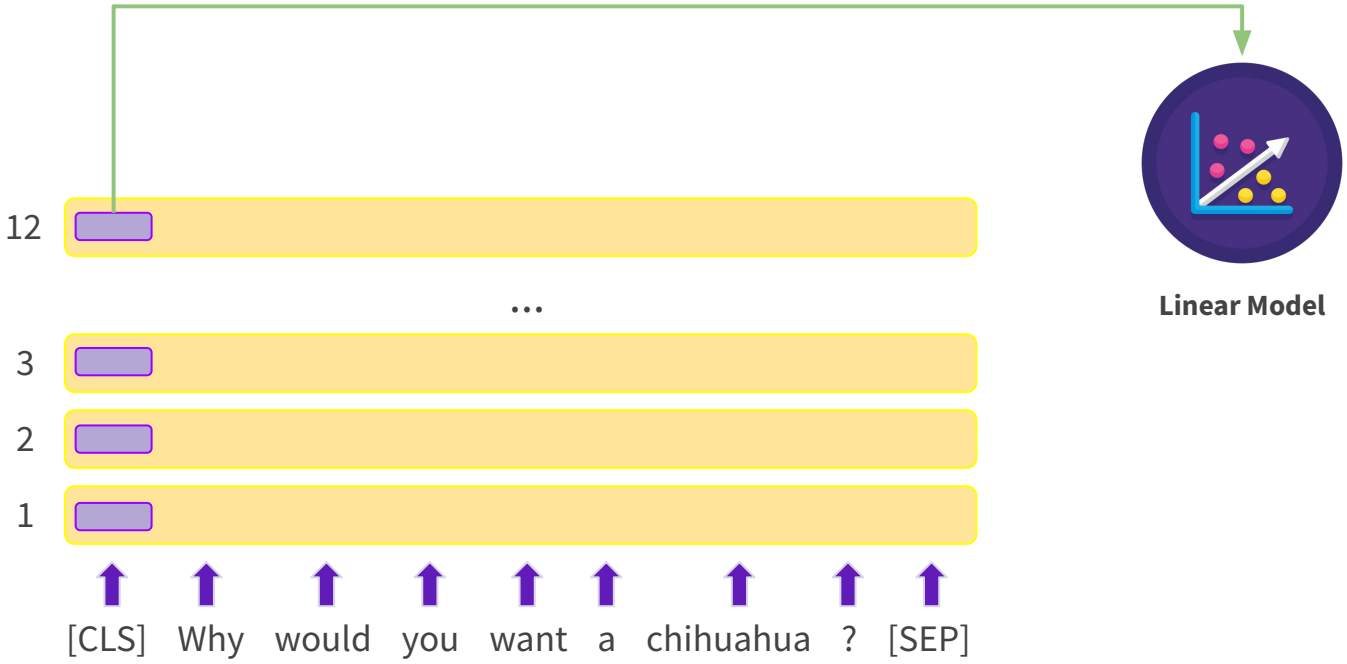
- Link: http://linguistic-profiling.italianlp.it/

Linguistic Feature
Raw Text Properties
Sentence Length
Word Length
Vocabulary Richness
Type/Token Ratio for words and lemmas
Morphosyntactic information
Distibution of UD and language–specific POS
Lexical density
Inflectional morphology
Inflectional morphology of lexical verbs and auxiliaries
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
Verb arity and distribution of verbs by arity
Global and Local Parsed Tree Structures
Depth of the whole syntactic tree
Average length of dependency links and of the longest link
Average length of prepositional chains and distribution by depth
Clause length
Relative order of elements
Order of subject and object
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chains and distribution by depth
Relative order of subordinate clauses

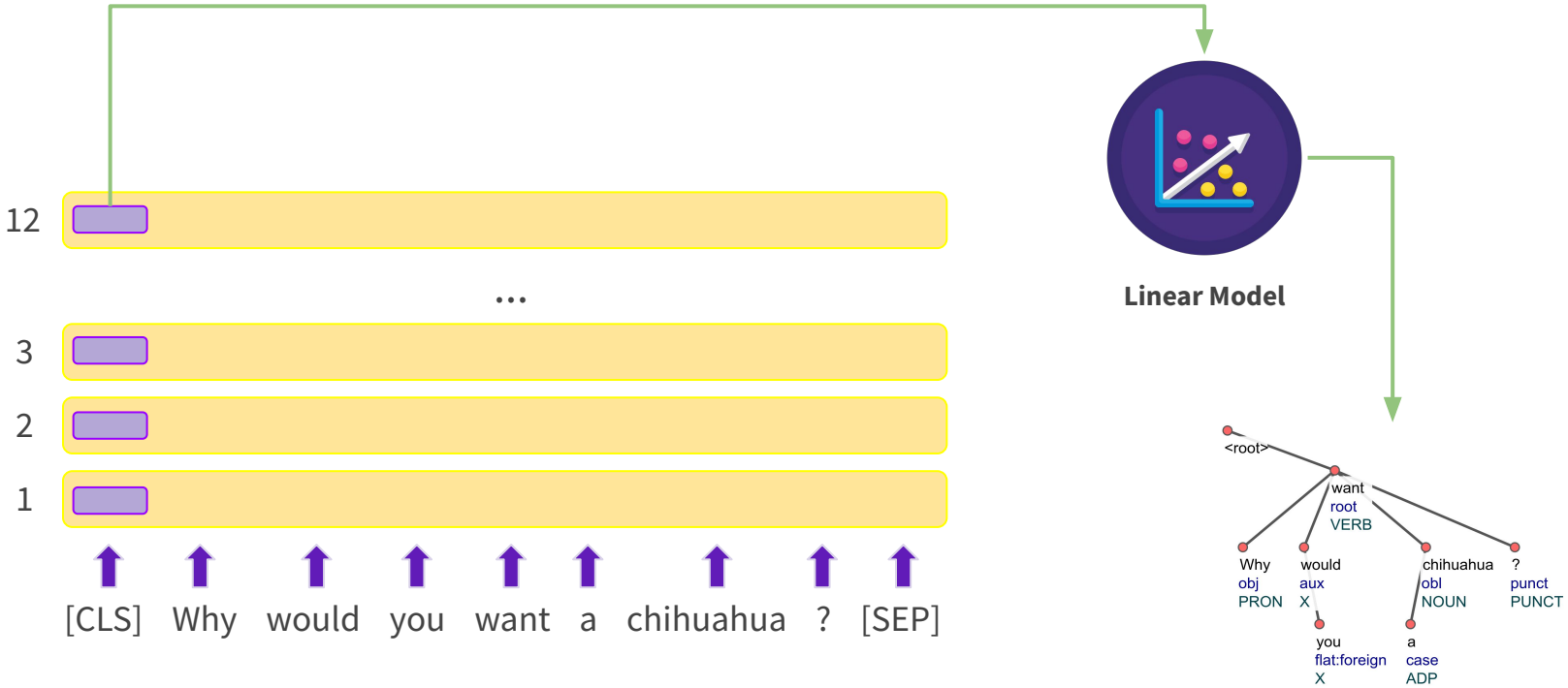
Profiling Neural Language Models



Profiling Neural Language Models



Profiling Neural Language Models



Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

- We investigated the linguistic knowledge implicitly encoded by BERT

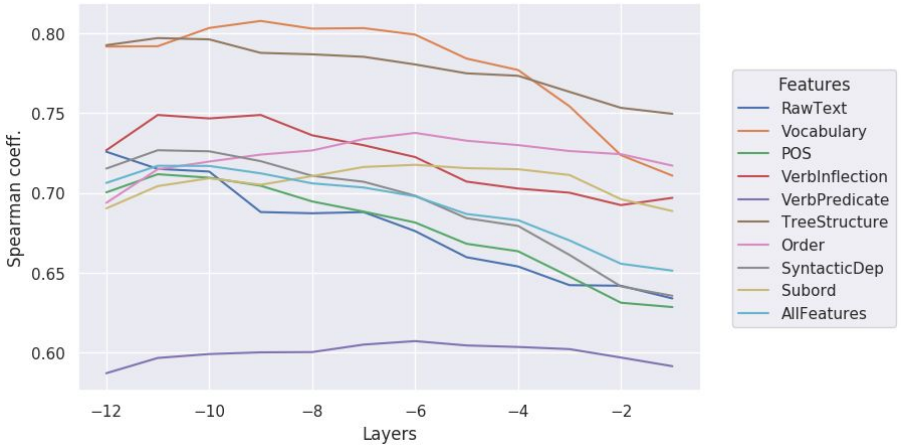
Research questions:

1. What kind of linguistic properties are encoded in a pre-trained version of BERT?

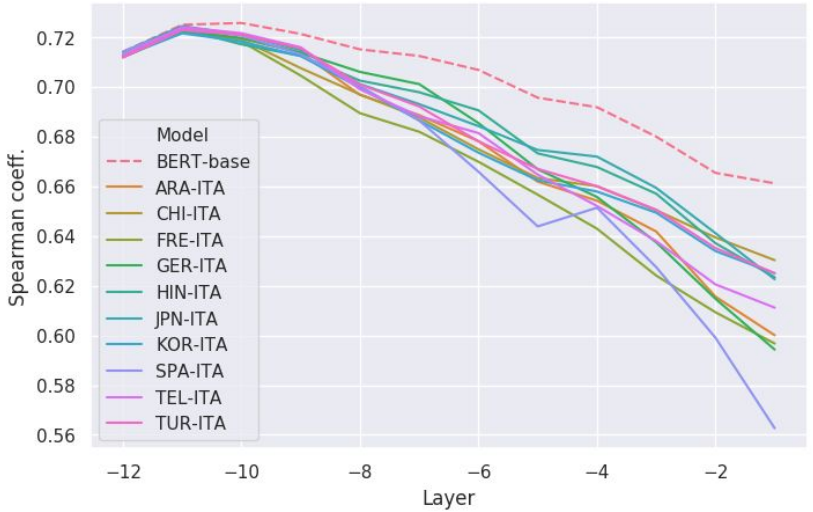
2. How this knowledge is modified after a fine-tuning process?
    a. Fine-tuning on the Natural Language Identification Task

Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

Pre fine-tuning:



Post fine-tuning:



Evaluating Large Language Models via Linguistic Profiling

- Motivations:
  - Large Language Models (LLMs) demonstrated remarkable capabilities in solving multiple tasks and in generating coherent and contextually relevant texts
  - Such capabilities have been extensively evaluated against several benchmarks, as evidenced by the success of platforms such as the OpenLLM Leaderboard
  - A comprehensive evaluation of **LLMs' linguistic abilities in generation**, independent of specific tasks and possibly cross-cutting across them, is still missing

Miaschi A., Dell'Orletta F., Venturi G. (2024). Evaluating Large Language Models via Linguistic Profiling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2024, Miami, Florida)

Evaluating Large Language Models via Linguistic Profiling

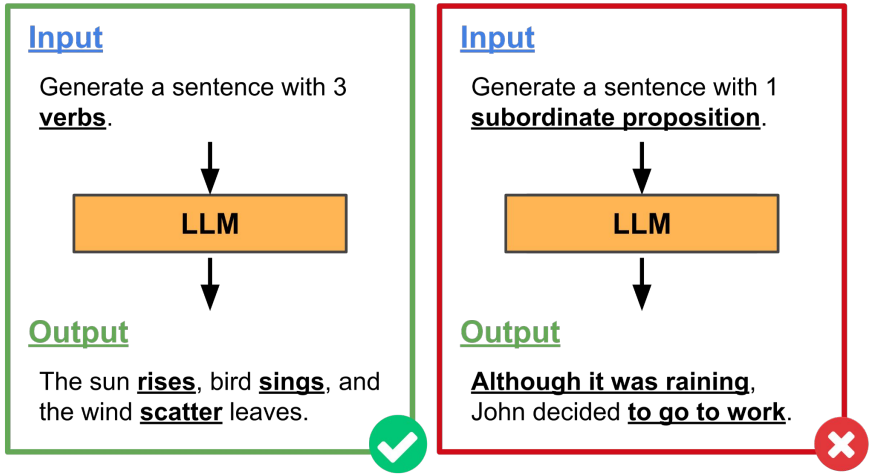
- Motivations:
  - Large Language Models (LLMs) demonstrated remarkable capabilities in solving multiple tasks and in generating coherent and contextually relevant texts
  - Such capabilities have been extensively evaluated against several benchmarks, as evidenced by the success of platforms such as the OpenLLM Leaderboard
  - A comprehensive evaluation of **LLMs' linguistic abilities in generation**, independent of specific tasks and possibly cross-cutting across them, is still missing

How effectively can LLMs generate sentences that adhere to targeted linguistic constraints representing various morpho-syntactic and syntactic phenomena?

Miaschi A., Dell'Orletta F., Venturi G. (2024). Evaluating Large Language Models via Linguistic Profiling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2024, Miami, Florida)

Our Approach

- We evaluate the ability of several LLMs to generate sentences with targeted (morpho-)syntactic linguistic constraints
- We prompted the models to generate sentences containing these constraints within a fixed prompt structure:
  - For each property/constraint, we asked the models to generate a fixed number of sentences having a precise value of that property
- Given the well-known difficulty of LLMs in producing texts with precise numerical constraints, we decided to constrain the models on increasing values of linguistic properties



Linguistic Properties and Values Selection

- We relied on a set of linguistic properties as constraints encompassing diverse morpho-syntactic and syntactic phenomena of a sentence
- We relied on the largest English Universal Dependency (UD) treebank, i.e. English Universal Dependency (EWT) (Silveira et al., 2014)
    - Extraction of the linguistic properties with the Profiling-UD tool (Brunato et al., 2020)
    - In the few-shot configuration, we used 5 exemplar sentences extracted from EWT
- We asked each model to generate a fixed number of sentences following a set of increasing values for each linguistic property
    - We generate 50 sentences for every value within the set of five values, thus obtaining a total of 250 sentences per property.

Models and Evaluation

Models:

Model	Parameters
Gemma	2B
Gemma	7B
LLaMA-2	7B
LLaMA-2	14B
Mistral	7B

Evaluation:

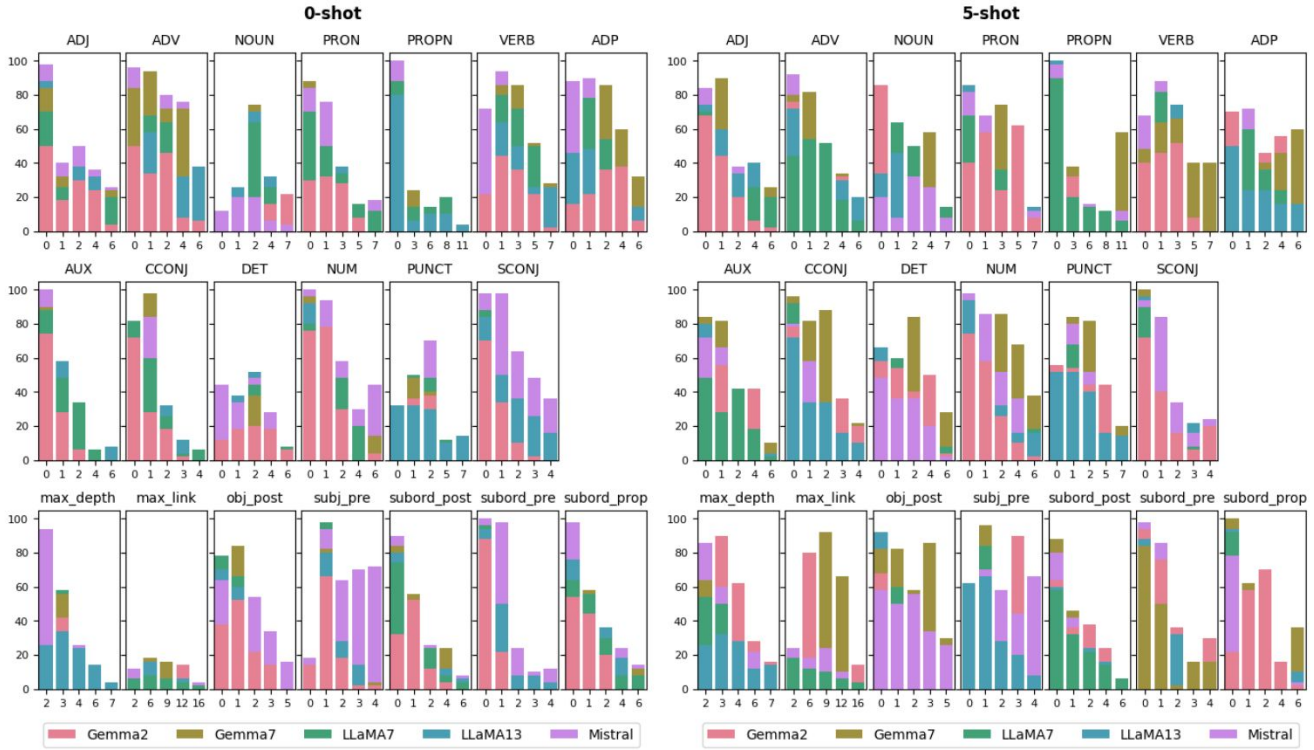
- We used two different metrics:
  - **Success Rate (SR)**: fraction of times the model generated a sentence whose property value exactly corresponds to the one provided.
  - **Spearman coefficient**: correlation coefficients between the increasing property values extracted from EWT and those extracted from the sentences generated by the models.

Success Rate Results

Ling. properties	Gemma2	Gemma7	LLaMA7	LLaMA13	Mistral
			Success Rate		
Morphosyntax			**0-shot**		
ADJ	25.2	36.8	33.6	42	50
ADV	28.8	70.8	34.4	38.8	74
NOUN	8.8	26	23.2	29.6	12.4
PRON	19.6	22.8	36.4	34	41.6
PROPN	25.6	29.2	28	22	22
VERB	25.2	50.8	46.8	37.2	57.6
ADP	23.6	54.4	31.2	31.6	64.4
AUX	21.6	23.6	35.2	37.2	29.2
CCONJ	24	33.2	35.6	35.2	33.2
DET	14.8	15.6	14.8	25.6	32
NUM	37.6	48	43.2	40.8	65.2
PUNCT	14.8	19.2	26	23.6	29.2
SCONJ	23.2	27.6	27.6	42.4	68.8
Avg	22.52	35.23	32	33.85	44.58
Syntax			**0-shot**		
max_depth	13.6	17.6	16.4	20.4	29.2
max_link	9.2	7.2	5.2	6.8	3.6
obj_post	25.2	36.4	35.2	36.4	40.8
subj_pre	20.4	21.2	22.8	26.4	63.6
subord_post	20	36.8	29.2	29.6	32.8
subord_pre	22	23.2	24	32.8	48.8
subord_prop	23.6	37.6	33.2	37.2	41.6
Avg	19.14	25.71	23.71	27.09	37.2

Ling. properties	Gemma2	Gemma7	LLaMA7	LLaMA13	Mistral
			Success Rate		
Morphosyntax			**5-shot**		
ADJ	28	47.6	34.4	42.8	45.6
ADV	33.2	47.2	34.8	41.2	51.6
NOUN	43.6	20.4	34.4	28.4	18.8
PRON	38.4	45.6	34	39.2	39.6
PROPN	30.4	40.4	28.4	29.6	29.2
VERB	29.2	51.6	38.4	37.6	52
ADP	44.8	47.2	28.8	26	42
AUX	31.6	45.6	27.6	38.4	35.6
CCONJ	38	63.6	34	33.2	34.4
DET	41.2	37.6	31.6	30	28.4
NUM	34	71.6	44.8	43.2	57.6
PUNCT	42	40	34	34.8	31.6
SCONJ	30.8	43.2	31.2	40.8	50.4
Avg	35.78	46.28	33.57	35.78	39.75
Syntax			**5-shot**		
max_depth	52	24.4	30.4	22.4	38.8
max_link	22.8	47.2	10	10.8	15.6
obj_post	31.6	67.6	32	43.6	44.8
subj_pre	51.2	42.4	41.6	36.8	50
subord_post	33.2	34	26.4	27.6	34
subord_pre	47.6	33.6	34	31.6	45.6
subord_prop	33.6	50.4	34.8	32.8	34
Avg	38.86	42.8	29.89	29.37	37.54

How do LLMs Follow Constraints Across Values?



Spearman Results

Ling. properties	Gemma2	Gemma7	LLaMA7	LLaMA13	Mistral
			Spearman		
Morphosyntax			0-shot		
ADJ	0.59	0.73	0.74	0.79	0.92
ADV	##	0.88	0.52	0.65	0.95
NOUN	0.63	0.72	0.62	0.66	0.93
PRON	0.26	0.35	0.58	0.80	0.91
PROPN	##	0.66	0.60	0.67	0.88
VERB	0.56	0.83	0.78	0.71	0.76
ADP	0.55	0.89	0.48	0.64	0.96
AUX	##	0.29	0.32	0.56	0.96
CCONJ	0.27	0.33	0.35	0.33	0.42
DET	0.28	0.36	##	0.28	0.79
NUM	0.49	0.74	0.60	0.62	0.94
PUNCT	0.24	0.54	0.63	0.61	0.78
SCONJ	##	0.44	0.40	0.62	0.92
Avg	0.30	0.60	0.51	0.61	0.86
Syntax			0-shot		
max_depth	##	0.18	##	##	0.76
max_link	##	0.44	0.57	0.43	0.75
obj_post	0.21	0.47	0.37	0.38	0.59
subj_pre	##	##	0.37	0.13	0.84
subord_post	0.13	0.65	0.44	0.58	0.59
subord_pre	##	0.33	0.13	0.34	0.72
subord_prop	0.28	0.60	0.45	0.67	0.83
Avg	0.08	0.38	0.33	0.36	0.73

Ling. properties	Gemma2	Gemma7	LLaMA7	LLaMA13	Mistral
			Spearman		
Morphosyntax			5-shot		
ADJ	0.19	0.78	0.76	0.79	0.86
ADV	0.43	0.62	0.52	0.71	0.80
NOUN	0.87	0.76	0.77	0.75	0.90
PRON	0.63	0.65	0.78	0.85	0.81
PROPN	0.25	0.87	0.76	0.81	0.81
VERB	0.42	0.77	0.77	0.72	0.87
ADP	0.46	0.81	0.53	0.61	0.77
AUX	0.37	0.70	0.53	0.59	0.60
CCONJ	0.53	0.56	0.52	0.52	0.60
DET	0.49	0.77	0.65	0.65	0.65
NUM	##	0.63	0.72	0.74	0.77
PUNCT	0.60	0.70	0.73	0.79	0.69
SCONJ	0.26	0.66	0.62	0.71	0.74
Avg	0.42	0.71	0.67	0.71	0.76
Syntax			5-shot		
max_depth	0.80	0.56	0.39	0.40	0.78
max_link	0.40	0.86	0.64	0.52	0.70
obj_post	0.42	0.84	0.51	0.62	0.72
subj_pre	0.59	0.52	0.55	0.47	0.74
subord_post	0.58	0.59	0.53	0.54	0.77
subord_pre	0.12	0.24	0.33	0.35	0.56
subord_prop	0.39	0.79	0.68	0.66	0.74
Avg	0.47	0.63	0.52	0.51	0.71

Selected Findings

- Models tend to adhere slightly more accurately to **morphosyntactic constraints** rather then syntactic ones

- Models are capable of distinguishing when they are asked to generate a sentence **with or without a given feature**

- Constraining generation for a specific linguistic element does not always primarily enhance that element, suggesting that the **models are not simply creating longer sentences, but rather sentences with a varied (morpho)syntactic structure**

NLP for Digital Social Reading

Digital Social Reading

- **Digital Social Reading (DSR)**: a wide variety of practices related to the activity of reading and using digital technologies and platforms to share thoughts and impressions about books with others (Pianzola F., 2025)
- The popularity of these platforms has led to the creation of new social valences of reading (Namakura, 2013) and, most importantly, of massive corpora of user-generated book reviews (e.g. Koshua et al., 2017; Sabri and Weber, 2021)

Digital Social Reading

- **Digital Social Reading (DSR)**: a wide variety of practices related to the activity of reading and using digital technologies and platforms to share thoughts and impressions about books with others (Pianzola F., 2025)
- The popularity of these platforms has led to the creation of new social valences of reading (Namakura, 2013) and, most importantly, of massive corpora of user-generated book reviews (e.g. Koshua et al., 2017; Sabri and Weber, 2021)
- Still little is known about the diverse communication strategies adopted by readers to share their reading experiences with others in terms of stylistic variations between reviews written across different platforms or referring to books belonging to different genres

Tell me how you write and I'll tell you what you read

- In this work we studied the linguistic properties and lexicon of Italian book reviews published on two leading platforms for DSR, i.e. Amazon Books and Goodreads
- For the purpose of our work we introduced a novel corpus called *A Good Review* which covers reviews of 300 books belonging to six literary fiction genres and reviewed by users of Amazon and Goodreads

Alzetta C., Dell'Orletta F., Miaschi A., Prat E., Venturi G. (2023). Tell me how you write and I'll tell you what you read: a study on the writing style of book reviews. In *Journal of Documentation*, Volume 80

Tell me how you write and I'll tell you what you read

- In this work we studied the linguistic properties and lexicon of Italian book reviews published on two leading platforms for DSR, i.e. Amazon Books and Goodreads
- For the purpose of our work we introduced a novel corpus called *A Good Review* which covers reviews of 300 books belonging to six literary fiction genres and reviewed by users of Amazon and Goodreads

Our Approach:

- We automatically acquired a set of stylistic properties from the reviews and we analysed the variation of these features across the review's venue and the genre of the reviewed book
- We conducted a series of classification experiments using multiple approaches and feature configurations to predict:
    - If a review was posted on either Amazon or Goodreads;
    - The genre of the book being reviewed based on its review.

Alzetta C., Dell'Orletta F., Miaschi A., Prat E., Venturi G. (2023). Tell me how you write and I'll tell you what you read: a study on the writing style of book reviews. In *Journal of Documentation*, Volume 80

A Good Review Corpus

- A Good Review (Amazon and GOODreads REVIEWs) is collection of book reviews acquired from Amazon and Goodreads across 6 literary fiction genres:
  - thriller, historical fiction, romance, science fiction, horror and fantasy

Genre	Amazon Reviews	Amazon Sentences	Amazon Tokens	Goodreads Reviews	Goodreads Sentences	Goodreads Tokens
Fantasy	8,608	31,229	567,472	8,316	75,402	1,701,735
Historical Fiction	4,455	16,050	296,361	5,196	43,486	1,037,555
Horror	3,958	16,677	329,801	6,219	51,815	1,205,617
Romance	6,885	28,970	527,996	7,855	76,992	1,707,373
Science Fiction	7,070	26,595	505,177	6,255	56,875	1,336,165
Thriller	5,952	21,699	383,449	4,847	34,484	765,947
Total	36,928	141,220	2,610,256	38,688	339,054	7,754,392

Table 1. Dataset statistics for each genre and in total for each platform

Source(s): Authors' own creation

Analysis of reviews' style

Group	Feature	Amazon mean (stdev)	Goodreads mean (stdev)	r
RawText	Tokens	70.69 (±128.56)	200.43 (±286.13)	(+)0.416
	sentences	3.82 (±5.24)	8.76 (±11.70)	(+)0.385
	sent_length	16.38 (±10.87)	20.80 (±11.22)	(+)0.270
Vocab	ttr_F (100)	0.10 (±0.26)	0.35 (±0.37)	(+)0.333
	ttr_L (100)	0.09 (±0.22)	0.30 (±0.33)	(+)0.332
	lexical_density	0.56 (±0.14)	0.51 (±0.09)	(−)0.235
	NBIV	0.87 (±0.14)	0.84 (±0.12)	(−)0.230
	ttr_F (200)	0.04 (±0.15)	0.18 (±0.29)	(+)0.218
	ttr_L (200)	0.03 (±0.13)	0.15 (±0.25)	(+)0.218
POS	PROPN	2.10 (±7.37)	2.72 (±4.52)	(+)0.283
	ADJ	12.56 (±16.28)	8.09 (±8.05)	(−)0.220
	NUM	0.55 (±1.85)	0.99 (±3.37)	(+)0.217
Verb Inflection	aux_3stprs-plr	7.11 (±18.23)	10.70 (±17.56)	(+)0.215
	verbs_3stprs-plr	7.80 (±19.45)	11.48 (±18.20)	(+)0.213
Verb Predicate	verbal_heads	2.01 (±1.60)	2.51 (±1.54)	(+)0.237
	verb_edges_dist_5	4.51 (±10.94)	6.38 (±9.81)	(+)0.234
	avg_verb_edges	2.21 (±1.06)	2.53 (±0.86)	(+)0.214
	n_prep_chains	2.63 (±5.23)	7.32 (±11.09)	(+)0.366
	max_links_len	11.88 (±9.53)	18.97 (±13.66)	(+)0.351
Tree Structure	avg_links_len	2.25 (±0.67)	2.52 (±0.52)	(+)0.282
	avg_max_links_len	7.29 (±5.19)	8.99 (±4.92)	(+)0.252
	avg_prep_chain_len	0.78 (±0.59)	0.94 (±0.48)	(+)0.214
	avg_max_depth	3.67 (±1.85)	4.23 (±1.74)	(+)0.206
	root	12.06 (±17.81)	7.28 (±8.77)	(−)0.270
	det:poss	0.61 (±1.64)	0.84 (±1.41)	(+)0.234
Syntactic Dep	expl	0.78 (±1.58)	1.06 (±1.42)	(+)0.225
	iobj	0.59 (±1.48)	0.72 (±1.31)	(+)0.206
	ccomp	0.67 (±1.42)	0.84 (±1.32)	(+)0.201
	nummod	0.45 (±1.37)	0.74 (±2.45)	(+)0.200
Subord	avg_sub_chain_len	0.95 (±0.61)	1.11 (±0.52)	(+)0.229
	subord_dist_2	11.99 (±23.48)	15.54 (±19.86)	(+)0.212

Note(s): Features in each group are ordered by decreasing rank-biserial correlation value (r)

- Goodreads' users tend to write longer reviews characterized by a more complex and articulated writing style
- On Amazon, the readers' writing style is more homogeneous across genres compared to Goodreads
- Readers of the same genre tend to adopt different styles based on the platform

Classifying Reviews

Model	Accuracy	Precision	Recall	F-score
Majority class	0.31	0.39	0.61	0.48
Sent-length	0.53	0.51	0.61	0.56
Profiling	0.64	0.63	0.66	0.64
Ngrams	0.59	0.55	0.63	0.59
BERT	0.82	0.86	0.76	0.81
SVM (BERT)	*0.86*	*0.85*	*0.89*	*0.87*
SVM (BERT + Profiling)	*0.86*	*0.85*	*0.89*	*0.87*

Classifying Reviews

Genre	M	Random uniform A	Random uniform G	Sent length A	Sent length G	Profiling A	Profiling G	Ngrams A	Ngrams G	BERT A	BERT G	SVM (BERT) A	SVM (BERT) G	SVM (BERT + Profiling) A	SVM (BERT + Profiling) G
Hor	P	0.11	0.17	0.12	0	0.15	0.22	0.47	0.55	0.7	0.62	0.58	0.8	0.57	0.8
	R	0.18	0.18	0.06	0	0.15	0.17	0.45	0.51	0.5	0.59	0.58	0.77	0.58	0.76
	F	0.14	0.17	0.08	0	0.15	0.19	0.46	0.53	0.58	0.6	0.58	0.78	0.57	0.78
Hist-Fi	P	0.09	0.13	0	0	0.18	0.22	0.33	0.46	0.52	0.56	0.5	0.73	0.5	0.73
	R	0.16	0.15	0	0	0.19	0.26	0.46	0.45	0.37	0.52	0.5	0.77	0.5	0.76
	F	0.11	0.14	0	0	0.18	0.24	0.38	0.45	0.43	0.54	0.5	0.75	0.5	0.74
Sci-Fi	P	0.25	0.16	0.05	0.08	0.23	0.23	0.65	0.55	0.61	0.72	0.65	0.78	0.65	0.78
	R	0.17	0.16	0	0.01	0.18	0.19	0.54	0.51	0.53	0.63	0.57	0.77	0.57	0.78
	F	0.2	0.16	0.01	0.01	0.2	0.21	0.59	0.53	0.57	0.67	0.6	0.77	0.61	0.78
Thril	P	0.21	0.12	0	0.10	0.24	0.2	0.57	0.53	0.59	0.61	0.63	0.72	0.62	0.72
	R	0.18	0.17	0	0.17	0.24	0.24	0.49	0.53	0.54	0.58	0.56	0.76	0.56	0.76
	F	0.19	0.14	0	0.13	0.24	0.22	0.53	0.53	0.56	0.6	0.59	0.74	0.59	0.74
Rom	P	0.26	0.2	0.22	0.18	0.26	0.28	0.58	0.56	0.54	0.61	0.63	0.79	0.62	0.79
	R	0.17	0.17	0.03	0.22	0.27	0.34	0.5	0.58	0.51	0.68	0.6	0.79	0.59	0.77
	F	0.21	0.18	0.06	0.20	0.27	0.31	0.54	0.57	0.52	0.64	0.61	0.79	0.6	0.78
Fant	P	0.12	0.22	0.11	0.19	0.32	0.28	0.32	0.56	0.54	0.66	0.6	0.81	0.6	0.81
	R	0.16	0.17	0.83	0.47	0.36	0.24	0.53	0.59	0.79	0.68	0.73	0.78	0.71	0.79
	F	0.14	0.19	0.19	0.27	0.34	0.26	0.4	0.58	0.64	0.67	0.66	0.8	0.65	0.8
All	A	0.17	0.17	0.11	0.17	0.25	0.24	0.5	0.54	0.57	0.63	0.6	0.78	0.6	0.77
	P	0.17	0.17	0.08	0.09	0.23	0.24	0.49	0.53	0.58	0.63	0.6	0.77	0.59	0.77
	R	0.16	0.17	0.06	0.10	0.23	0.24	0.48	0.53	0.55	0.62	0.59	0.77	0.59	0.77
	F	0.16	0.17	0.06	0.10	0.23	0.24	0.48	0.53	0.55	0.62	0.59	0.77	0.59	0.77

Source(s): Authors' own creation

Revealing Author Identity through Reader Reviews

Explore whether the writing style of user-generated reviews, analyzed in terms of lexical and (morpho-)syntactic characteristics, can serve as a reliable source of information to predict the author of a reviewed book.

Alzetta C., Dell'Orletta F., Fazzone C., Miaschi A., Venturi G. (2023). Unmasking the Wordsmith: Revealing Author Identity through Reader Reviews. In *Proceedings of the 9th Italian Conference on Computational Linguistics* (CLiC-it 2023, Venice, Italy)

Revealing Author Identity through Reader Reviews

Explore whether the writing style of user-generated reviews, analyzed in terms of lexical and (morpho-)syntactic characteristics, can serve as a reliable source of information to predict the author of a reviewed book.

- Why?
  - Readers that share similar interests might also share some traits of their writing style.
  - Reading recommendations based on related authors are more effective than same-genre ones.
- **Book Author Prediction**: a novel task which consists of predicting the author of a book from the readers' reviews.

Alzetta C., Dell'Orletta F., Fazzone C., Miaschi A., Venturi G. (2023). Unmasking the Wordsmith: Revealing Author Identity through Reader Reviews. In *Proceedings of the 9th Italian Conference on Computational Linguistics* (CLiC-it 2023, Venice, Italy)

Literary Voices Corpus (LVC)

- LVC is a novel corpus of 11,202 book reviews written in Italian acquired from 2 social reading platforms

	Rowling	King	Tolkien	Austen	Maas	Brown	All
Goodreads							
Books	6	8	7	7	6	7	41
Reviews	1,100	1,100	1,100	1,100	1,100	1,100	6,600
Sentences Total	5,951	7,479	6,224	6,914	11,447	5,151	43,166
Tokens Total	155,653	202,027	180,680	214,921	302,687	129,684	1,185,652
Avg Sentences per Review	5.41	6.80	5.65	6.28	10.40	4.68	6.54
Avg Tokens per Review	141.50	183.66	164.25	195.38	275.17	117.89	179.64
Amazon							
Books	6	8	6	7	5	7	39
Reviews	800	800	800	749	653	800	4,602
Sentences Total	1,712	3,525	2,695	2,326	3,961	2,422	16,641
Tokens Total	21,899	69,078	48,275	40,875	81,668	40,719	302,514
Avg Sentences per Review	2.14	4.40	3.36	3.10	6.06	3.03	3.61
Avg Tokens per Review	27.37	86.34	60.34	54.57	125.06	50.89	65.73

Results

- All models outperformed a random uniform baseline on both Amazon and Goodreads
- Lexical information has more discriminative power than linguistic properties
- Adding stylistic properties does not improve the performance of a Language Model

	Rowling	King	Tolkien	Austen	Maas	Brown	All
Model				**Goodreads (accuracies)**			
Baseline	0.19	0.15	0.16	0.18	0.15	0.16	0.16
Profiling	0.21	0.18	0.26	0.27	0.40	0.25	0.26
Ngrams	0.42	0.36	0.46	0.51	0.46	0.44	0.44
BERT	**0.69**	**0.70**	**0.72**	**0.79**	**0.73**	**0.74**	**0.73**
SVM (BERT)	0.44	0.51	0.55	0.58	0.57	0.56	0.54
SVM (BERT + Profiling)	0.46	0.50	0.51	0.54	0.56	0.57	0.52
Average	0.44	0.45	0.50	**0.54**	**0.54**	0.51	0.50
				Amazon (accuracies)			
Baseline	0.16	0.15	0.17	0.16	0.16	0.14	0.16
Profiling	0.38	0.18	0.27	0.17	0.32	0.22	0.26
Ngrams	0.44	0.35	0.40	0.38	0.58	0.39	0.42
BERT	**0.57**	**0.60**	**0.56**	**0.64**	**0.72**	**0.61**	**0.61**
SVM (BERT)	0.39	0.40	0.45	0.45	0.63	0.43	0.46
SVM (BERT+Profiling)	0.41	0.42	0.39	0.46	0.56	0.36	0.43
Average	0.44	0.39	0.41	0.42	**0.56**	0.40	0.44

Selected Findings

- **Goodreads vs. Amazon**: Goodreads reviews are longer and stylistically more complex, while Amazon reviews show a more homogeneous style across genres.

- **Linguistic features**: Effective for distinguishing Amazon vs. Goodreads reviews, but less reliable for predicting book authors.

- **Broader impact**: These insights can contribute to the understanding of the complex and multifaceted phenomenon of DSR, taking as an innovative starting point the user-generated book reviews

Istituto di Linguistica Computazionale
"Antonio Zampolli"
Consiglio Nazionale delle Ricerche

Thanks for the attention!

https://alemiaschi.github.io/

@AlessioMiaschi

http://www.italianlp.it/

@ItaliaNLP_Lab

Evaluating Lexical Proficiency in Neural Language Models

- Few works focused on investigating and evaluating NLMs' abilities in tasks related to lexical proficiency

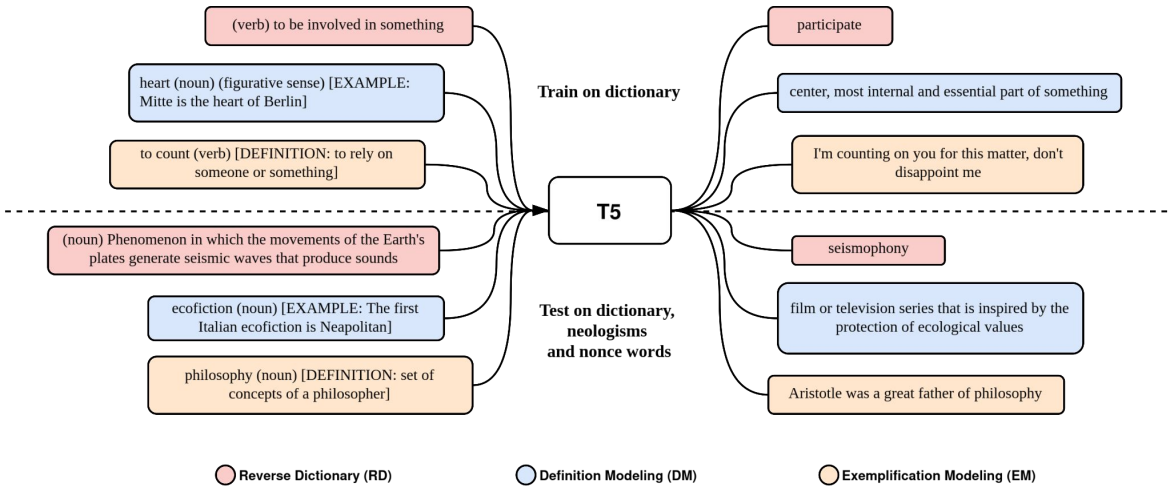
- Almost no study that goes beyond commonly lexicalized words

- We propose an evaluation framework for testing the lexical proficiency of LMs on different linguistic settings for the Italian language

Ciaccio C., Miaschi A., Dell'Orletta F. (2025). Evaluating Lexical Proficiency in Neural Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (ACL 2025, Vienna)

Our Approach

- Evaluation of Encoder-Decoder Models on a mixture of tasks that implicitly exposes the morpho-lexical link that relates lemmas to definitions



- **Reverse Dictionary:** generating a target word given a source definition
- **Definition Modeling:** generating a definition given a word
- **Exemplification Modeling:** generating a usage example given a word paired with a definition

Settings, Data and Models

- We conducted our evaluation across three different settings:
  - **Dictionary setting:** Evaluating against an unseen split of the models training dataset
  - **Neologism setting:** Evaluating against unseen neologisms that have zero to few occurrences in the models' **pretraining data**
  - **Nonce words setting:** assessing the linguistically creative abilities in creating, defining, and using nonce words (i.e. unseen words)

- Three different training/evaluation datasets:
  - **Dictionary dataset:** We developed a new resources starting from the April 2024 Wikizionario Dump + ONLI (*Osservatorio Neologico della Lingua Italiana*) neologism database
  - **Neologism dataset:** We collected a list of neologisms from various online dictionaries (appearing between 2021 to 2024) and kept only those with less then five occurrences in the pretraining dataset of our models
  - **Nonce words dataset:** We used GPT-4o to obtain a list of 100 unattested nonce words

Model	Lang	#P	#T	#T/#P
IT5-small	IT	60M	41B	683.33
IT5-base	IT	220M	41B	186.36
MT5-base	Multi	580M	6.3T	10,862.06
IT5-large	IT	738M	41B	55.55

Table 2: Models used in experiments along with the pre-training languages (*Lang*), number of parameters (*#P*), number of training tokens (*#T*) and the number of tokens per parameter (*#T/#P*).

Results

		Reverse Dictionary					Definition Modeling				Exemplification Modeling	
		Acc@1/10/100	R1	R2	CER↓	SBERT	R1	R2	RL	SBERT	PPL pred. ↓	PPL target
Dict.	**IT5-small**	*.29/.4/.53*	*41.33*	*31.19*	*50.58*	*0.68*	36.85	*23.98*	34.87	*0.61*	144.49	80.26
	IT5-base	.37/.52/.66	48	37.01	46	0.71	**39.58**	**26.54**	**37.42**	**0.65**	118.26	
	MT5-base	.33/.46/.57	43.64	33.73	47.95	0.7	*36.43*	24.58	*34.71*	0.62	*161.8*	
	IT5-large	**.39/.56/.69**	**49.7**	**38.8**	**43.83**	**0.73**	38.97	25.94	36.94	**0.65**	**112.66**	
	Avg	.34/.48/.61	45.67	35.18	47.09	0.7	37.96	25.26	35.98	0.63	134.3	
Neo.	**IT5-small**	*.06/.12/.13*	*25.39*	16.37	*71.95*	*0.55*	*18.36*	*3.44*	14.8	*0.45*	60.6	53.38
	IT5-base	.09/**.16**/.21	**33.06**	19.99	61.47	**0.6**	**21.21**	**5.36**	**16.92**	**0.53**	53.6	
	MT5-base	.08/.15/.18	26.82	*14.23*	**59.98**	0.59	18.43	3.66	*14.4*	0.48	*79.52*	
	IT5-large	**.1/.16/.27**	32.42	**20.64**	63.2	**0.6**	20.69	4.34	16.36	**0.53**	**43.44**	
	Avg	.08/.14/.19	29.4	17.8	64.05	0.58	19.67	4.2	15.62	0.5	59.15	
Nonce	**IT5-small**	—	—	—	—	—	18.91	*2.83*	15.13	*0.49*	68.35	64.28
	IT5-base	—	—	—	—	—	**21.79**	**4.19**	**17.13**	0.56	67.31	
	MT5-base	—	—	—	—	—	*18.1*	2.93	*14.15*	0.51	*84.33*	
	IT5-large	—	—	—	—	—	21.09	3.78	16.6	**0.58**	**48.05**	
	Avg	—	—	—	—	—	19.97	3.42	15.72	0.53	67.01	

Table 3: Results obtained by all the models for all the tasks (RD, DM and EM) and the three linguistically different settings: *Dict.*, *Neo.* and *Nonce*.

Results - Human Evaluation

- We collected human judgments over 100 pairs of definitions (taken from the nonce words dataset) and nonce words (generated by our models)
    - We asked 5 Italian native speakers to read each definition-word pair and express two judgments about the nonce word according to the **perceived novelty** and the **adhesion to the definition**

Results - Human Evaluation

- We collected human judgments over 100 pairs of definitions (taken from the nonce words dataset) and nonce words (generated by our models)
  - We asked 5 Italian native speakers to read each definition-word pair and express two judgments about the nonce word according to the **perceived novelty** and the **adhesion to the definition**

	Adhesion	Novelty	α
IT5-small	3.06±1.45	3.11±1.3	.51/.14
IT5-base	3.01±1.32	3.61±1.37	.29/.34
MT5-base	3.37±1.32	2.98±1.31	.37/.15
IT5-large	3.37±1.42	3.11±1.15	.41/.18
GPT-4o	3.86±1.09	3.32±1.15	.17/.07

Table 5: Mean and standard deviation for the adhesion and novelty scores given by human annotators. The column α reports the Krippendorff's Alpha between annotators for adhesion/novelty.

Results - Human Evaluation

- We collected human judgments over 100 pairs of definitions (taken from the nonce words dataset) and nonce words (generated by our models)
  - We asked 5 Italian native speakers to read each definition-word pair and express two judgments about the nonce word according to the **perceived novelty** and the **adhesion to the definition**

	Adhesion	Novelty	α
IT5-small	3.06±1.45	3.11±1.3	.51/.14
IT5-base	3.01±1.32	3.61±1.37	.29/.34
MT5-base	3.37±1.32	2.98±1.31	.37/.15
IT5-large	3.37±1.42	3.11±1.15	.41/.18
GPT-4o	3.86±1.09	3.32±1.15	.17/.07

Table 5: Mean and standard deviation for the adhesion and novelty scores given by human annotators. The column α reports the Krippendorff's Alpha between annotators for adhesion/novelty.

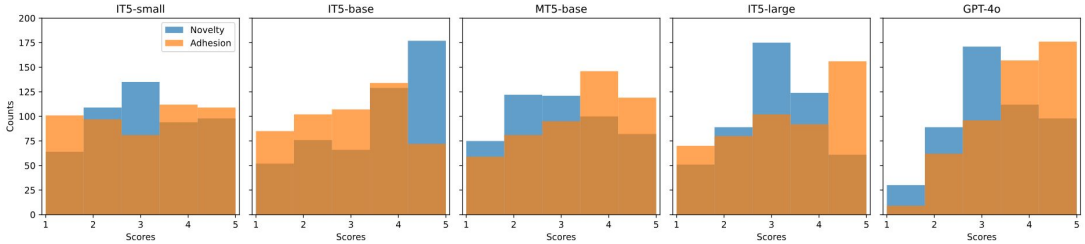


Figure 1: Distribution of novelty and adhesion human scores across the 5 values of the Likert scale for all models.

Results

Definitions	Model	Predicted Word	Adhesion	Novelty
Veicolo progettato per esplorazioni su superfici planetarie, adatto a terreni extraterrestri. [trad. *Vehicle designed for exploration on planetary surfaces, suitable for extraterrestrial terrain.*]	IT5-small	planetaro	3.0	4.2
	IT5-base	elioplano [trad. *helioplane*]	2.2	4.6
	MT5-base	cosmoplano [trad. *cosmoplane*]	3.2	4.0
	IT5-large	astroveicolo [trad. *astrovehicle*]	4.6	3.2
	GPT-4o	roverastro [trad. *astrorover*]	3.6	3.4
Vela navigabile che raccoglie dati geologici mentre si sposta su laghi o mari, utilizzata in esplorazioni scientifiche. [trad. *Navigable sail that collects geological data as it moves across lakes or seas, used in scientific exploration.*]	IT5-small	geonauta [trad. *geonaut*]	4.6	2.4
	IT5-base	ecovela [trad. *ecosail*]	4.4	1.8
	MT5-base	vettolaghiera	2.0	4.4
	IT5-large	idrovedetta [trad. *hydropatrol*]	4.6	2.8
	GPT-4o	geonave [trad. *geoship*]	4.0	3.2
Una tavola o superficie capace di mostrare visivamente il passare del tempo, evidenziando i cambiamenti avvenuti su di essa. [trad. *A table or surface capable of visually showing the passage of time, highlighting the changes that have occurred on it.*]	IT5-small	cromatopompa	1.2	3.8
	IT5-base	cronopalestra [trad. *chronogym*]	2.0	5.0
	MT5-base	retrotavola [trad. *retrotable*]	2.2	3.0
	IT5-large	cronotavolo [trad. *chronotable*]	4.4	3.0
	GPT-4o	cronotavola [trad. *chronotable*]	3.6	3.6
Forma d'arte che utilizza nebbie artificiali e giochi di luce per creare installazioni immersive. [trad. *An art form that uses artificial fog and light effects to create immersive installations.*]	IT5-small	immersivismo [trad. *immersivism*]	3.8	2.4
	IT5-base	metacaduta [trad. *metafall*]	2.0	4.6
	MT5-base	fotoart [trad. *photoart*]	3.4	2.6
	IT5-large	nebbiografia [trad. *foggraphy*]	4.4	3.0
	GPT-4o	nebbioarte [trad. *fogart*]	3.6	3.6
Fenomeno in cui i movimenti delle placche terrestri generano onde sismiche che producono suoni dissonanti, studiato in geologia e acustica. [trad. *Phenomenon in which the movements of the earth's plates generate seismic waves that produce dissonant sounds, studied in geology and acoustics.*]	IT5-small	biogeoacustica [trad. *biogeoacoustics*]	4.4	3.4
	IT5-base	sismofonia [trad. *seismophony*]	3.0	4.0
	MT5-base	sismismo [trad. *seismism*]	3.0	4.0
	IT5-large	sismofonia [trad. *seismophony*]	4.2	3.2
	GPT-4o	sismofonia [trad. *seismophony*]	4.2	2.0

Table 6: Sample of generated nonce words (we tried to provide a translation when possible), along with adhesion and novelty average scores, for all the models. The definitions are those generated by GPT-4o.

"Astroveicolo"

Selected Findings

- Larger, monolingual models generally outperformed their multilingual counterparts

- Despite the drop in performance with low-frequency neologisms and nonce words, the rank between models remained consistent

- The models' ability to generate novel and coherent nonce words further indicates LMs are capable of **learning approximations of word formation rules**, rather than relying solely on memorization