Linguistic Profiling of Large Language Models

CNR-IVI workshop on AI Technologies, June 9 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 (RTDA) 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.

About me and... the team!

I am a full-time researcher (RTDA) 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.

ItaliaNLP Lab — Italian Natural Language Processing Lab@ILC — www.italianlp.it

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

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
- Cristiano Ciaccio
- Chiara Fazzone
- Ruben Piperno
- Marta Sartor

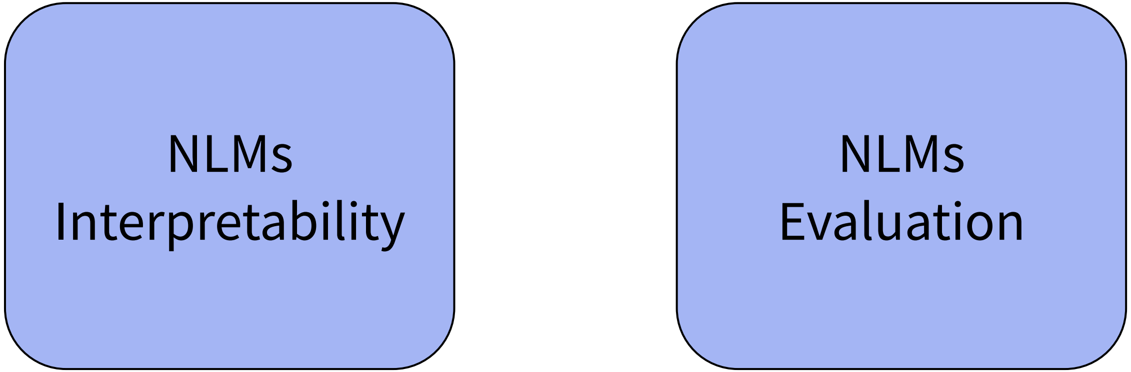
PhD Students:
- Luca Dini
- Lucia Domenichelli
- Michele Papucci

+ Master/Undergraduate/Visiting Students

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

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

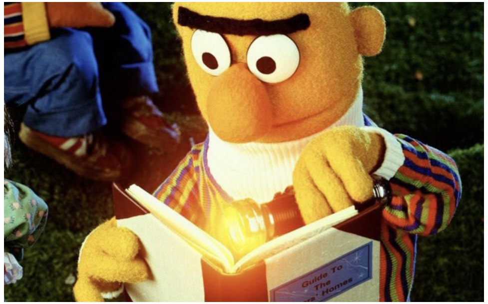
NLMs Interpretability

NLMs Evaluation

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."

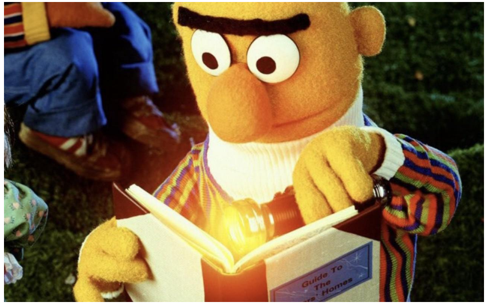
Belinkov and Glass, Analysis Methods in Neural Language Processing: A Survey (2019). In Transactions of ACL, Volume 7, pages 49-72.



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

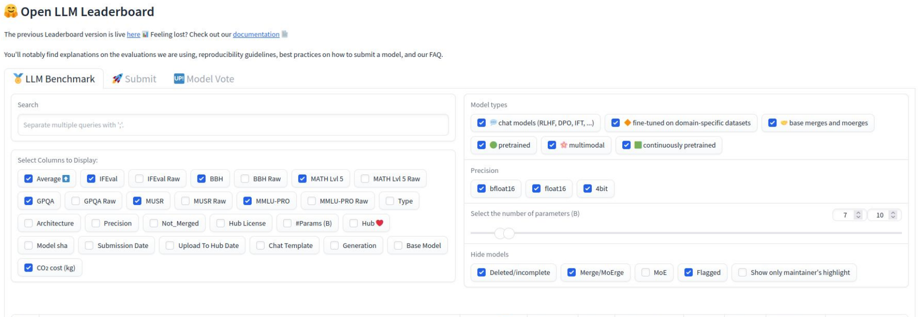
- The rapid development and widespread adoption of state-of-the-art Neural Language Models (NLMs) have increased the need for studies focused on their **interpretability** and the **evaluation** of their abilities

NLMs Interpretability

NLMs Evaluation

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)



T	Model	Average	IFEval	BBH	MATH Lvl 5	GPQA	MUSR	MMLU-PRO	CO_2 cost (kg)
	dfurman/CalmeRys-78B-Orpo-v0.1	51.24	81.63	61.92	40.71	20.02	36.37	66.8	13
	MaziyarPanahi/calme-2.4-rys-78b	50.71	80.11	62.16	40.41	20.36	34.57	66.69	12.98
	rombodawg/Rombos-LLM-V2.5-Qwen-72b	45.91	71.55	61.27	50.68	19.8	17.32	54.83	16.03
	zetasepic/Qwen2.5-72B-Instruct-abliterated	45.29	71.53	59.91	46.15	20.92	19.12	54.13	18.81
	dnhkng/RYS-XLarge	45.13	79.96	58.77	41.24	17.9	23.72	49.2	13.58
	rombodawg/Rombos-LLM-V2.5-Qwen-32b	44.57	68.27	58.26	41.99	19.57	24.73	54.62	17.91
	MaziyarPanahi/calme-2.1-rys-78b	44.56	81.36	59.47	38.9	19.24	19	49.38	14.33
	MaziyarPanahi/calme-2.3-rys-78b	44.42	80.66	59.57	38.97	20.58	17	49.73	13.3
	MaziyarPanahi/calme-2.2-rys-78b	44.26	79.86	59.27	39.95	20.92	16.83	48.73	13.52

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

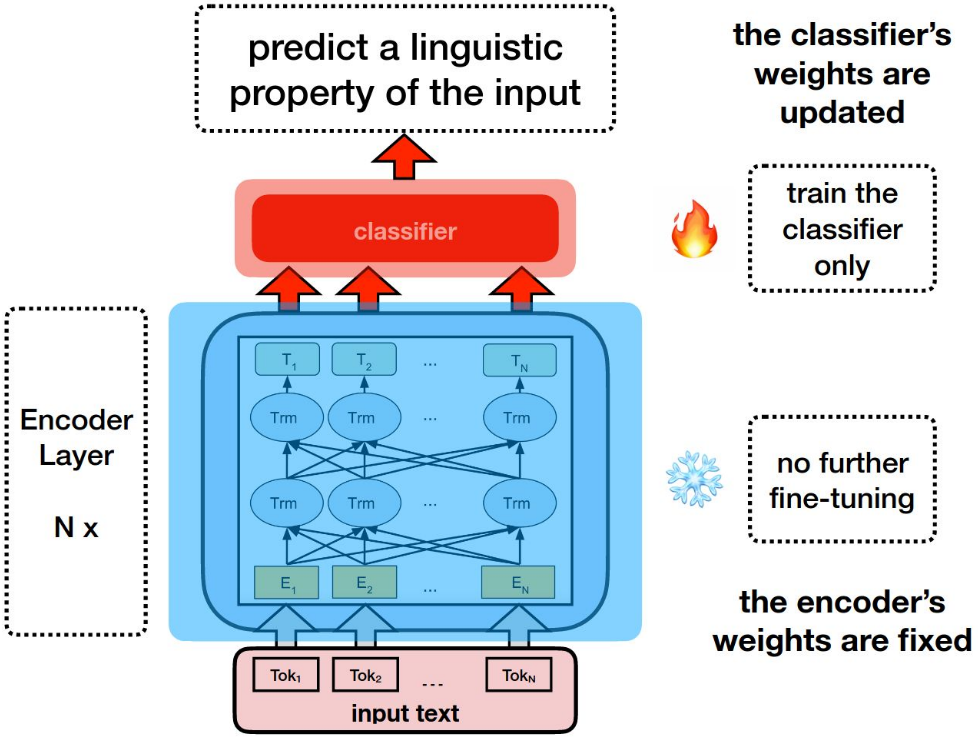
- The "*linguistic profiling*" methodology (van Halteren, 2004) assumes that wide counts of linguistic features are particularly helpful in the resolution of several NLP tasks, e.g.:
  - Text Profiling (e.g. text readability, textual genres)
  - Author Profiling (e.g. author's age and native language)

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

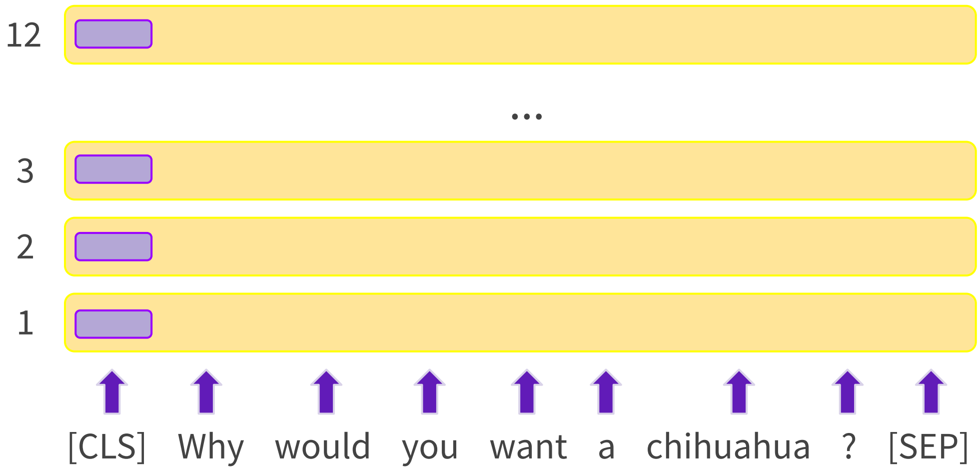
- ProfilingUD (Brunato et al., 2020) is a web–based application that performs linguistic profiling of a text, or a large collection of texts, for multiple languages

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

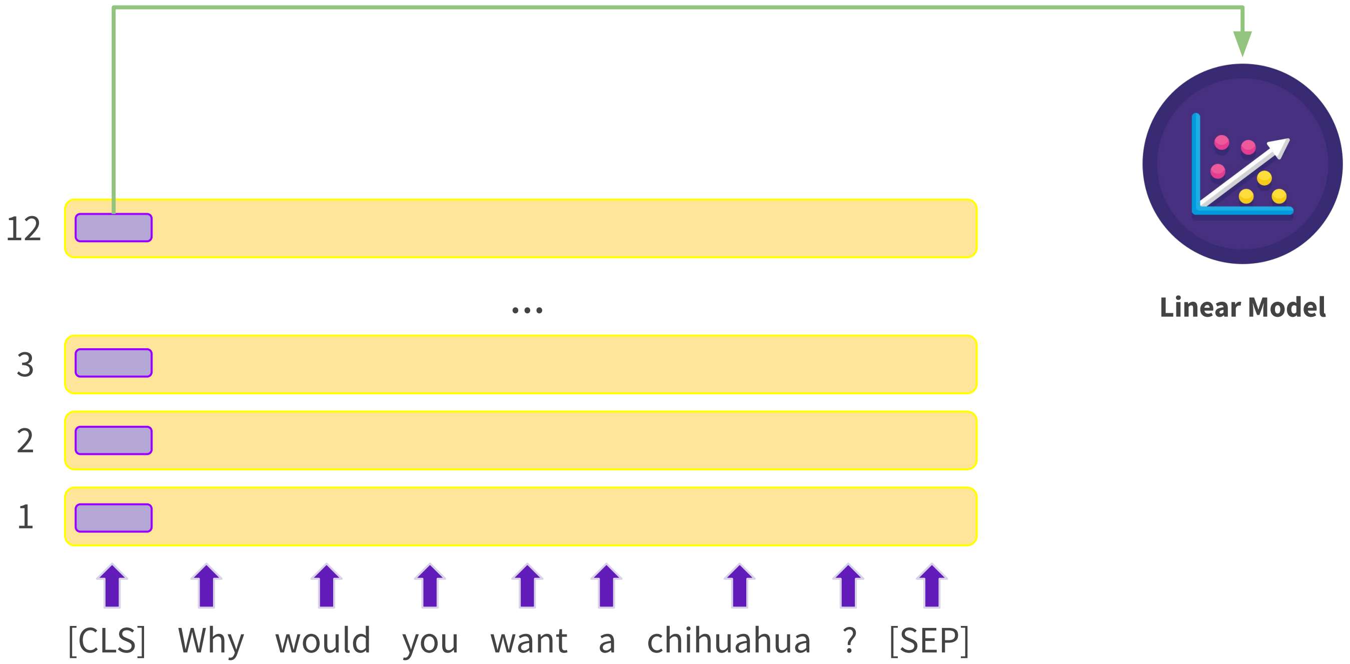
- Link: http://linguistic-profiling.italianlp.it/

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

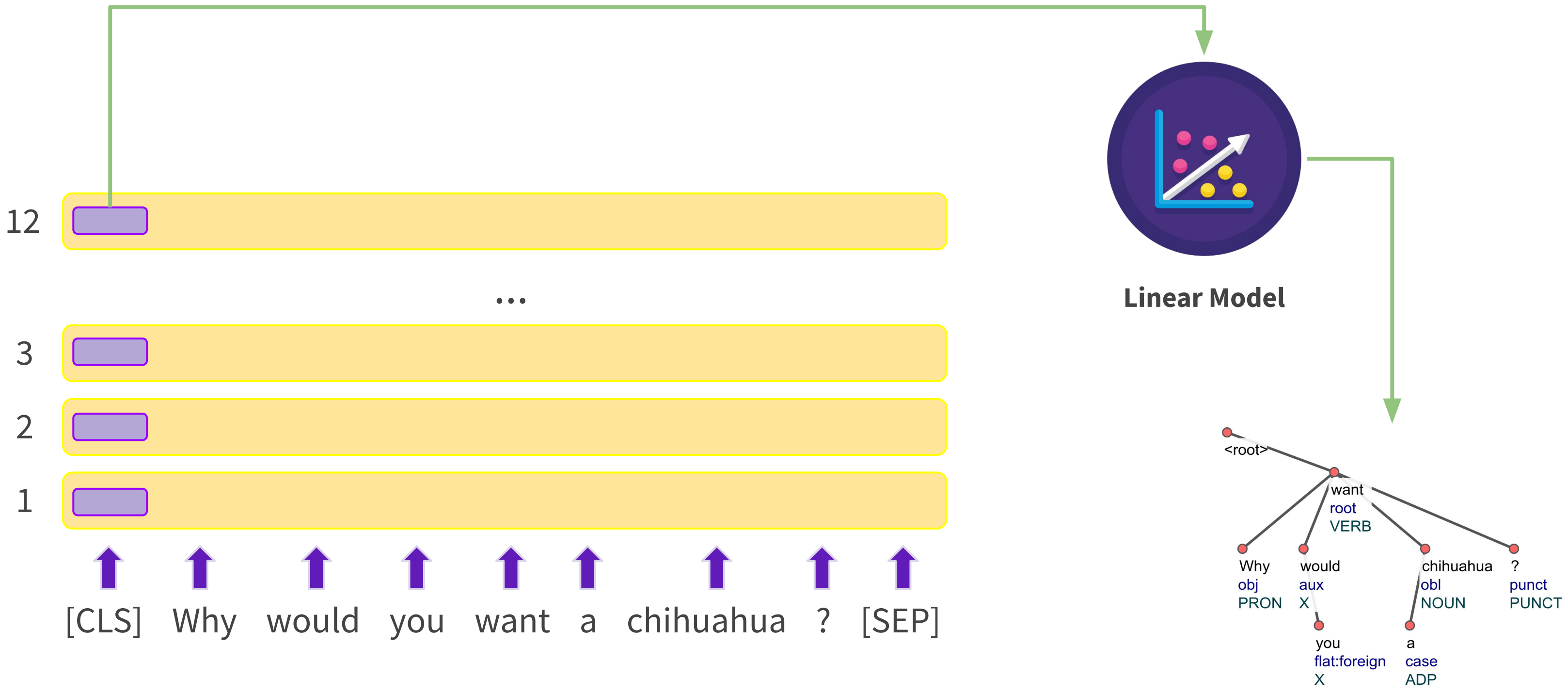
Profiling Neural Language Models



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 (Devlin et al., 2018)

Research questions:

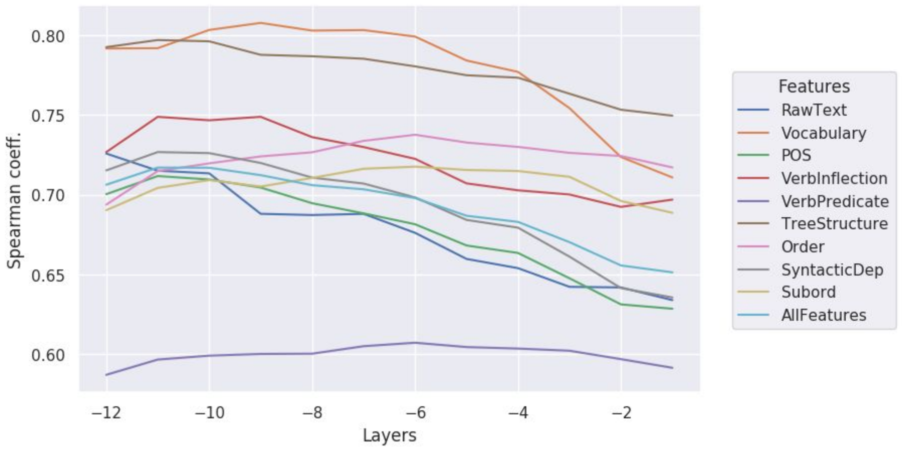
1. What kind of linguistic properties are encoded in a pre-trained version of BERT?

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

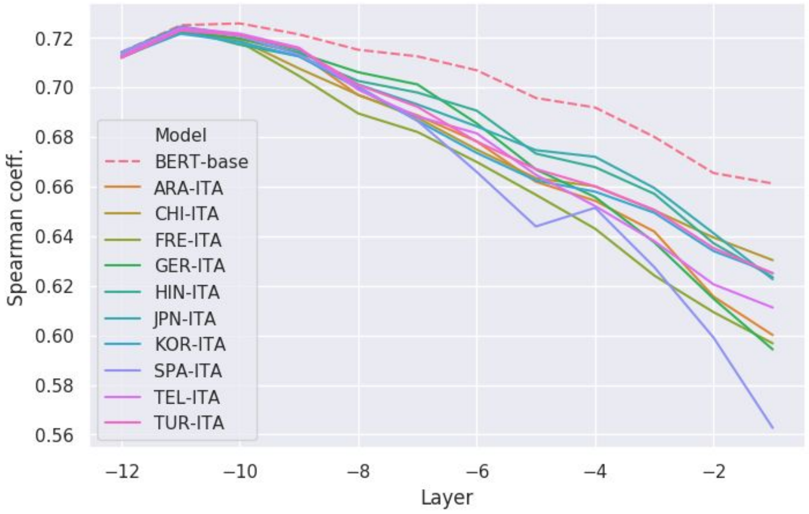
Miaschi A., Brunato D., Dell’Orletta F., Venturi G. (2020). Linguistic Profiling of a Neural Language Models. In *Proceedings of the 28th International Conference on Computational Linguistics* (COLING 2020, Barcelona) **[Outstanding paper for COLING 2020]**

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

Pre fine-tuning:



Post fine-tuning:



Linguistic Knowledge Can Enhance Encoder-Decoder Models

- Motivations:
  - Understanding "how linguistic concepts that were common as features in NLP systems are captured in neural networks" (Belinkov & Glass, *Transactions of the Association for Computational Linguistics 2019*) has been the focus of many recent studies
  - Fine-tuning on a intermediate supporting task and then on the target task consecutively is highly beneficial to improve pre-trained model's performance (Weller et al., *ACL 2022*)

Miaschi A., Dell'Orletta F., Venturi G. (2024). Linguistic Knowledge Can Enhance Encoder-Decoder Models (*If You Let It*). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024, Turin)

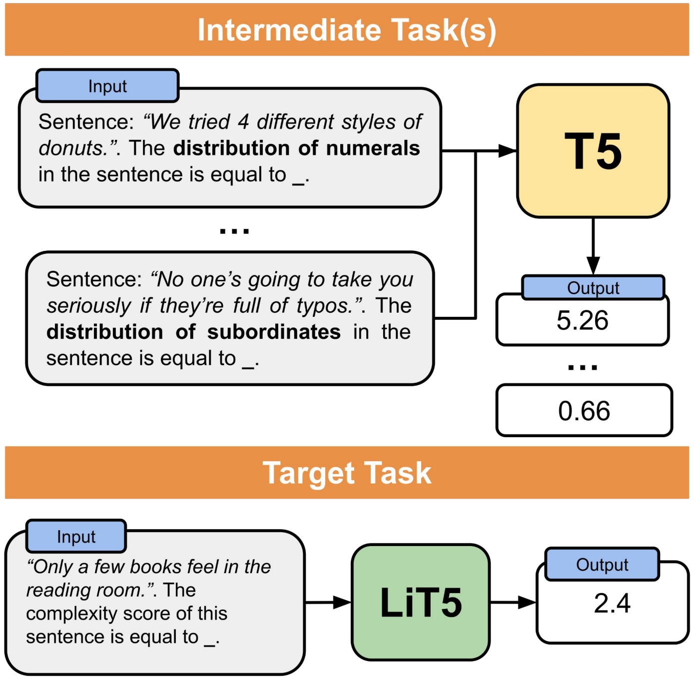
Linguistic Knowledge Can Enhance Encoder-Decoder Models

- Motivations:
  - Understanding "how linguistic concepts that were common as features in NLP systems are captured in neural networks" (Belinkov & Glass, *Transactions of the Association for Computational Linguistics 2019*) has been the focus of many recent studies
  - Fine-tuning on a intermediate supporting task and then on the target task consecutively is highly beneficial to improve pre-trained model's performance (Weller et al., *ACL 2022*)

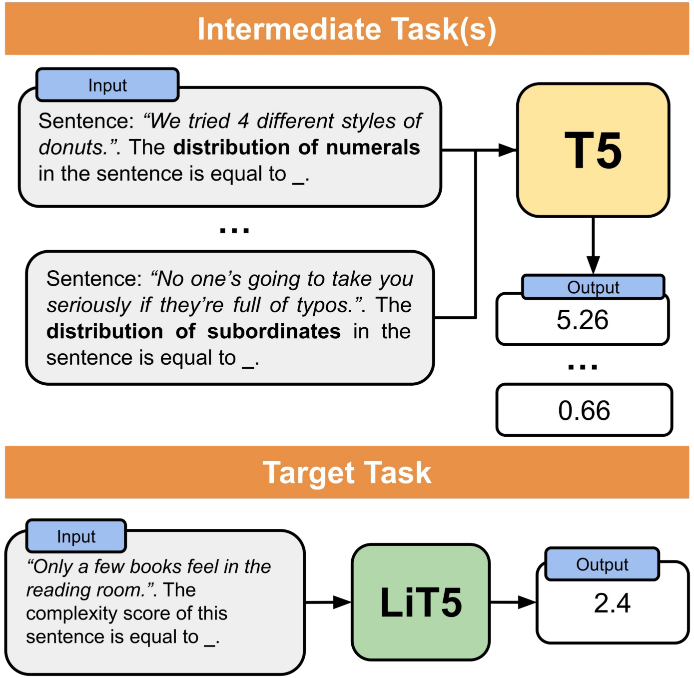
Does a step of intermediate fine-tuning on linguistic tasks enhance the prediction on a target task that strongly relies on linguistic knowledge?

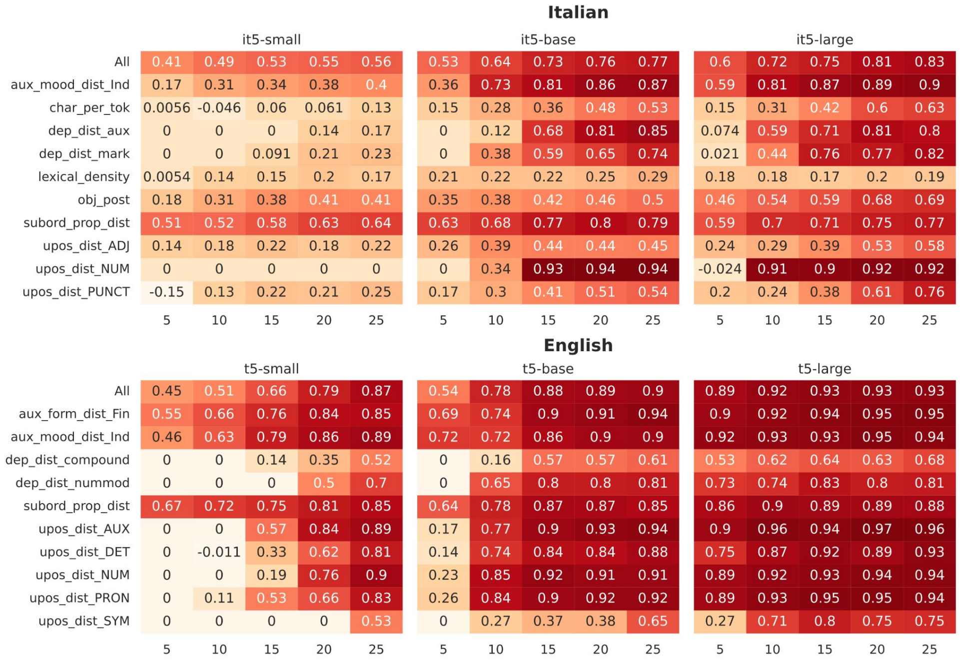
Miaschi A., Dell'Orletta F., Venturi G. (2024). Linguistic Knowledge Can Enhance Encoder-Decoder Models (*If You Let It*). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024, Turin)

Linguistic Knowledge Can Enhance Encoder-Decoder Models

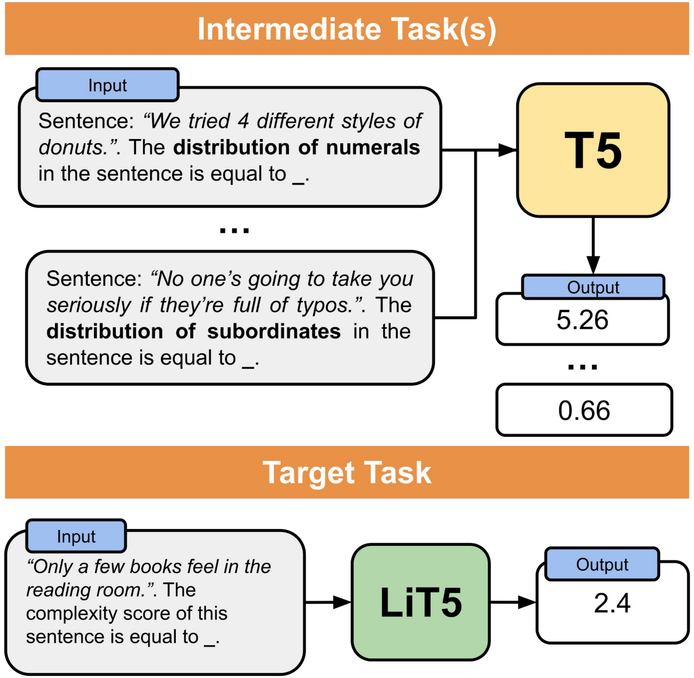


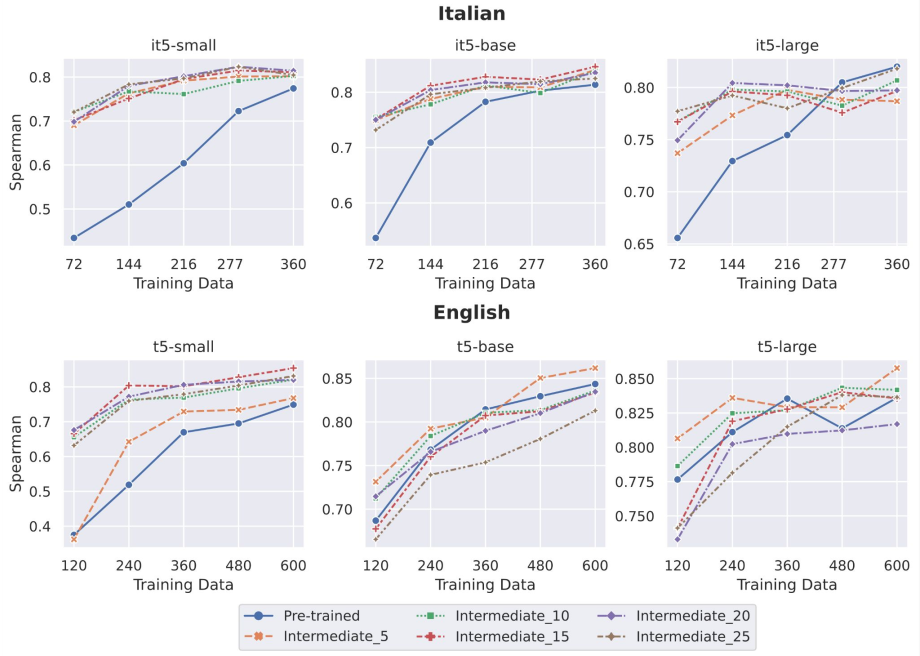
Linguistic Knowledge Can Enhance Encoder-Decoder Models





Linguistic Knowledge Can Enhance Encoder-Decoder Models





Selected Findings

- Informing models linguistically over several epochs allows them to progressively improve their degree of language proficiency.

- The method of linguistic enhancement is particularly effective, especially when applied to smaller models and in scenarios with limited availability of target training data.

- Small models, refined through intermediate fine-tuning, can frequently surpass the performance of larger models that have not undergone this intermediate refinement process.

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?

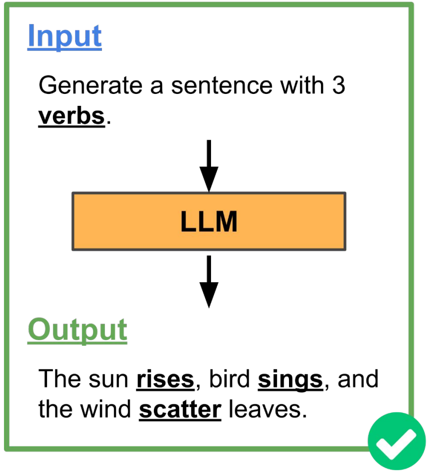
Miaschi A., Dell'Orletta F., Venturi G. (2024). Evaluating Large Language Models via Linguistic Profiling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2024, Miami, Florida)

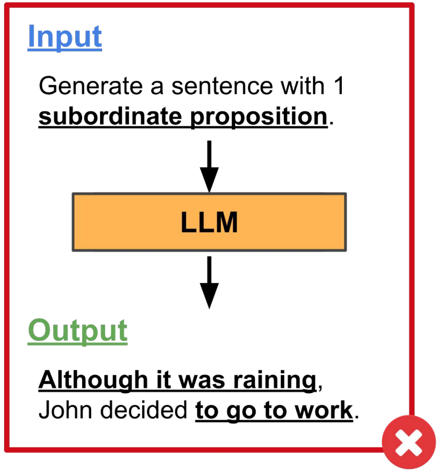
Our Approach

- 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.

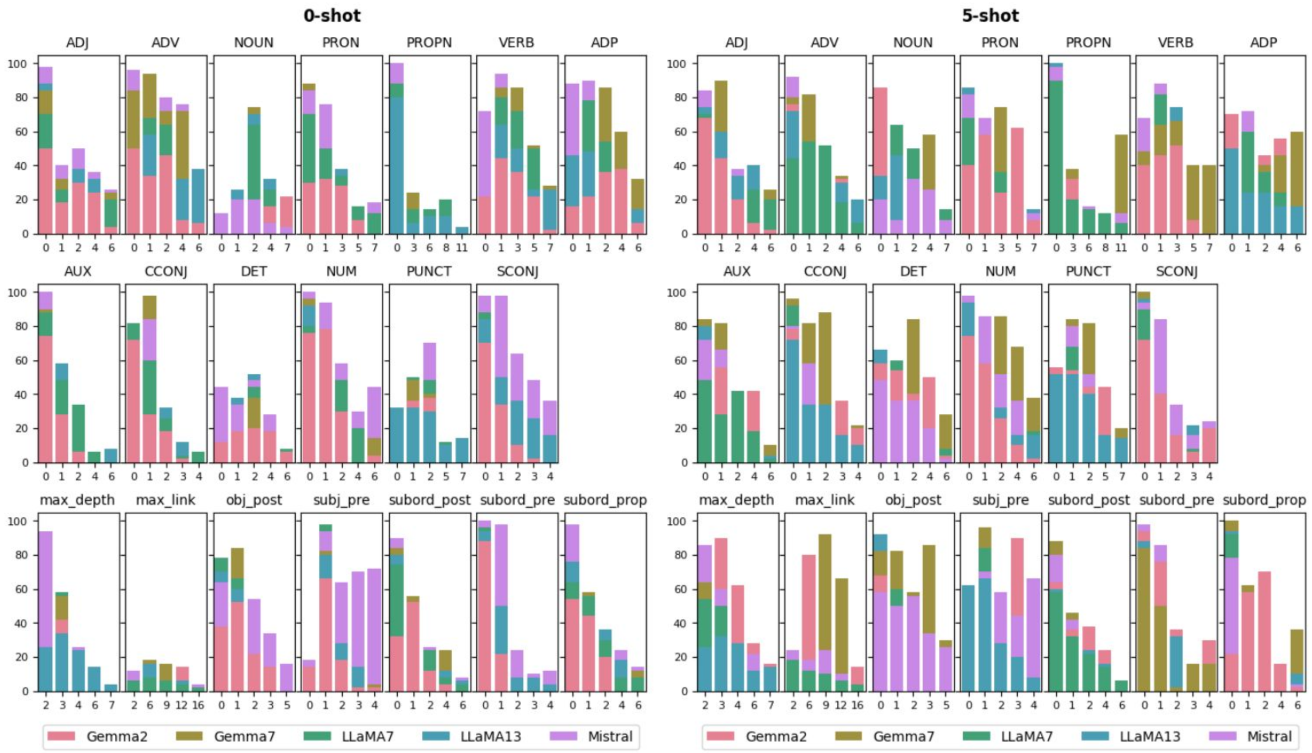
  - **Spearman coefficient**: correlation coefficients between the increasing property values extracted from EWT and those extracted from the sentences generated by the models.

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**

- The differences between the scores of the two tested metrics seem to confirm that **they offer two distinct perspectives on models' behaviour**

Conclusion and Future Directions

- LLMs have reached astonishing performance in almost all NLP tasks
- Their success has led to a growing interest in their evaluation, alongside studies analyzing their behavior and internal mechanisms
- Despite significant progress, there is still a lot to do!

Ongoing and Future Directions:

- Studying and evaluating generalization of LLMs across different scenarios, domains and languages (Hupkes et al., 2023)
- Mechanistic Interpretability (Elhage et al, 2021; Olsson et al., 2022)
- Analyzing and controlling the "linguistic profile" of generated texts to develop more robust Machine-Generated Text (MGT) detection systems → *"Stress-testing Machine Generated Text Detection: Shifting Language Models Writing Style to Fool Detectors"* (Pedrotti et al., 2025), accepted at Findings of ACL 2025

Conclusion and Future Directions

- LLMs hav
- Their suc ... de studies analyzin
- Despite s

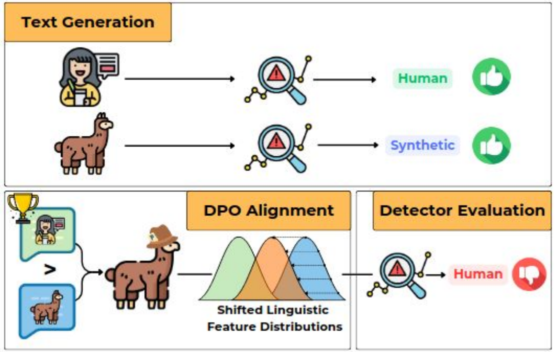
Ongoing and Fut

- Studying a ... nguages (Hupkes et al., 2023)
- Mechanisti
- Analyzing a ... achine-Generated Text (MGT) ... *guage Models Writing Styl*

Stress-testing Machine Generated Text Detection: Shifting Language Models Writing Style to Fool Detectors

Andrea Pedrotti[α], Michele Papucci[β,γ], Cristiano Ciaccio[γ], Alessio Miaschi[γ], Giovanni Puccetti[α], Felice Dell'Orletta[γ], Andrea Esuli[α]

[α] Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (CNR-ISTI)
{name.surname}@isti.cnr.it
[β] Department of Computer Science, University of Pisa
[γ] ItaliaNLP Lab, Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC)
{name.surname}@ilc.cnr.it



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