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

Evaluating Linguistic Abilities of Neural Language Models

Genova, December 9 2024

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.

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

RTDA:
- 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/

Outline

1. An introduction to Language Models (LMs)
2. Neural Language Models (NLMs)
3. Transformer-based LMs
4. Interpreting and Evaluating NLMs
5. Conclusion and Future Directions

An introduction to Language Models (LMs)

Language Models

- In the context of numerous studies in Computational Linguistics (CL) and Natural Language Processing (NLP), it is assumed that language can be viewed as a *probabilistic system*

- To describe and explain the functioning of a probabilistic system, it is necessary to define a *(probabilistic) model*

- A **language model**, therefore, is nothing more than a system capable of assigning a probability to word sequences

Probabilistic Language Models

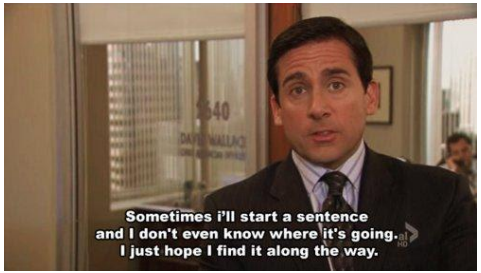
- Given a sequence of words w_1, ..., w_n, we can represent the sequence as:

$$p(w_1, ..., w_n) = p(w_1)p(w_2|w_1)...p(w_n|w_1, ..., w_{n-1})$$

Probabilistic Language Models

- Given a sequence of words $w_1, ..., w_n$, we can represent the sequence as:

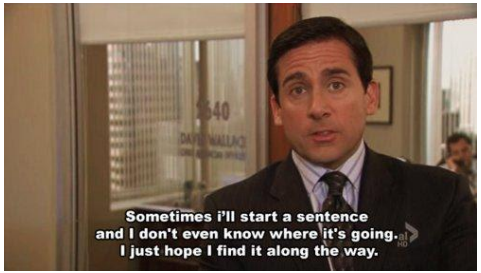
$$p(w_1, ..., w_n) = p(w_1)p(w_2|w_1)...p(w_n|w_1, ..., w_{n-1})$$



Probabilistic Language Models

- Given a sequence of words $w_1, ..., w_n$, we can represent the sequence as:

$$p(w_1, ..., w_n) = p(w_1)p(w_2|w_1)...p(w_n|w_1, ..., w_{n-1})$$



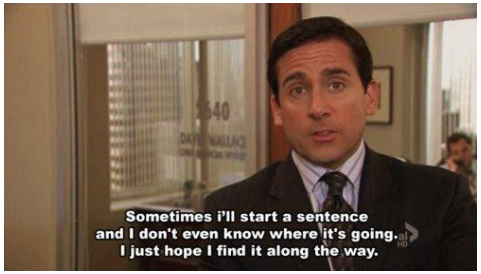
- As a consequence, the probability of the next word in a sequence given the preceding context can be defined as:

$$p(w_n|w_1, ..., w_{n-1}) = \frac{Count(w_1, ..., w_{n-1}, w_n)}{Count(w_1, ..., w_{n-1})}$$

Probabilistic Language Models

- Given a sequence of words w_1, ..., w_n, we can represent the sequence as:

$$p(w_1, ..., w_n) = p(w_1)p(w_2|w_1)...p(w_n|w_1, ..., w_{n-1})$$



- As a consequence, the probability of the next word in a sequence given the preceding context can be defined as:

$$P(the|its\ water\ is\ so\ transparent\ that) = \frac{C(its\ water\ is\ so\ transparent\ that\ the)}{C(its\ water\ is\ so\ transparent\ that)}$$

Probabilistic Language Models (ngrams)

- N-grams LMs can be exploited to approximate the probability of the next word as follows:

$$p(w_i|w_1, ..., w_{t-1}) \approx p(w_i|w_{i-N}, ..., w_{i-1})$$

Probabilistic Language Models (ngrams)

- N-grams LMs can be exploited to approximate the probability of the next word as follows:

$$p(w_i|w_1, ..., w_{t-1}) \approx p(w_i|w_{i-N}, ..., w_{i-1})$$

- As *N* increases, the approximation becomes more accurate, but the complexity grows exponentially.
- Conversely, when *N=1*, the model requires less information, but its performance is significantly lower.

Probabilistic Language Models (ngrams)

Before

P(I saw a cat on a mat) =

P(I)
· P(saw | I)
· P(a | I saw)
· P(cat | I saw a)
· P(on | I saw a cat)
· P(a | I saw a cat on)
· P(mat | I saw a cat on a)

After (3-gram)

P(I saw a cat on a mat) =

P(I) → P(I)
· P(saw | I) → · P(saw | I)
· P(a | I saw) → · P(a | I saw)
· P(cat | ~~I~~ saw a) → · P(cat | saw a)
· P(on | ~~I saw~~ a cat) → · P(on | a cat)
· P(a | ~~I saw a~~ cat on) → · P(a | cat on)
· P(mat | ~~I saw a cat~~ on a) → · P(mat | on a)

ignore use

Language Modeling, NLP Course, Lena Voita,
https://lena-voita.github.io/nlp_course/language_modeling.html

Probabilistic Language Models (ngrams)

- N-gram-based language models, however, have several limitations:
    - Regardless of the value assigned to *N*, the model will always be an approximation of the true probability distribution.
    - Due to the exponential growth in complexity, the choice of *N* will always fall on particularly low values (usually 2 or 3).
    - An N-gram model cannot **generalize to new word sequences**.

Word representations

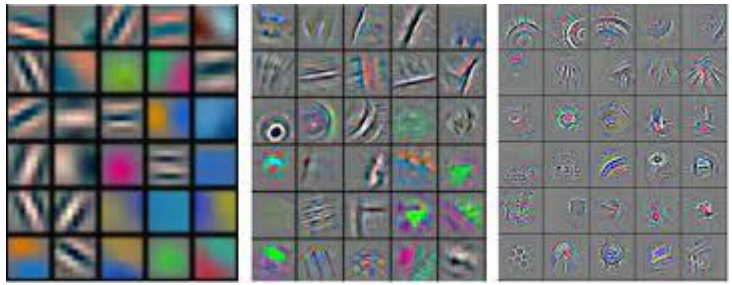
- Words can be considered the basic units of a language model

- To understand a language, it is first necessary to know the meaning of the words that compose it

- To comprehend a language, a (computational) language model should be able to *represent* the words of that language

A *representation* problem

- *Representation learning* is a central problem in the context of Artificial Intelligence, neuroscience, and semantics

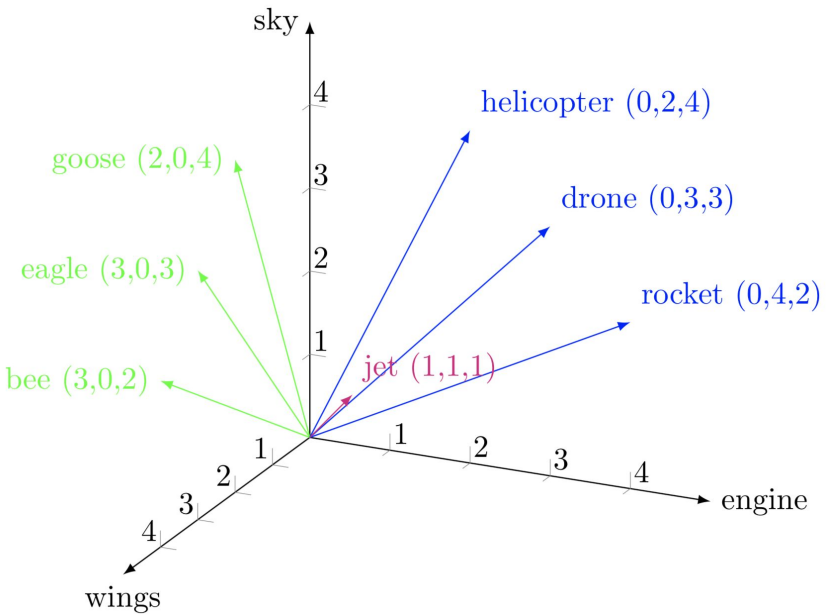


representation

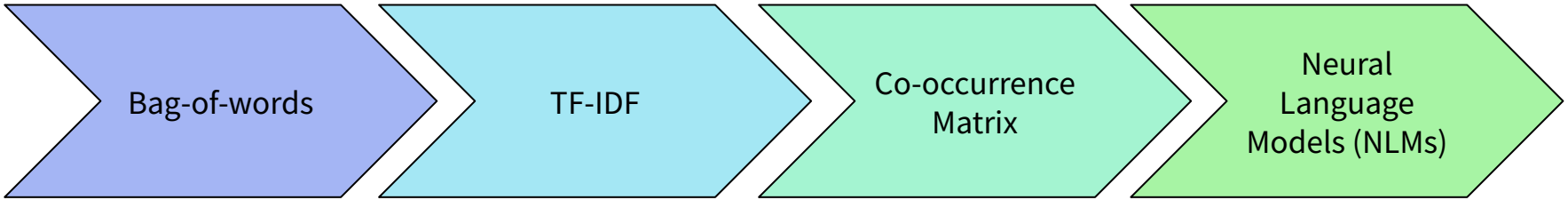
“monkey”

Word representations

- From a computational perspective, the most intuitive method to represent a word is to associate it with a **vector of numbers**



Word representations



Neural Language Model (NLM)

Neural Language Model (NLM)

- A NLM is a Neural Network (NN) trained to approximate the **language modeling** function

Neural Language Model (NLM)

- A NLM is a Neural Network (NN) trained to approximate the **language modeling** function

- A probabilistic LM defines the probability of a sequence $s = [w_1, w_2, \ldots, w_n]$ as:

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

Neural Language Model (NLM)

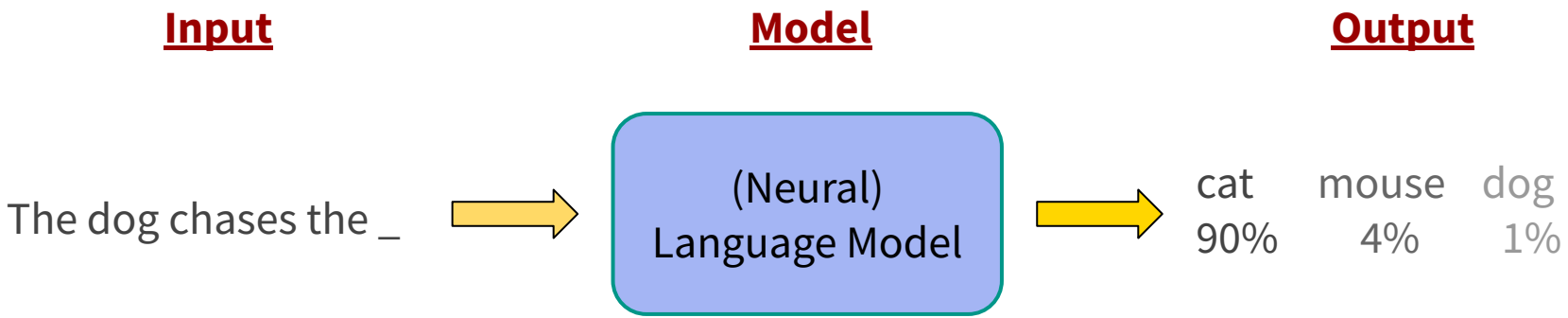
- A NLM is a Neural Network (NN) trained to approximate the **language modeling** function

- A probabilistic LM defines the probability of a sequence $s = [w_1, w_2, \ldots, w_n]$ as:

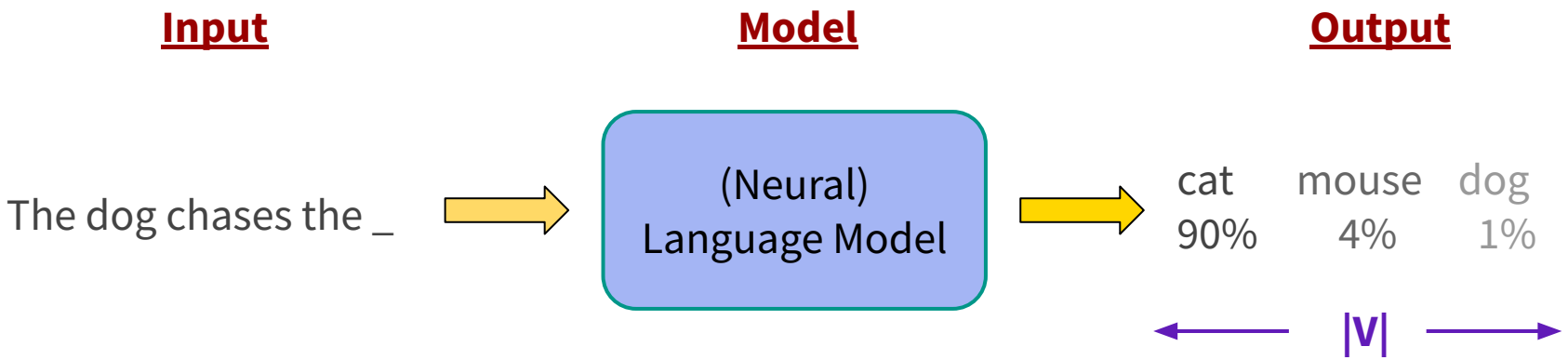
$$P(s) = \prod_{i=1}^{N} P(w_i | w_1, w_2, ..., w_{i-1})$$

- Bengio et al. (2003) proposed a model that approximate the LM function relying on the architecture of a NN → **Neural Probabilistic Language Model**

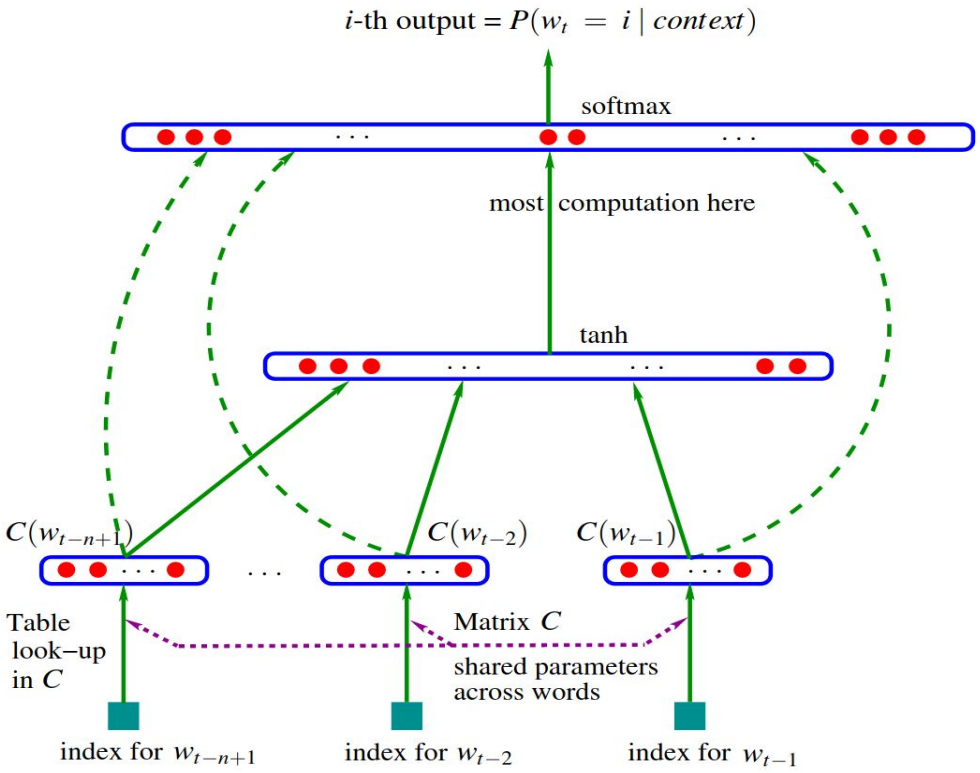
Neural Language Model (NLM)



Neural Language Model (NLM)



Neural Language Model (NLM)



Transformer Model

Transformer

- The most widely used architecture nowadays is the **Transformer**, first introduced in: Attention is All you Need (Vaswani et al., 2017)

Transformer

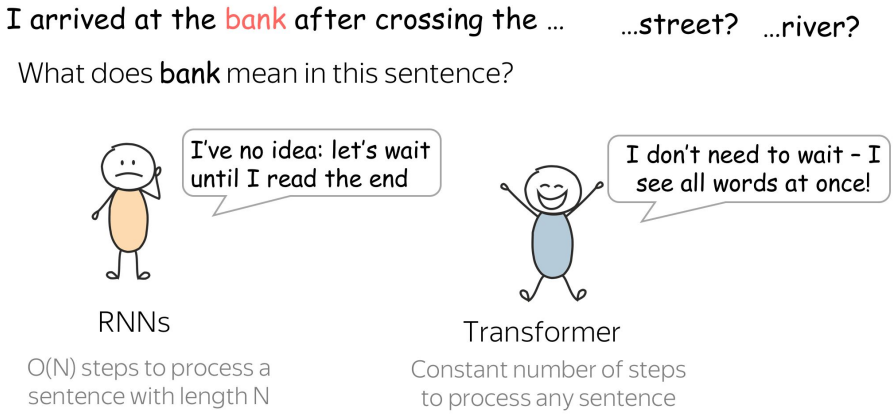
- The most widely used architecture nowadays is the **Transformer**, first introduced in: Attention is All you Need (Vaswani et al., 2017)

- The Transformer is a neural network (Encoder-Decoder) that leverages a specific mechanism, **Attention**, to focus on key portions of a sentence and create contextual word representations.

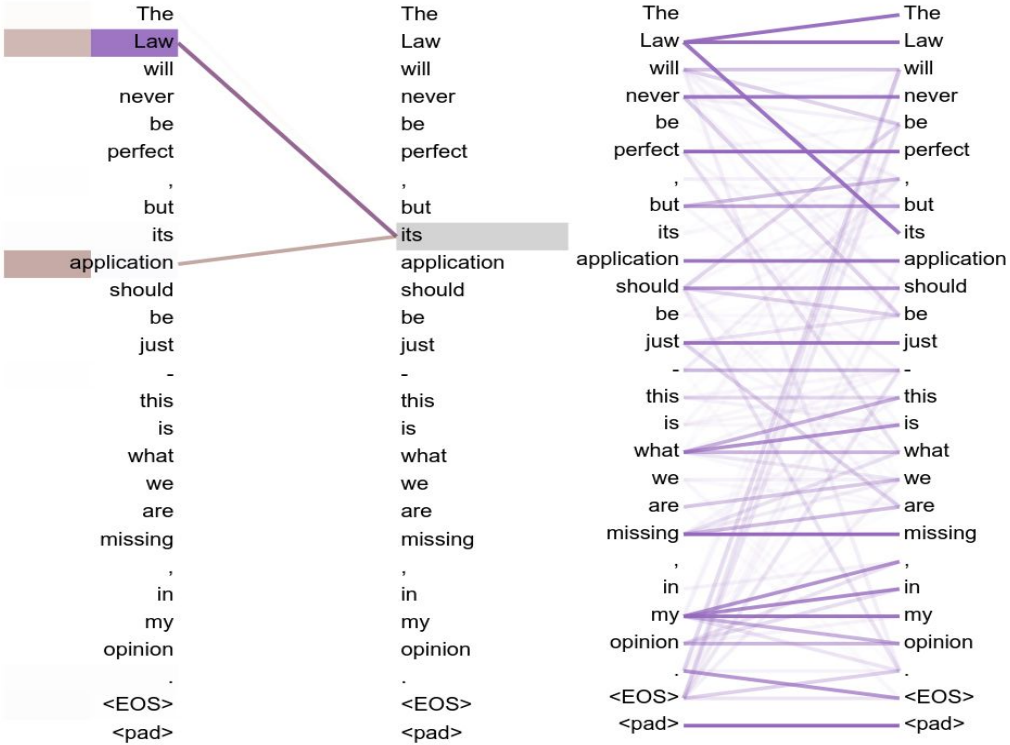
Transformer

- The most widely used architecture nowadays is the **Transformer**, first introduced in: [Attention is All you Need (Vaswani et al., 2017)]

- The Transformer is a neural network (Encoder-Decoder) that leverages a specific mechanism, **Attention**, to focus on key portions of a sentence and create contextual word representations.

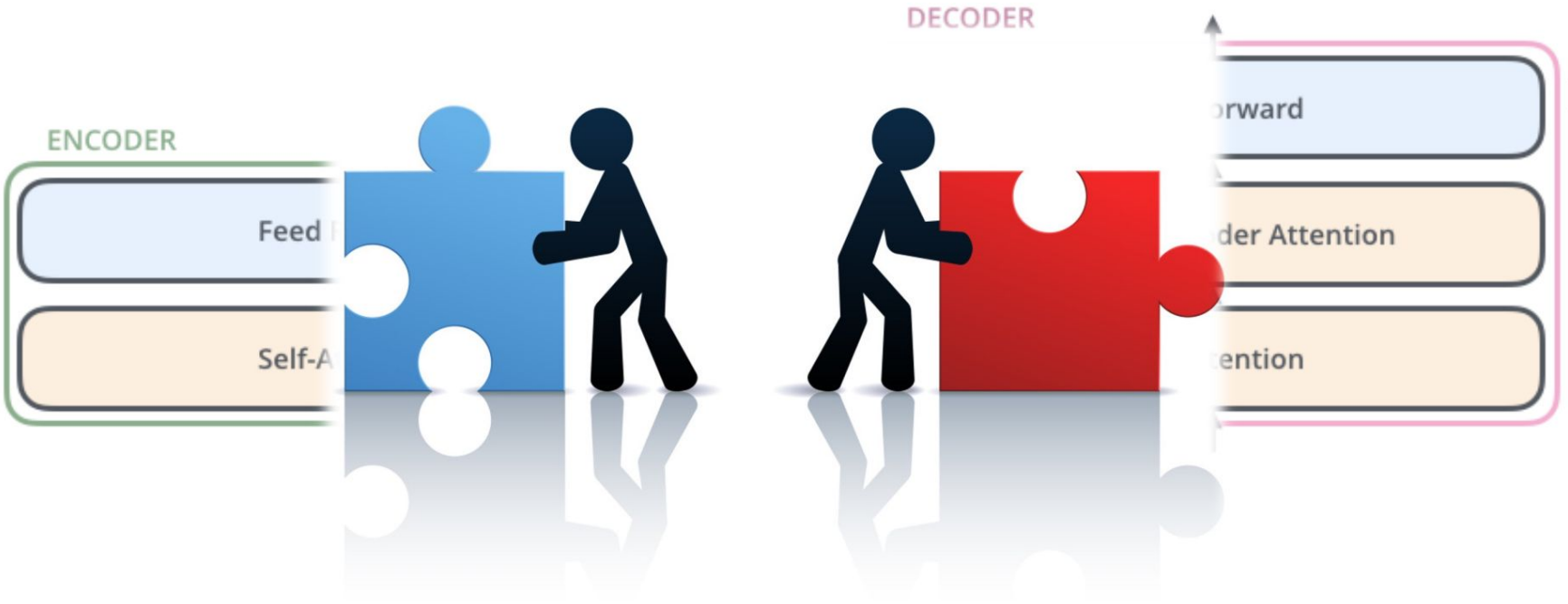


Transformer - Attention



Transformer-based NLMs

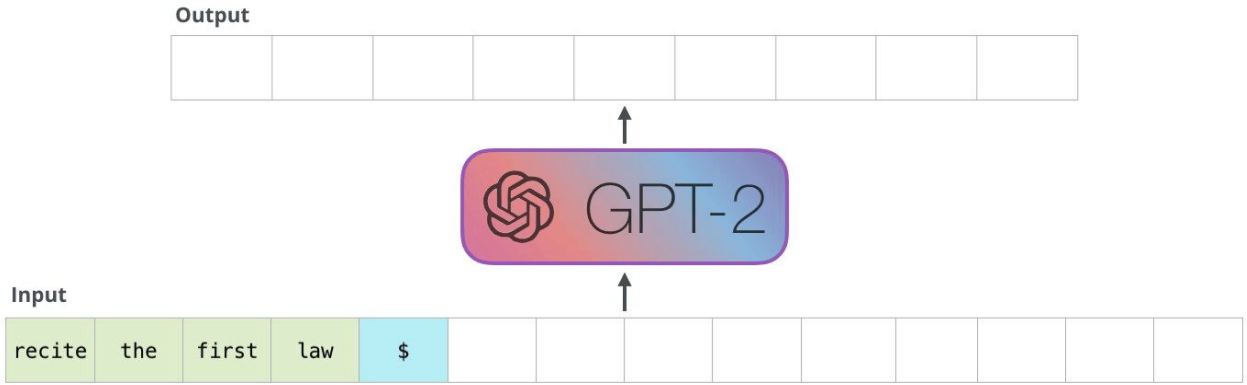
Transformer-based NLMs



Source: [Large Language Models and How To Instruct Them (CLiC-it 2023 Tutorial)]

GPT (Radford et al, 2018), GPT-2 (Radford et al, 2019), etc

- Decoder Transformer model
- Trained on the **Language Modeling** (**LM**) task
- Generative model

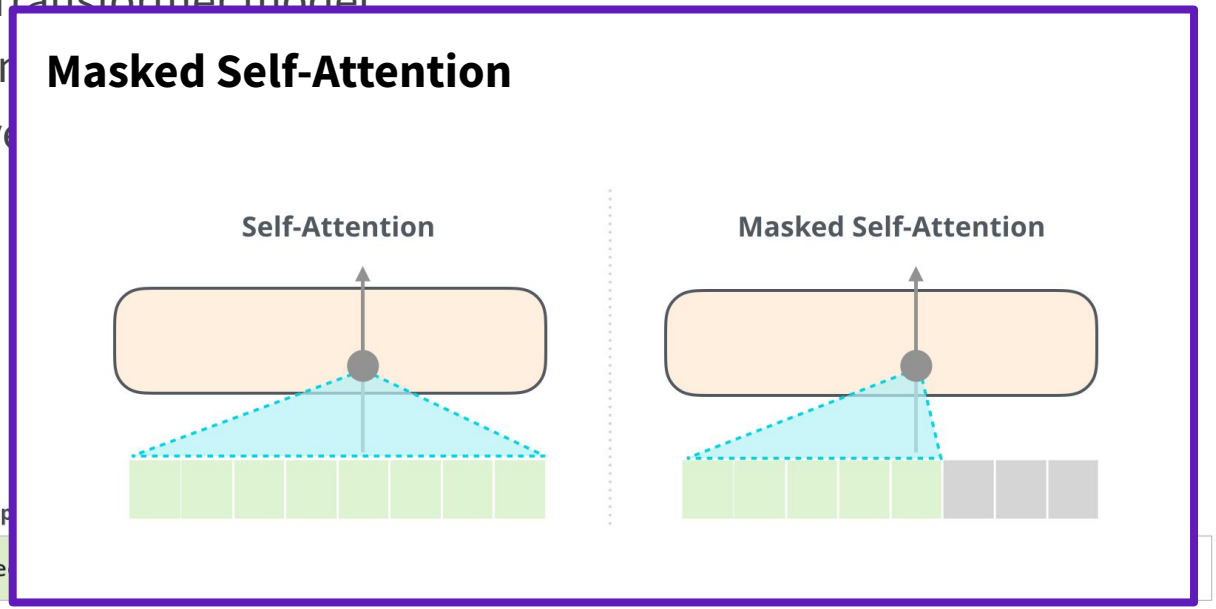


Improving Language Understanding by Generative Pre-Training (Radford et al,, 2018), https://openai.com/research/language-unsupervised

Language Models are Unsupervised Multitask Learners (Radford et al., 2019), https://openai.com/research/better-language-models

GPT (Radford et al, 2018), GPT-2 (Radford et al, 2019), etc

- Decoder Transformer model
- Trained on
- Generative

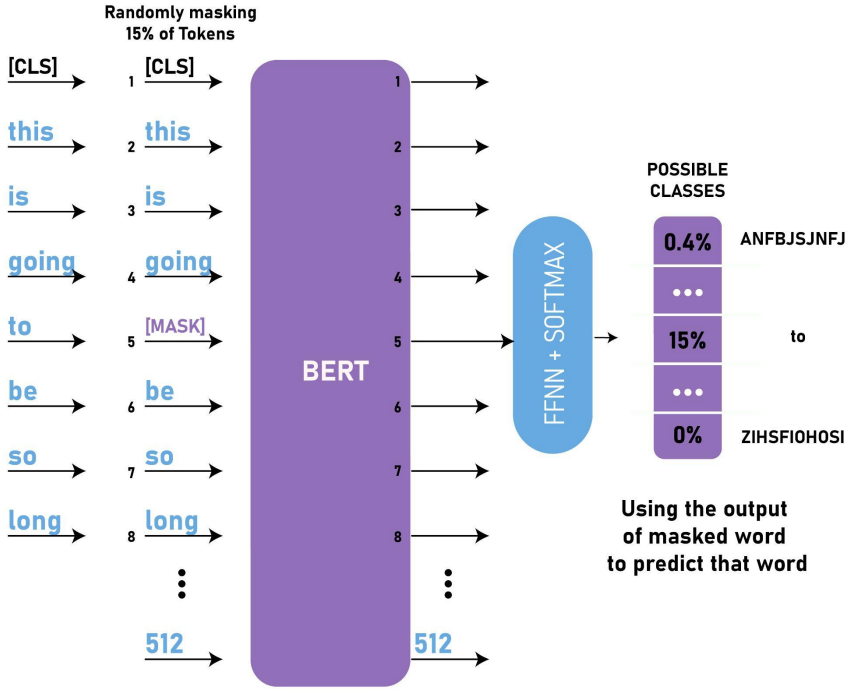


Improving Language Understanding by Generative Pre-Training (Radford et al., 2018), https://openai.com/research/language-unsupervised

Language Models are Unsupervised Multitask Learners (Radford et al., 2019), https://openai.com/research/better-language-models

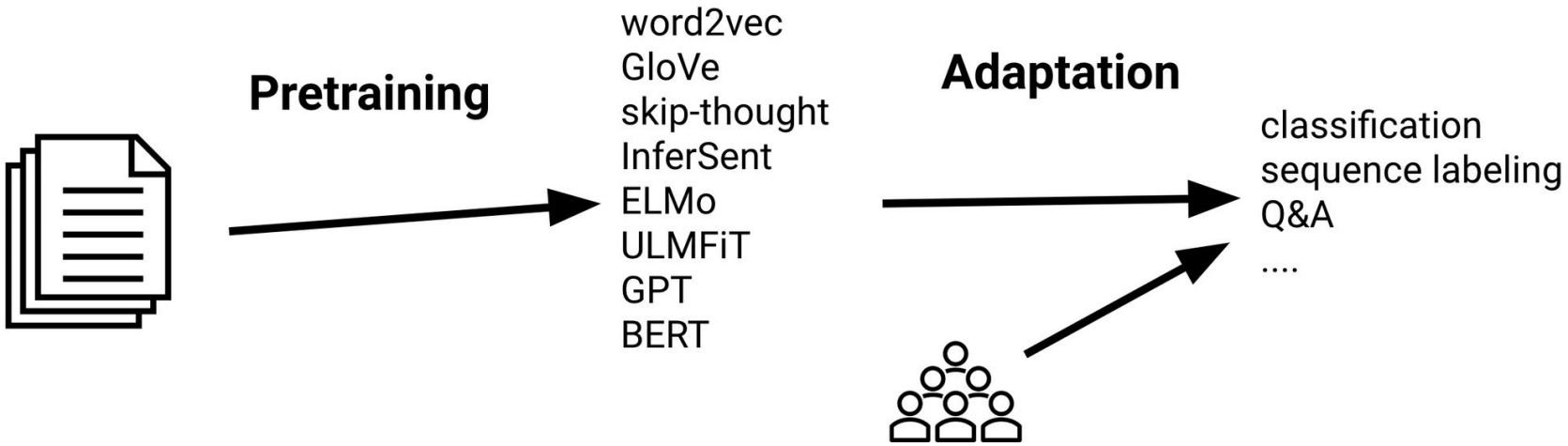
BERT (Devlin et al., 2019)

- Encoder Transformer model (12/24 layers)
- Trained on the **Masked Language Modeling** (**MLM**)
- The model can be further trained (fine-tuning) for solving different NLP tasks:
  - Sentiment analysis;
  - Question answering;
  - Textual entailment;
  - etc.



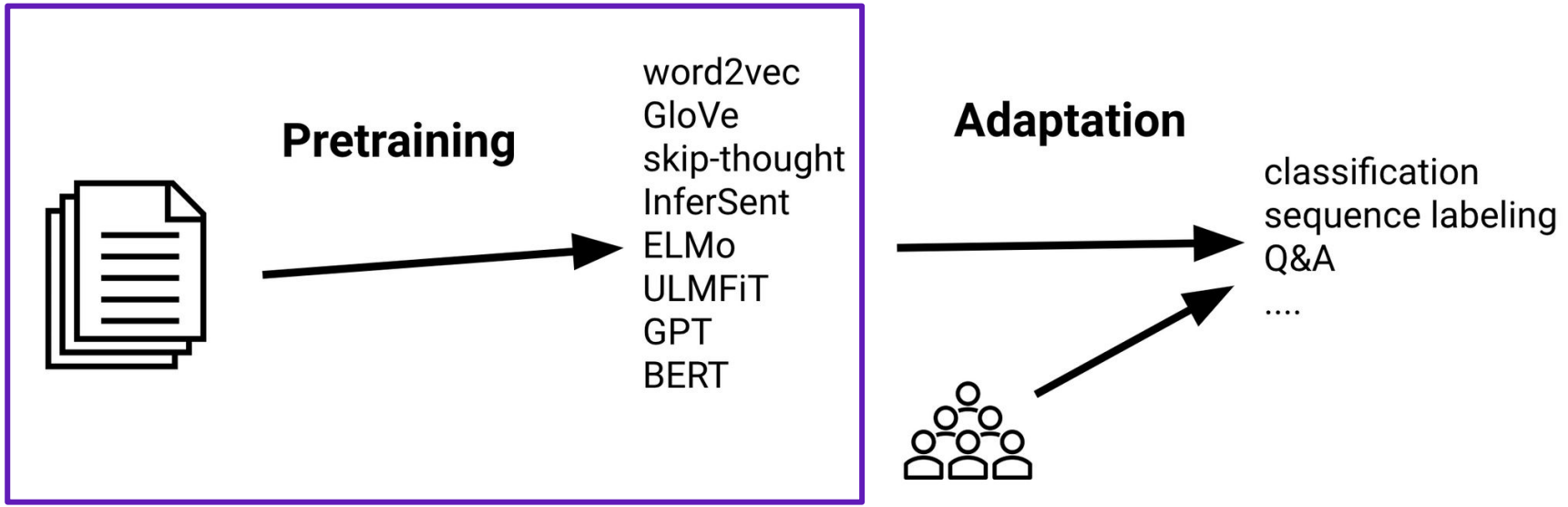
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2019), https://aclanthology.org/N19-1423.pdf

Transfer Learning



The State of Transfer Learning in NLP: https://ruder.io/state-of-transfer-learning-in-nlp/

Transfer Learning



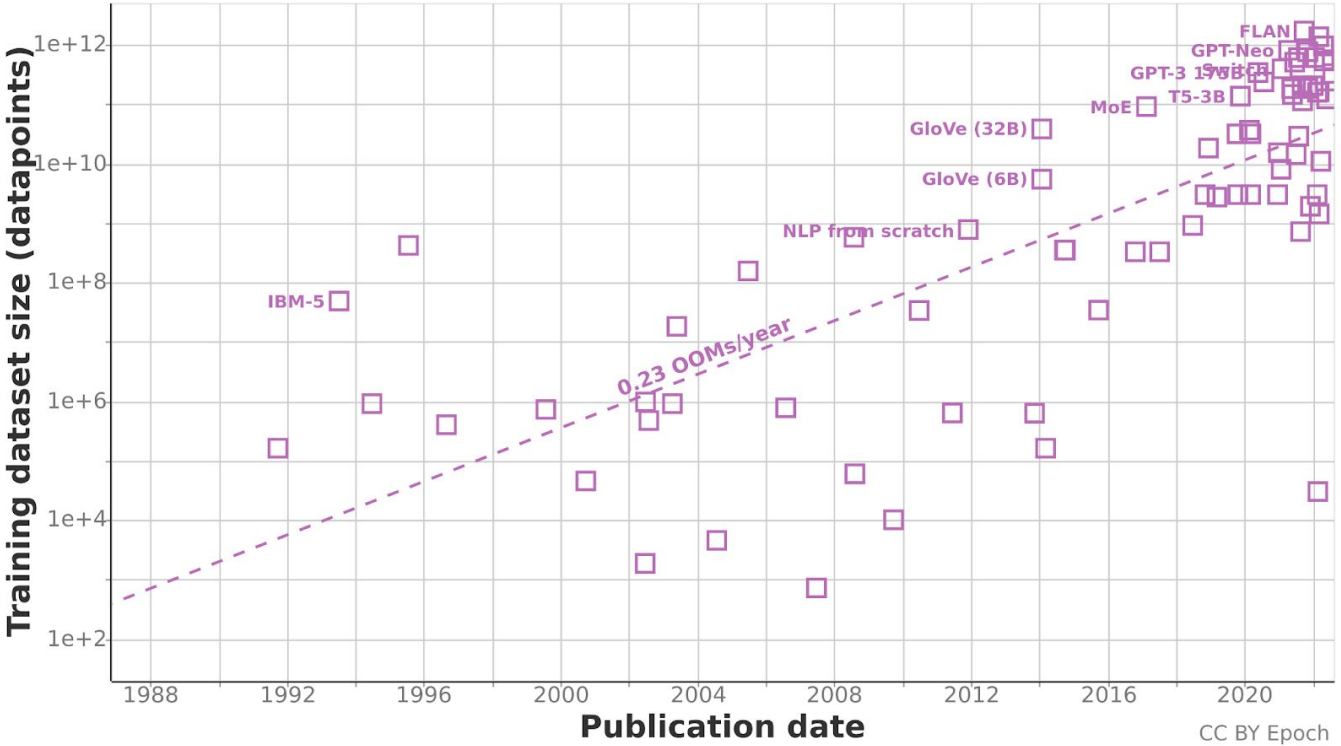
The State of Transfer Learning in NLP: https://ruder.io/state-of-transfer-learning-in-nlp/

Pre-training

- During the "*Pre-training*" phase, the model is trained in an unsupervised manner (e.g. LM, MLM) on a huge collection of raw text

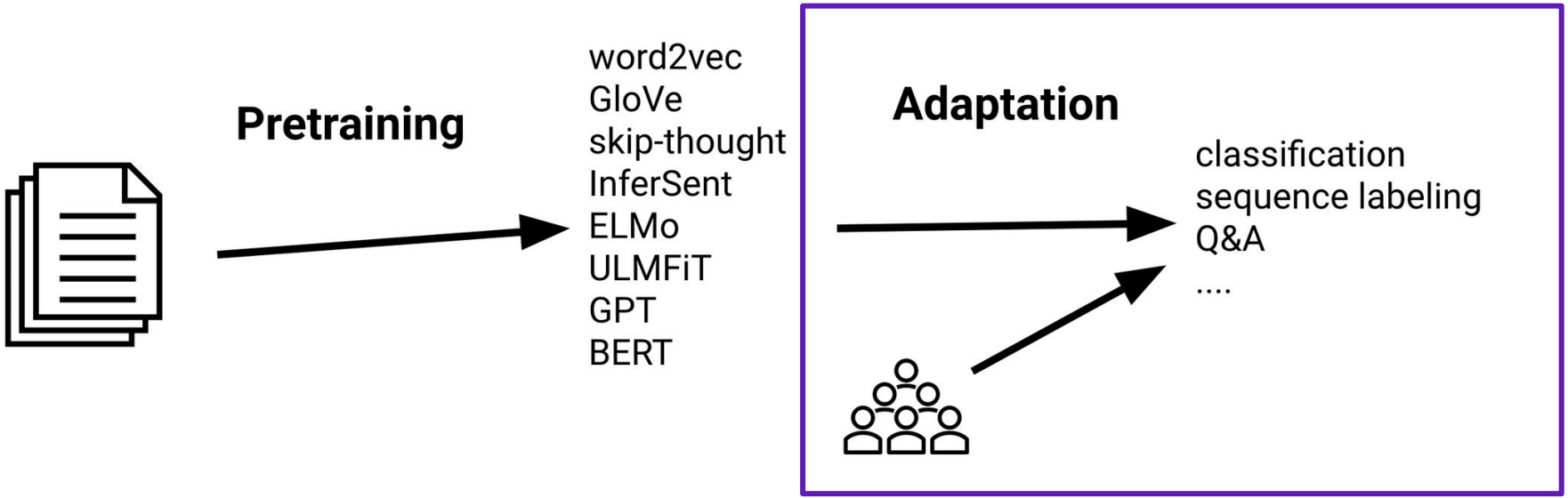
- Some exaomples:
  - **BERT training**: BookCorpus (800M words) + English Wikipedia (2500M words)
  - **GPT-3 training**: CommonCrawl + WebText2 + Books1 + Books2 + Wikipedia (around 500B words)

Pre-training



Source: https://www.lesswrong.com/posts/asqDCb9XzXnLjSfgL/trends-in-training-dataset-sizes

Transfer Learning



The State of Transfer Learning in NLP: https://ruder.io/state-of-transfer-learning-in-nlp/

Prompting → Large Language Models (LLMs)

- In recent years, the development of NLMs has shifted towards the creation of generative models:
  - Main goal: framing any task (e.g., classification, translation, question answering, etc.) as a **generation task**

Prompting → Large Language Models (LLMs)

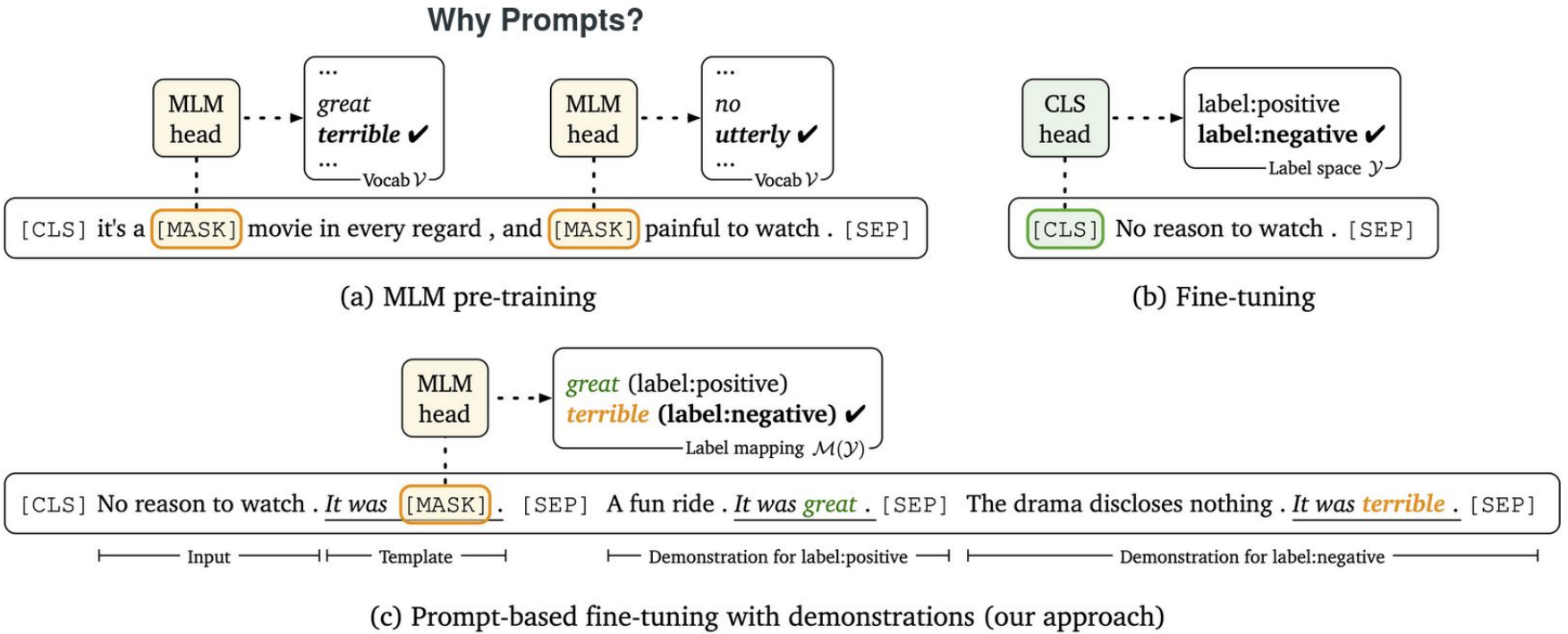
- In recent years, the development of NLMs has shifted towards the creation of generative models:
  - Main goal: framing any task (e.g., classification, translation, question answering, etc.) as a **generation task**

Prompting

"A prompt is a piece of text inserted in the input examples, so that the original task can be formulated as a (masked) language modeling problem."

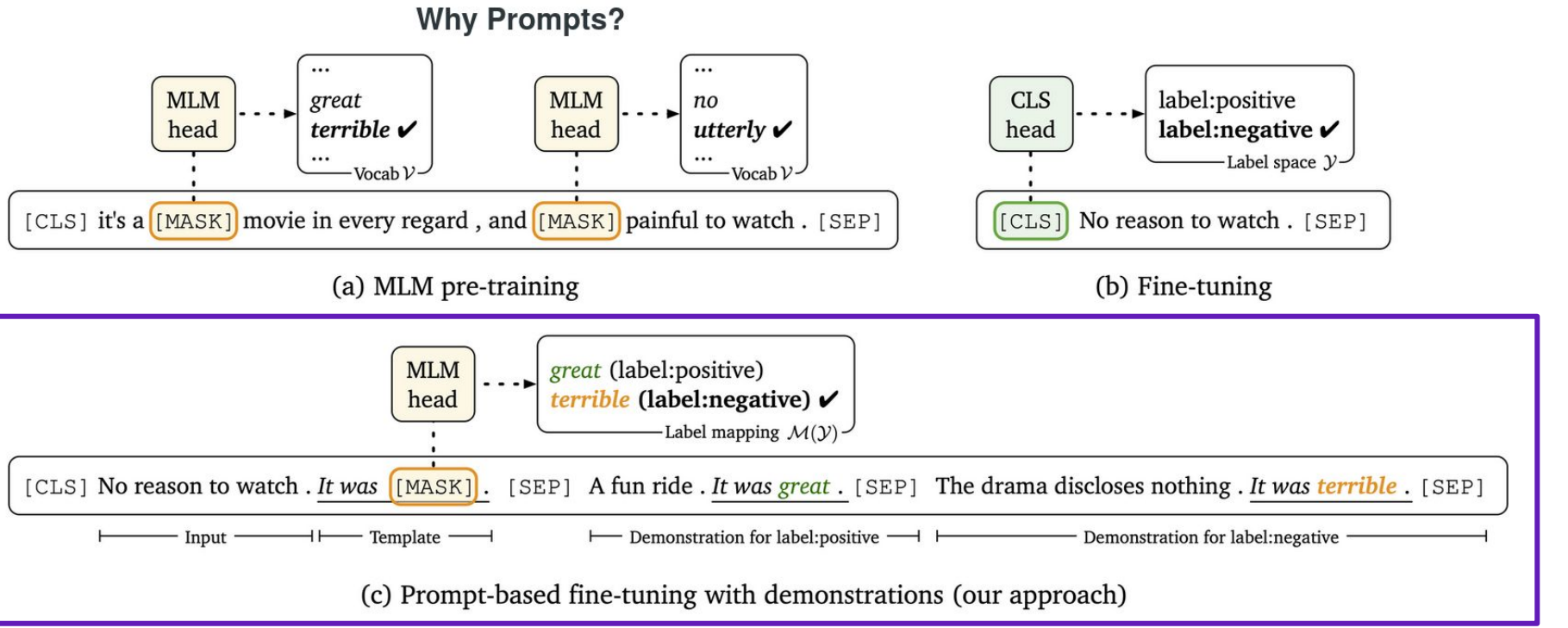
(Prompting: Better Ways of Using Language Models for NLP Tasks, The Gradient)

Prompting → Large Language Models (LLMs)



Source: [Prompting: Better Ways of Using Language Models for NLP Tasks, The Gradient]

Prompting → Large Language Models (LLMs)



Source: [Prompting: Better Ways of Using Language Models for NLP Tasks, The Gradient]

Instruction Tuning e RLHF: from GPT-3 to InstructGPT

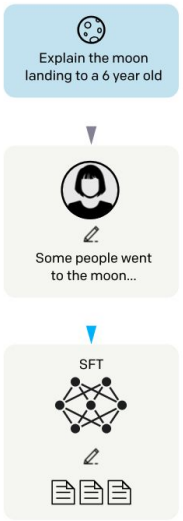
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



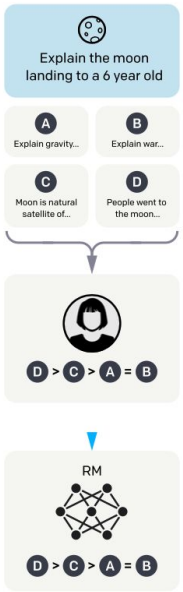
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

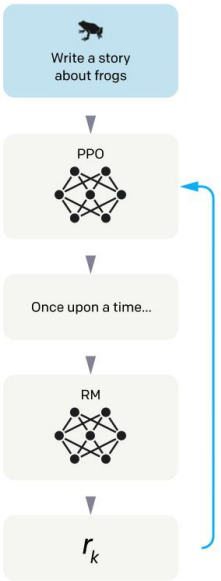
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

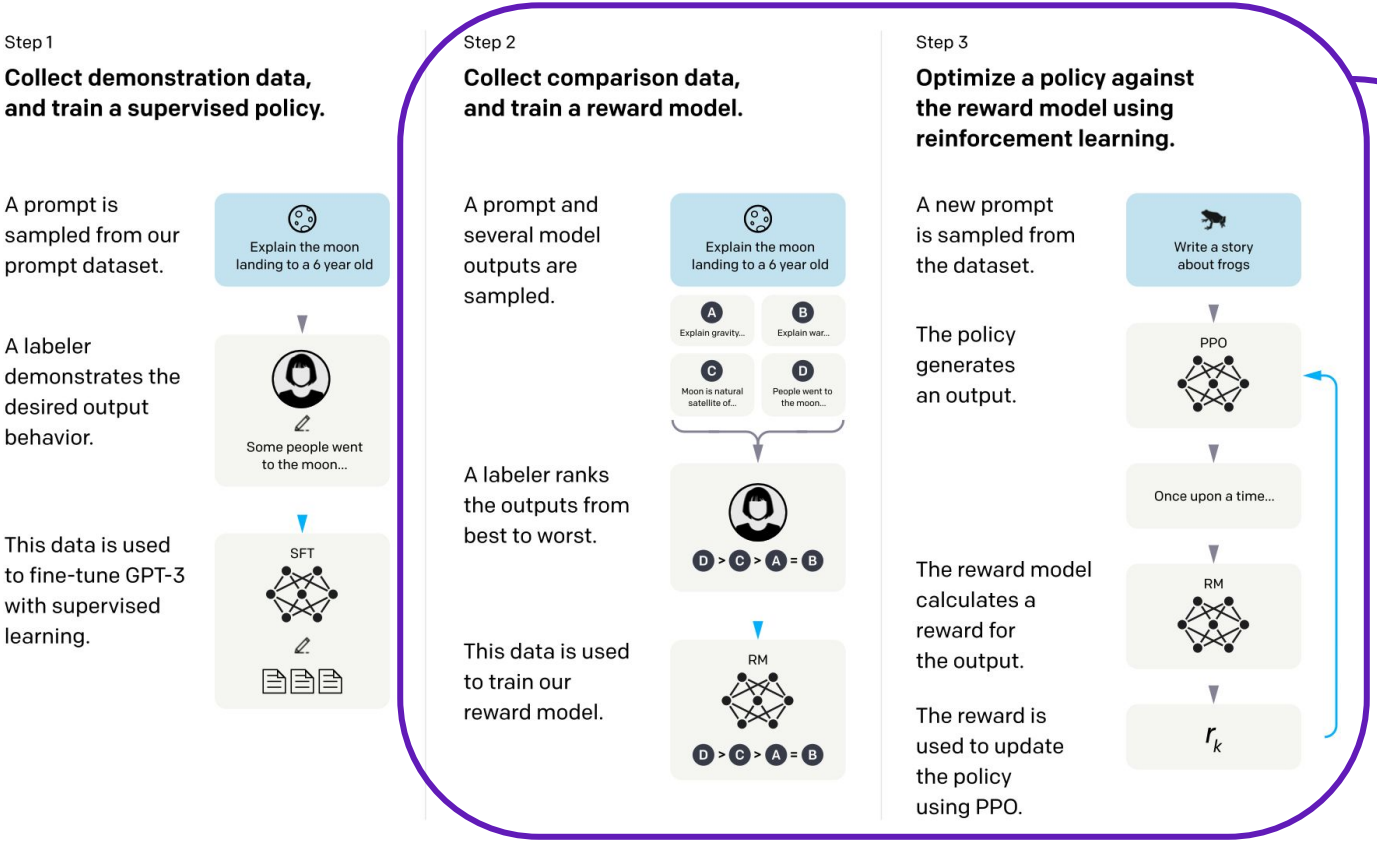
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



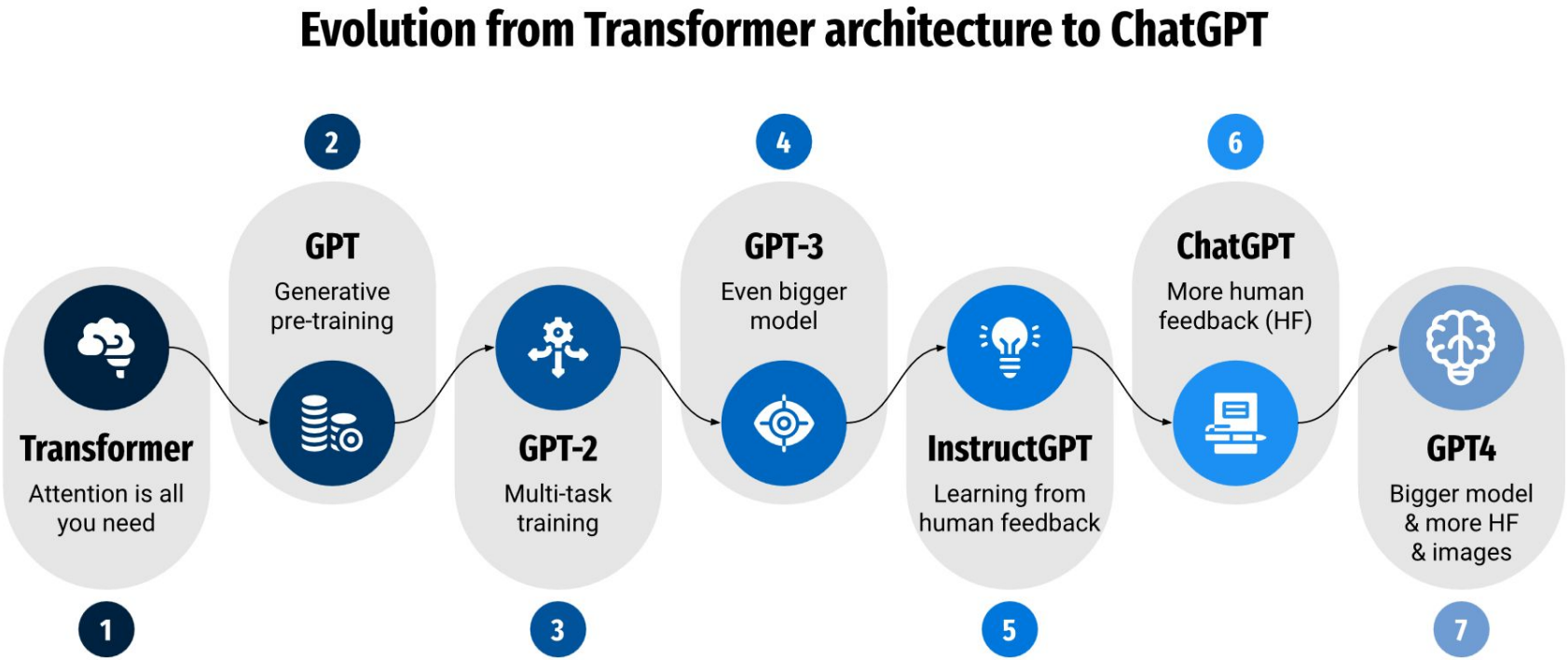
Instruction Tuning e RLHF: from GPT-3 to InstructGPT



Reinforcement Learning from Human Feedback (RLHF)

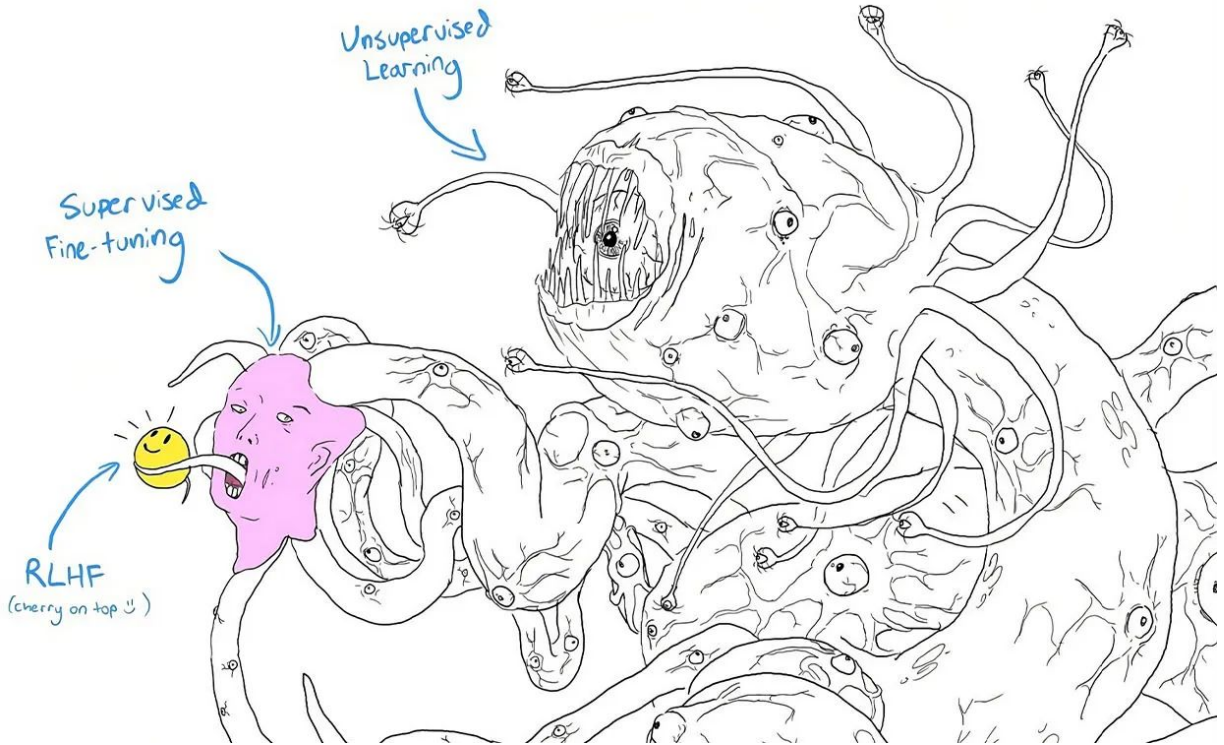
https://huggingface.co/blog/rlhf

From Transformer to GPT4



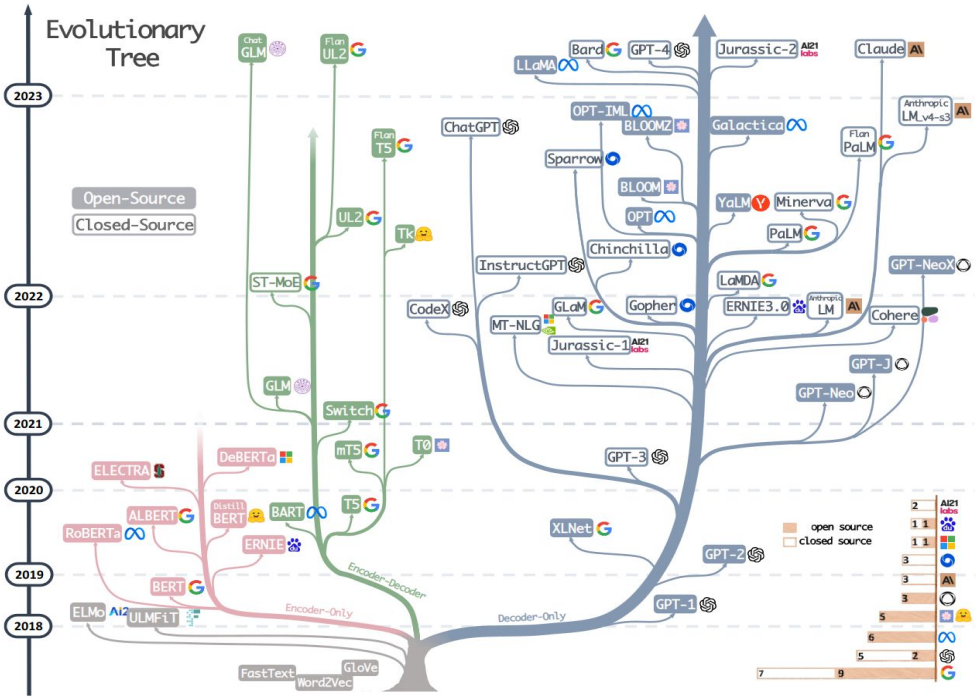
ChatGPT: Jack of all trades, master of none (Kocoń et al., 2023), https://www.sciencedirect.com/science/article/pii/S156625352300177X

From Transformer to GPT4



From: https://medium.com/@mataciunasdeividas/the-simple-explanation-of-chatgpt-llm-rlhf-using-shoggoth-with-smiley-face-meme-947a0e9fb441

“Evolutionary Tree”

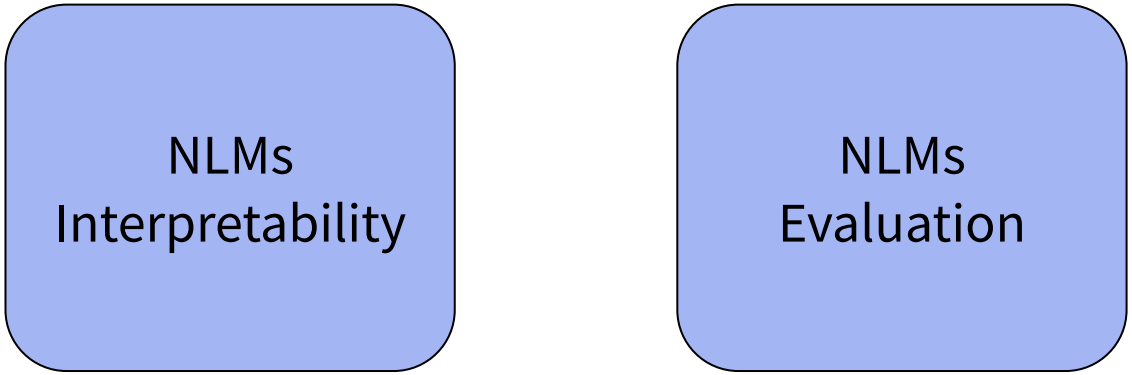


Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond (Yang et al., 2024), https://dl.acm.org/doi/10.1145/3649506

Interpreting and Evaluating NLMs

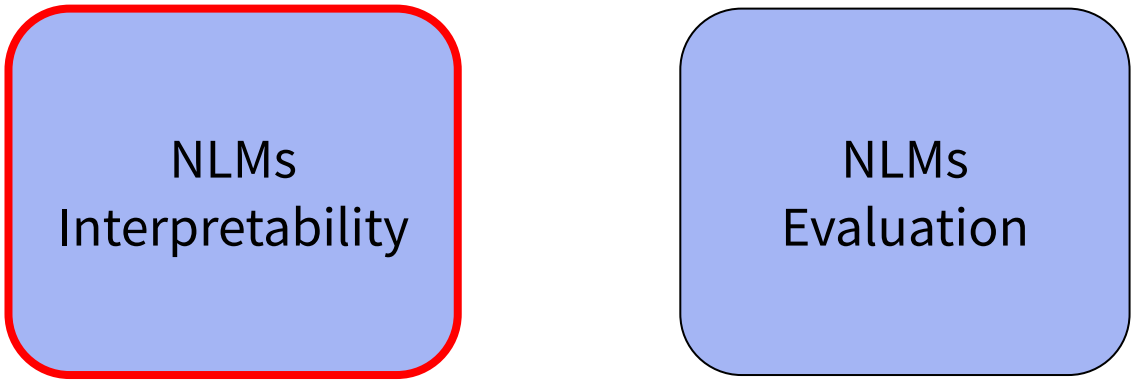
Interpreting and Evaluating NLMs

- 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



The Case for Interpretability

- The development of powerful state-of-the-art NLMs comes at the cost of **interpretability**, since complex NN models offer little transparency about their inner workings and their abilities

Objectives:

- **Understand the nature of AI systems** → be faithful to what influences the AI decisional process
- **Empower AI system users** → derive actionable useful insights from AI choices

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."

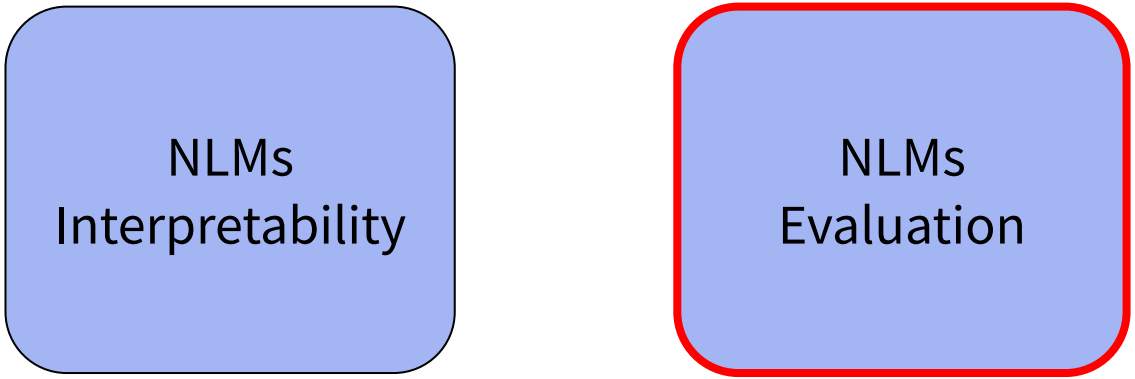
Belinkov and Glass, Analysis Methods in Neural Language Processing: A Survey (2019). In Transactions of ACL, Volume 7, pages 49-72.

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?

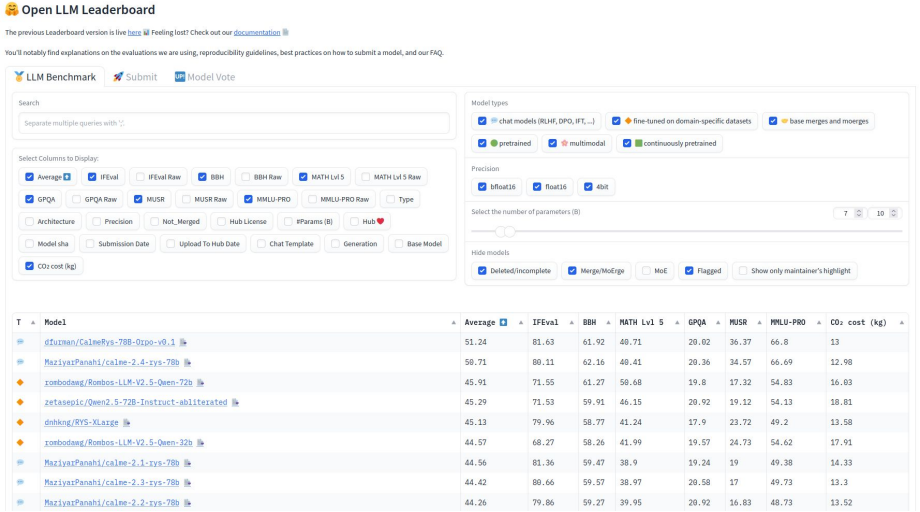
Interpreting and Evaluating NLMs

- 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

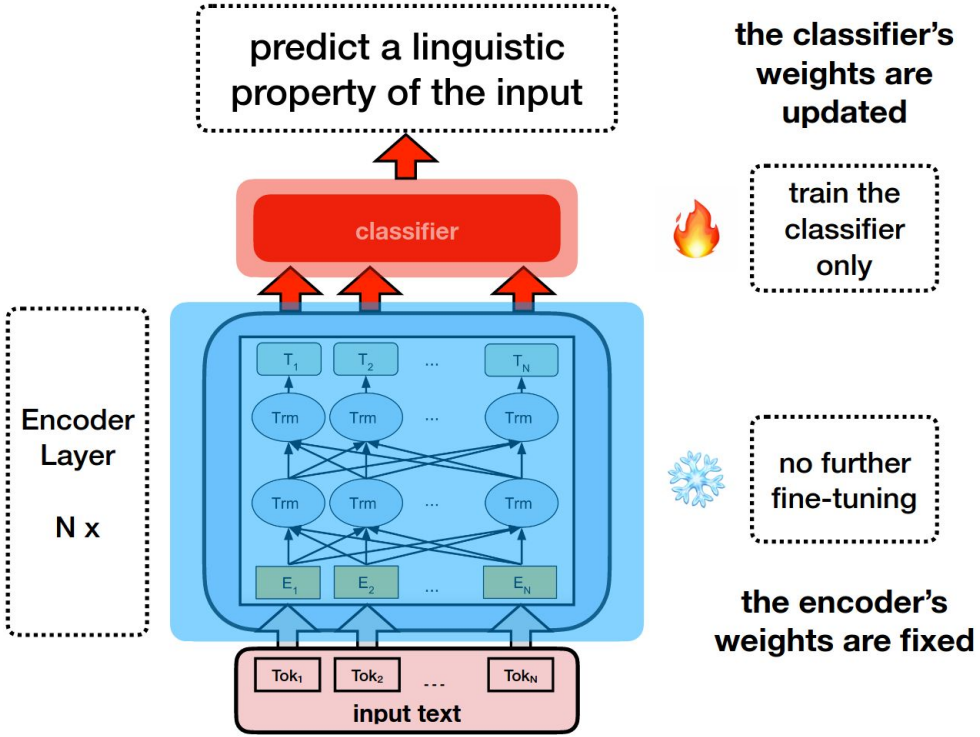
- The "*linguistic profiling*" methodology (van Halteren, 2004) assumes that wide counts of linguistic features are particularly helpful in the resolution of several NLP tasks, e.g.:
  - Text Profiling (e.g. text readability, textual genres)
  - Author Profiling (e.g. author's age and native language)

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

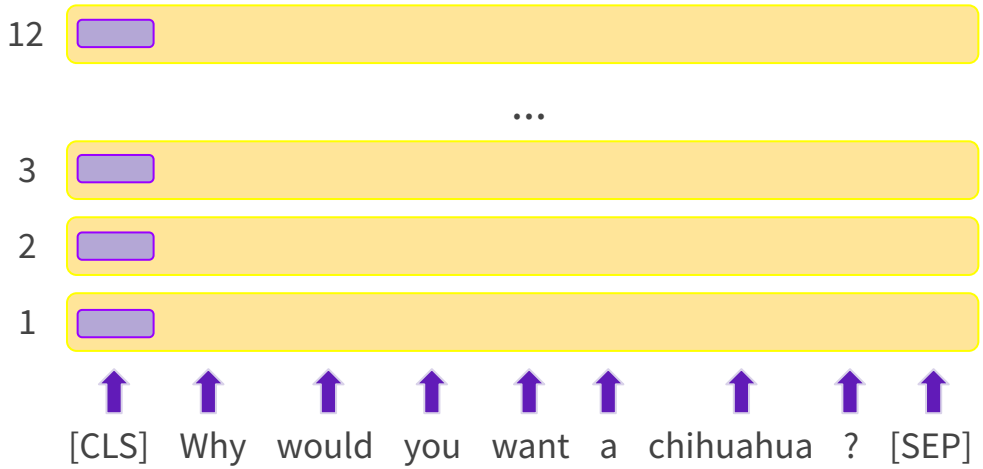
- ProfilingUD (Brunato et al., 2020) is a web–based application that performs linguistic profiling of a text, or a large collection of texts, for multiple languages

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

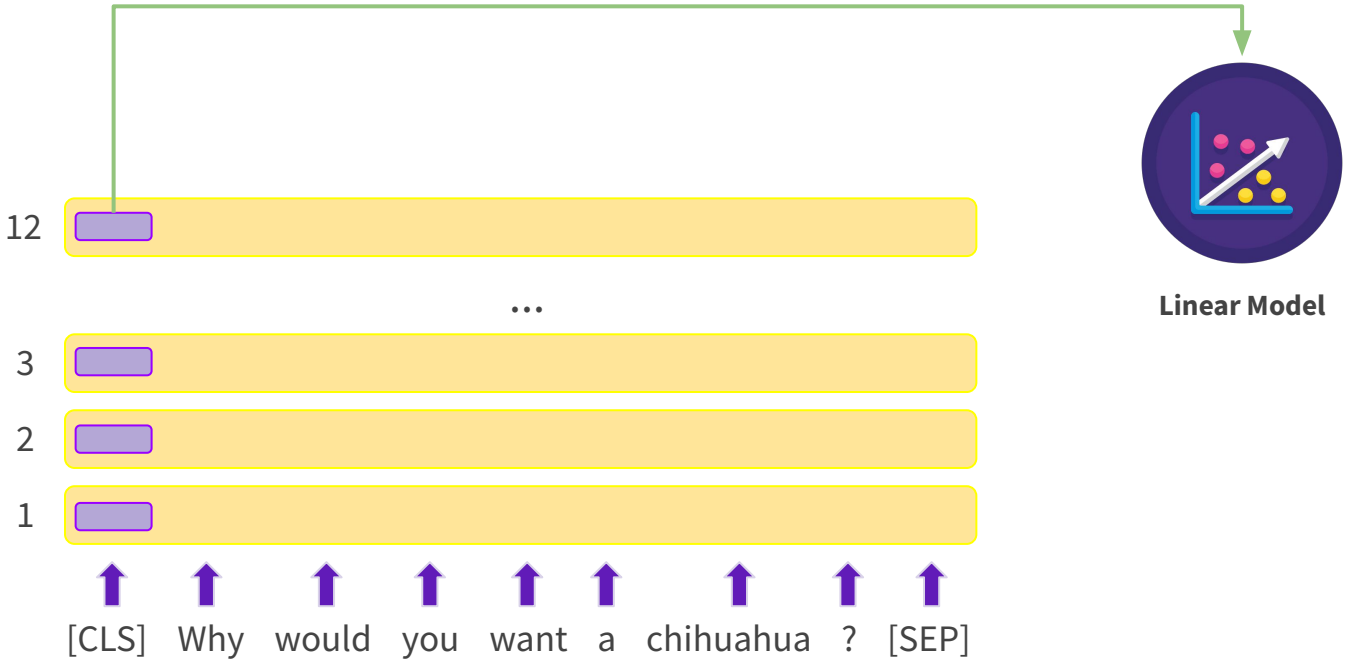
- Link: http://linguistic-profiling.italianlp.it/

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

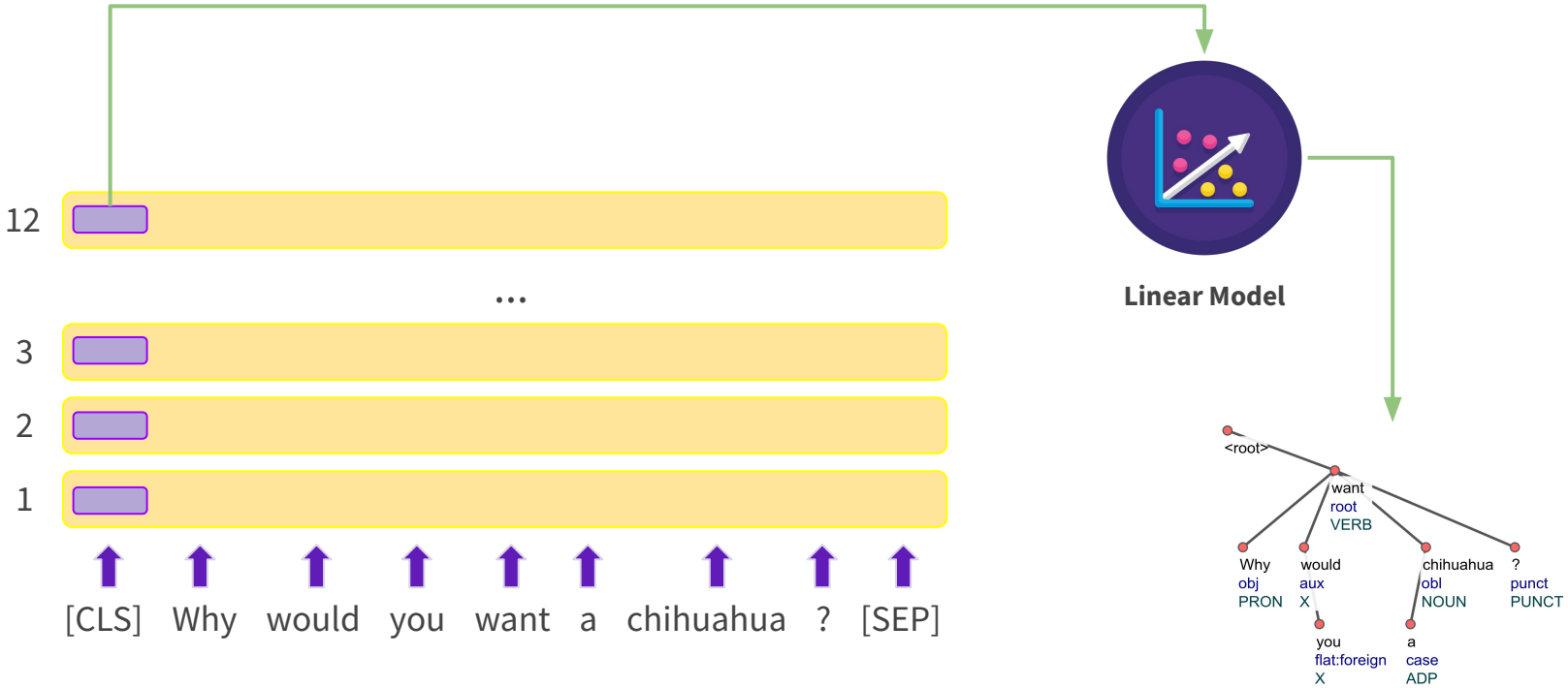
Profiling Neural Language Models



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

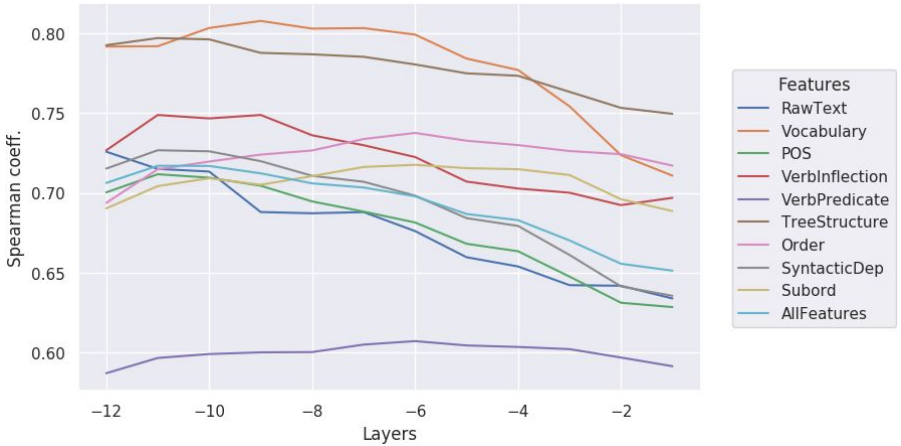
Research questions:

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

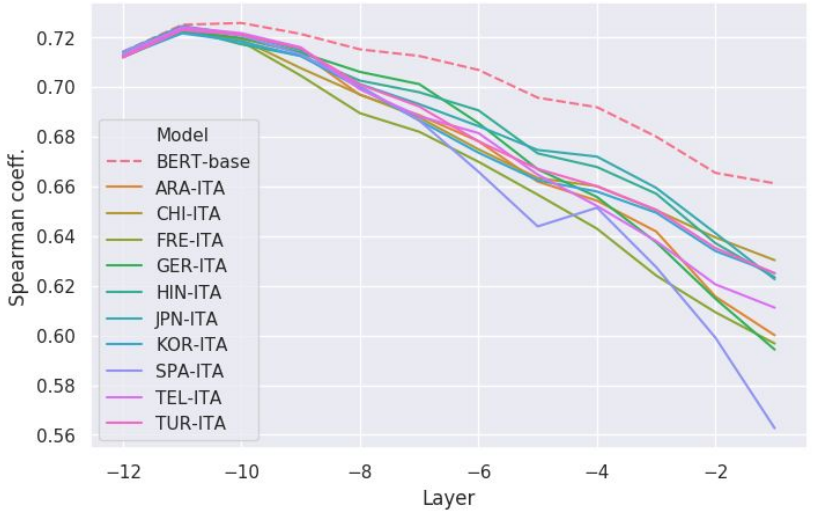
2. How this knowledge is modified after a fine-tuning process?
    a. Fine-tuning on the Natural Language Identification Task

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

Pre fine-tuning:



Post fine-tuning:



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?

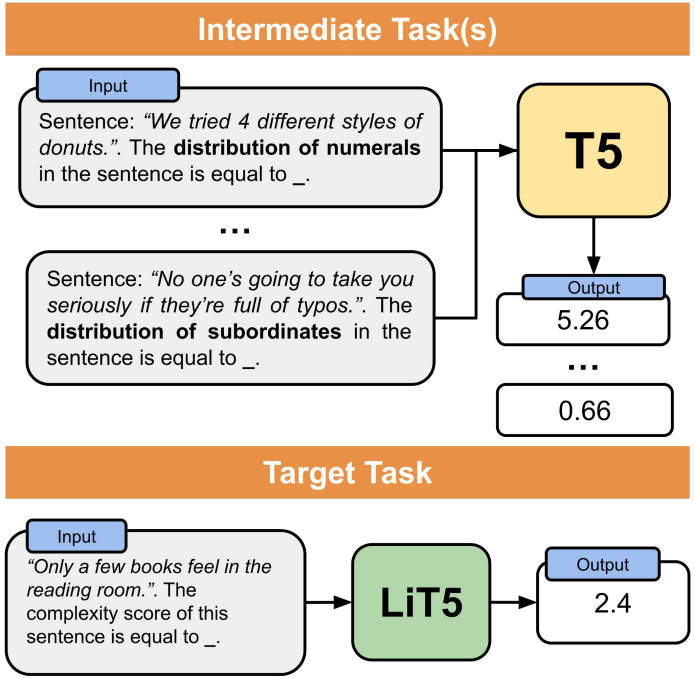
Miaschi A., Dell’Orletta F., Venturi G. (2024). Linguistic Knowledge Can Enhance Encoder-Decoder Models (*If You Let It*). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024, Turin)

Our Approach

- Two-step approach:
  - Fine-tune the T5 models on several intermediate tasks
    - Multi- and single-task fine-tuning
  - Fine-tune the Linguistically-Informed (LI) models on the target task

- We saved checkpoints every 5 epochs, in order to monitor the impact of the approach at increasing snapshots of the models

- We tested the approach both in Italian and English and in a cross-lingual scenario

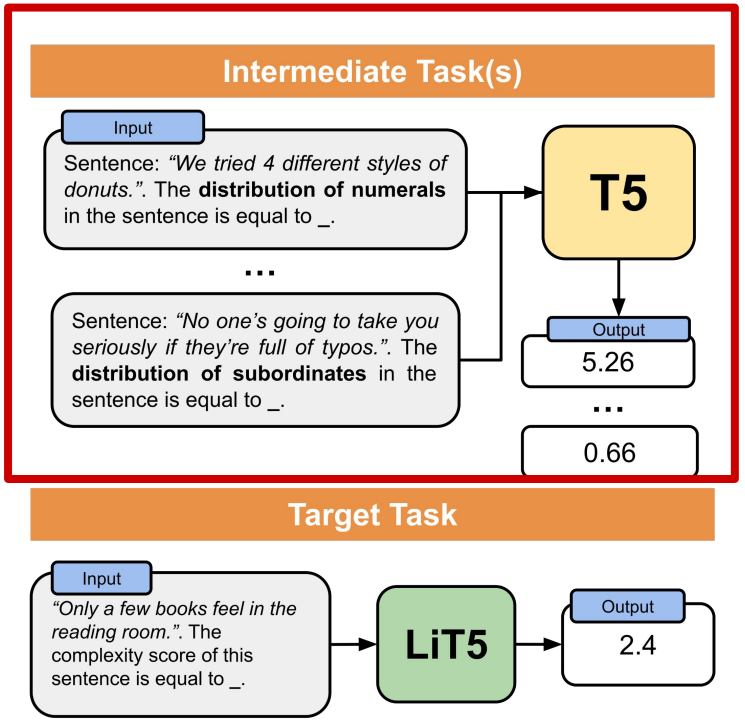


Our Approach

- Two-step approach:
  - **Fine-tune the T5 models on several intermediate tasks**
    - **Multi- and single-task fine-tuning**
  - Fine-tune the Linguistically-Informed (LI) models on the target task

- We saved checkpoints every 5 epochs, in order to monitor the impact of the approach at increasing snapshots of the models

- We tested the approach both in Italian and English and in a cross-lingual scenario

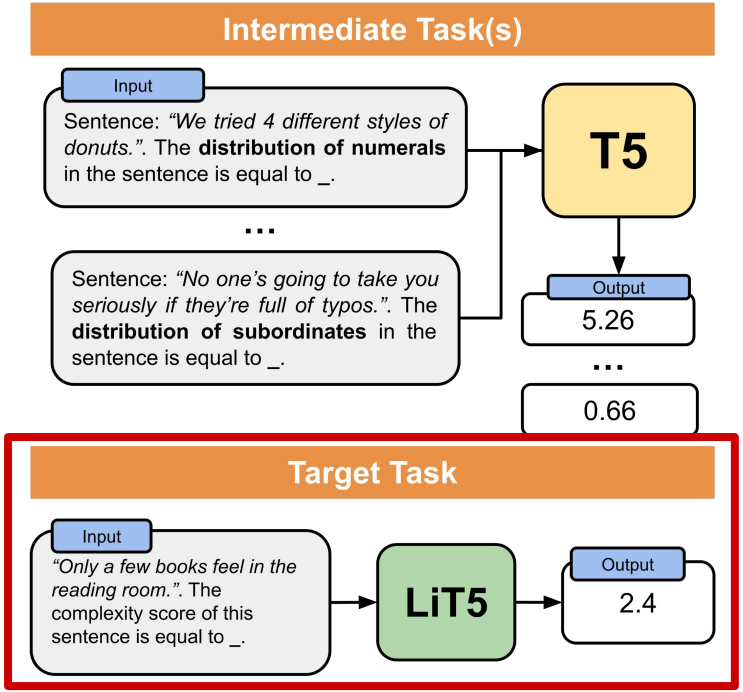


Our Approach

- Two-step approach:
  - Fine-tune the T5 models on several intermediate tasks
    - Multi- and single-task fine-tuning
  - **Fine-tune the Linguistically-Informed (LI) models on the target task**

- We saved checkpoints every 5 epochs, in order to monitor the impact of the approach at increasing snapshots of the models

- We tested the approach both in Italian and English and in a cross-lingual scenario

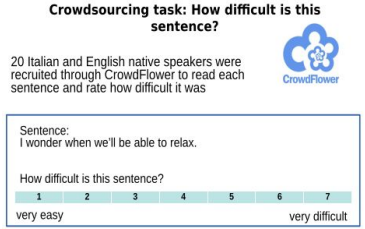


Data

- Intermediate tasks:
  - 10 morpho- and syntactic characteristics of a sentence
    - selected on the degree of correlation between sentence-level complexity judgments and their values

Profiling-UD: extraction of feature values from ITA e ENG UD treebank

Features Italian	Corr	Features English	Corr
char_per_tok	0.28	upos_dist_NUM	0.35
upos_dist_ADJ	0.21	dep_dist_nummod	0.31
upos_dist_NUM	0.19	upos_dist_SYM	0.27
lexical_density	0.17	upos_dist_AUX	0.25
dep_dist_aux	0.17	dep_dist_compound	0.25
dep_dist_mark	0.16	upos_dist_PRON	0.24
aux_mood_dist_Ind	0.14	upos_dist_DET	0.23
obj_post	0.14	subord_prop_dist	0.17
upos_dist_PUNCT	0.13	aux_form_dist_Fin	0.16
subord_prop_dist	0.12	aux_mood_dist_Ind	0.14



  - Prediction of their distribution in the Italian and English versions of the Universal Dependency Treebanks extracted with Profiling-UD

- Target task:
  - corpus of 1,440 Italian and 2,400 English sentences manually rated by 20 crowdsourced workers for the level of perceived complexity on 1-7 Likert scale (Brunato et al., EMNLP 2018)

Models and Evaluation

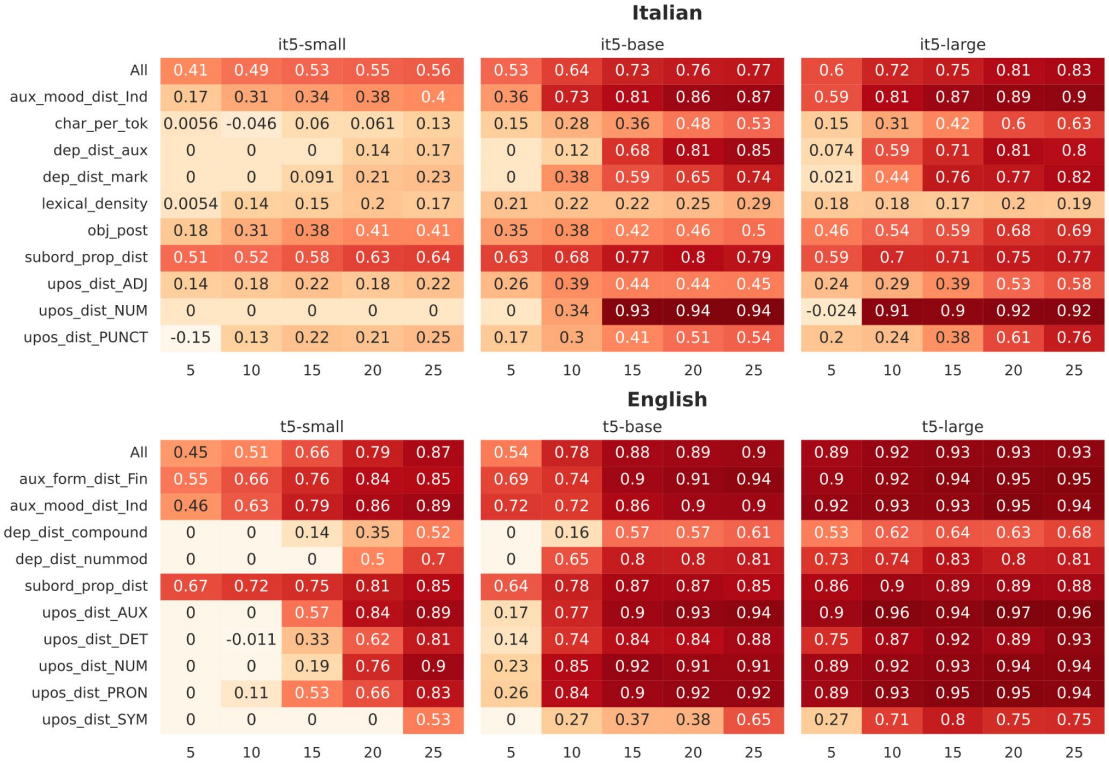
Models:

Language	Model	Parameters
English	t5-small	60M
	t5-base	220M
	t5-large	770M
Italian	it5-small	60M
	it5-base	220M
	it5-large	738M

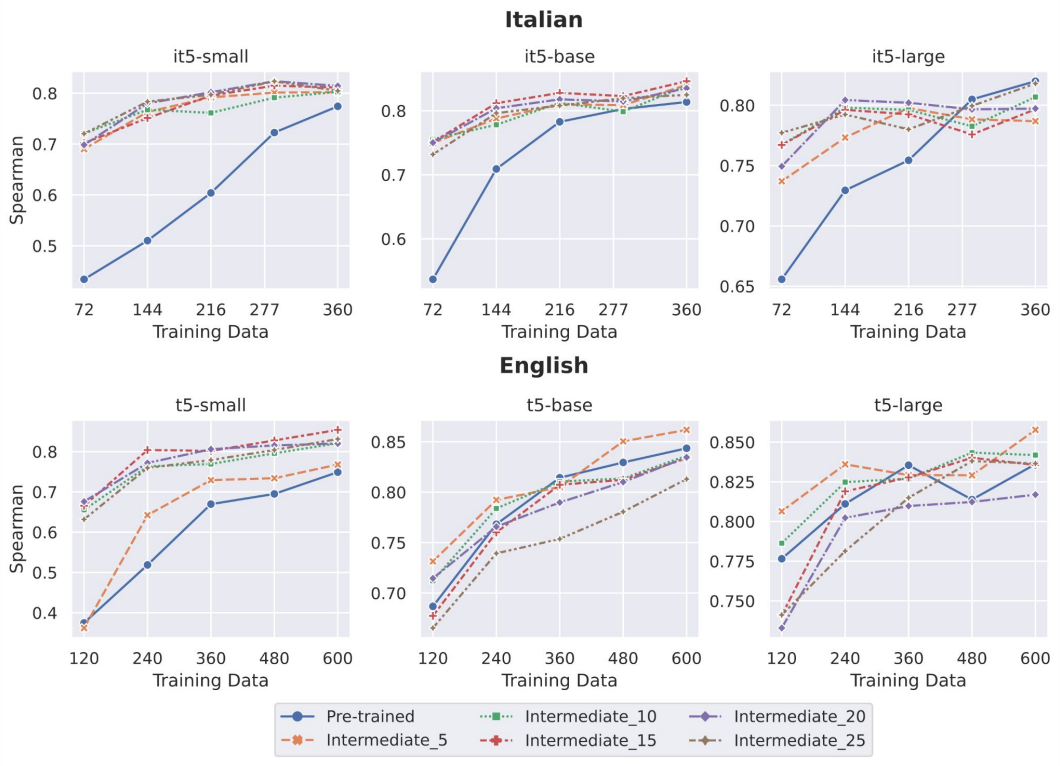
Evaluation:

- We used Spearman correlation score as evaluation metric:
  - **Intermediate tasks**: Correlation between the gold value of each feature in the Italian or English treebank and the predicted value of the models for the intermediate tasks.
  - **Target task**: Correlation between average judgments of complexity and the complexity scores obtained with the fine-tuned LiT5 models.

Enhancing T5 with Linguistic Features



Predicting Complexity with LI 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

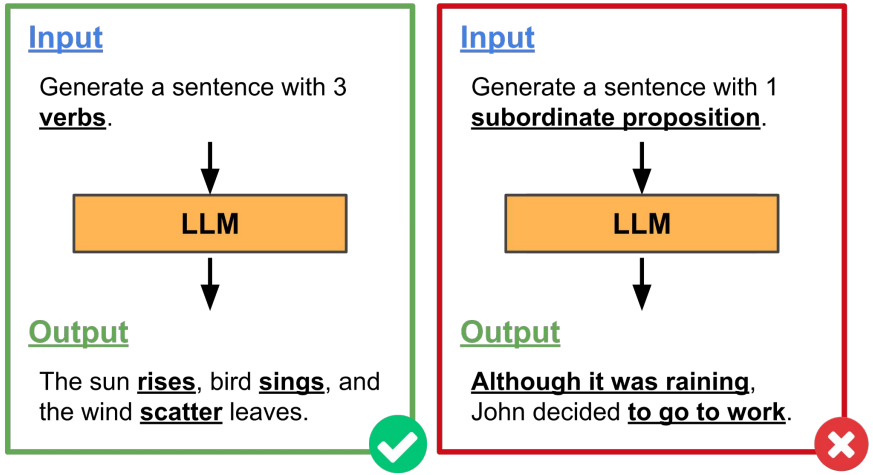
- Motivations:
  - Large Language Models (LLMs) demonstrated remarkable capabilities in solving multiple tasks and in generating coherent and contextually relevant texts
  - Such capabilities have been extensively evaluated against several benchmarks, as evidenced by the success of platforms such as 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:

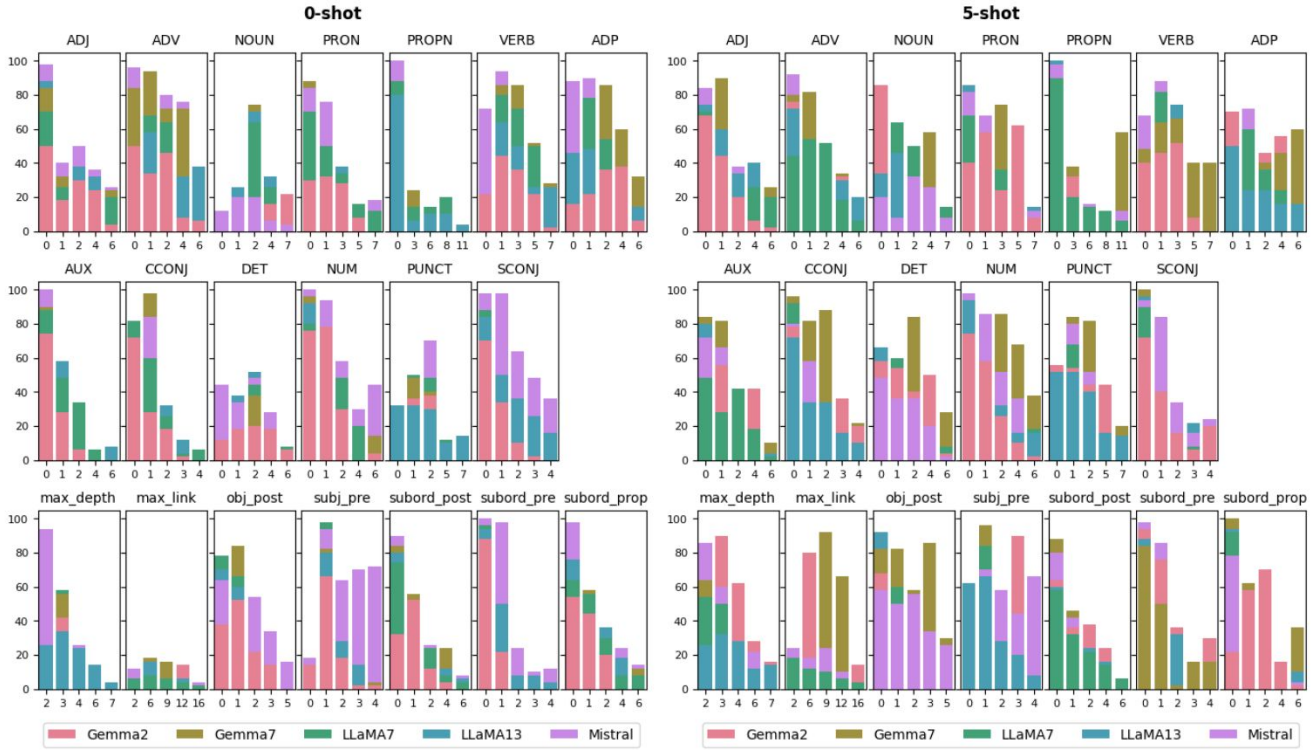
- We used two different metrics:
  - **Success Rate (SR)**: fraction of times the model generated a sentence whose property value exactly corresponds to the one provided.
  - **Spearman coefficient**: correlation coefficients between the increasing property values extracted from EWT and those extracted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

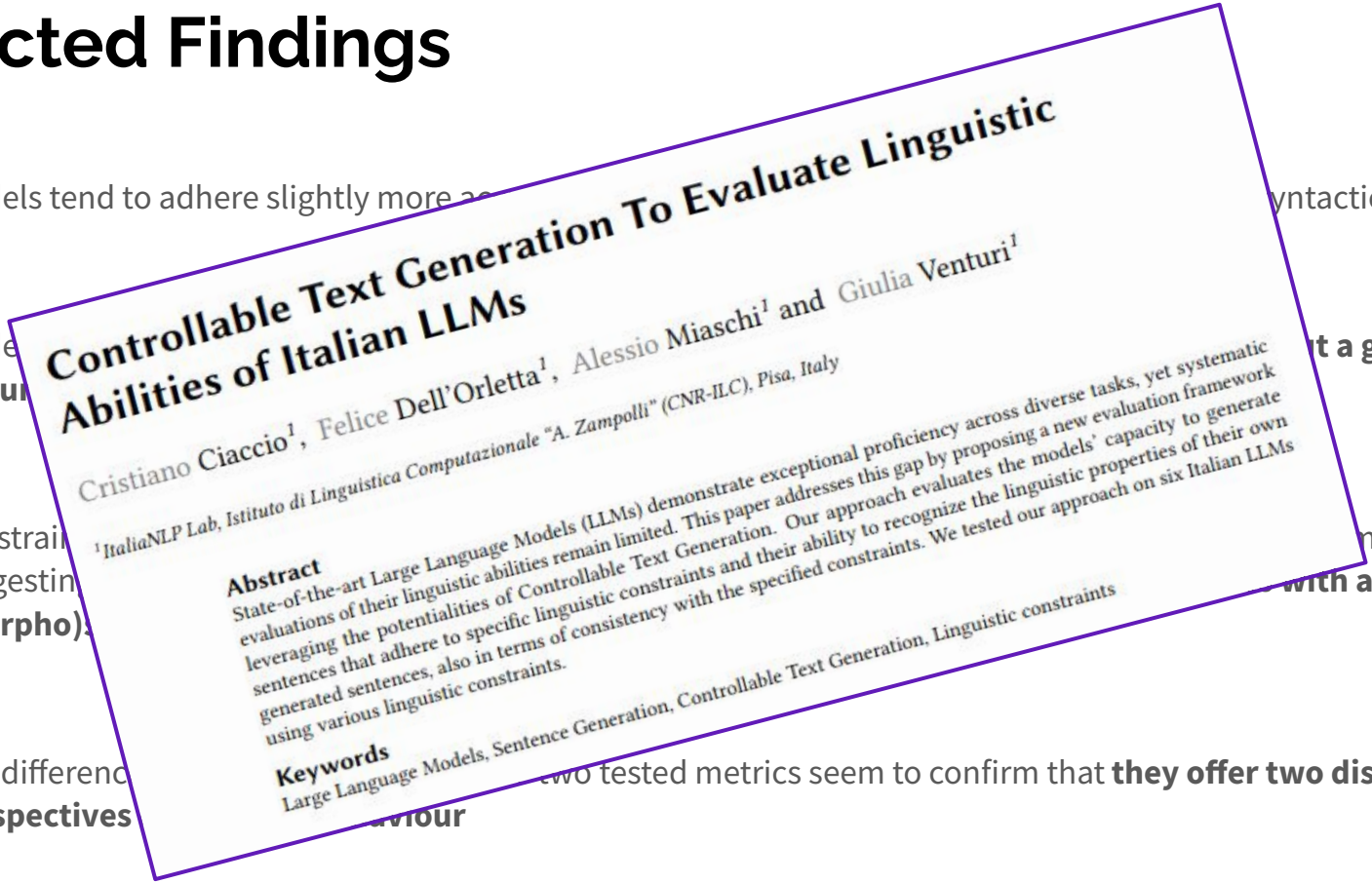
- Models are capable of distinguishing when they are asked to generate a sentence **with or without a given feature**

- 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**

Selected Findings

- Models tend to adhere slightly more ac… yntactic ones
- Mode… **t a given featu…**
- Constrai… ment, suggestin… **with a varied (morpho)…**
- The differenc… two tested metrics seem to confirm that **they offer two distinct perspectives … aviour**



Controllable Text Generation To Evaluate Linguistic Abilities of Italian LLMs

Cristiano Ciaccio[1], Felice Dell'Orletta[1], Alessio Miaschi[1] and Giulia Venturi[1]

[1]ItaliaNLP Lab, Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), Pisa, Italy

Abstract

State-of-the-art Large Language Models (LLMs) demonstrate exceptional proficiency across diverse tasks, yet systematic evaluations of their linguistic abilities remain limited. This paper addresses this gap by proposing a new evaluation framework leveraging the potentialities of Controllable Text Generation. Our approach evaluates the models' capacity to generate sentences that adhere to specific linguistic constraints and their ability to recognize the linguistic properties of their own generated sentences, also in terms of consistency with the specified constraints. We tested our approach on six Italian LLMs using various linguistic constraints.

Keywords

Large Language Models, Sentence Generation, Controllable Text Generation, Linguistic constraints

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

Ciaccio C., Miaschi A., Dell'Orletta F. (2024). Evaluating Lexical Proficiency in Neural Language Models (*submitted*).

Evaluating Lexical Proficiency in Neural Language Models

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

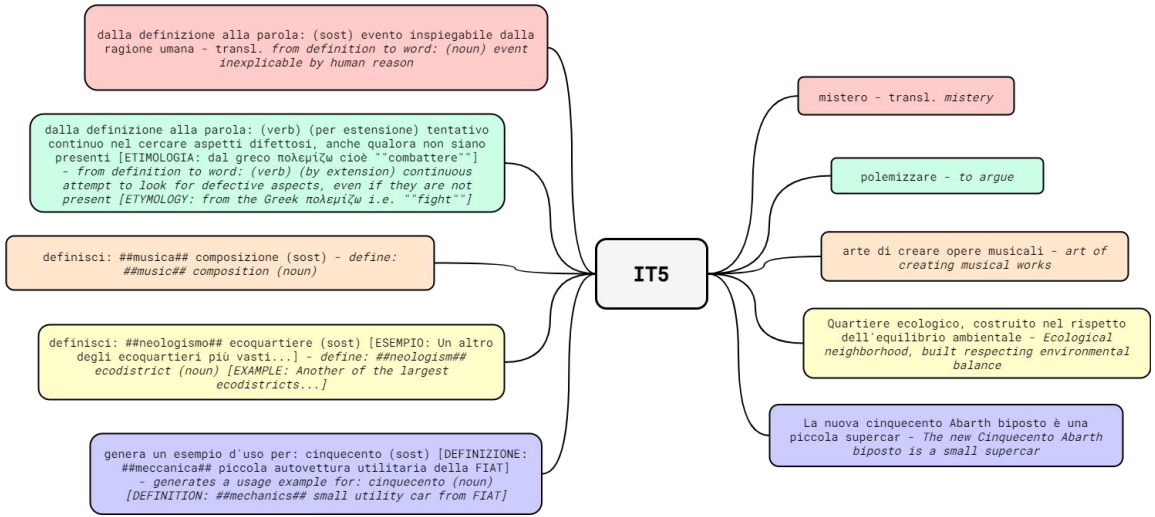
- Almost no study that goes beyond commonly lexicalized words

- 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. (2024). Evaluating Lexical Proficiency in Neural Language Models (*submitted*).

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.

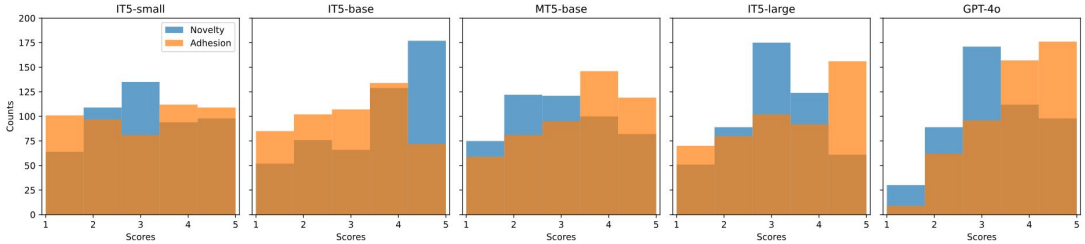


Figure 1: Distribution of novelty and adhesion human scores across the 5 values of the Likert scale for all models.

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

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!

Future Directions:

- Studying and evaluating generalization of LLMs across different scenarios, domains and languages (Hupkes et al., 2023)
- Testing models' behaviour and performance on complex and "creative task": *"The philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative"* (Colton and Wiggins, 2012) → creativity as a step towards Artificial General Intelligence (AGI) [Computational Creativity, Tim Van de Cruys]
- Mechanistic Interpretability (Elhage et al, 2021; Olsson et al., 2022)

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

References

- Bengio, Yoshua, et al. (2003). "A neural probabilistic language model." *The journal of machine learning research* 3, pages 1137-1155
- Vaswani, Ashish, et al. (2017). "Attention is all you need." *Advances in Neural Information Processing Systems* (NEURIPS)
- Radford, Alec. "Improving language understanding by generative pre-training." (2018)
- Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.
- Devlin, Jacob, et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*
- 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)
- Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a Tool for Linguistic Profiling of Texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association
- 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)
- 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)
- Ciaccio C., Dell'Orletta F., Miaschi A., Venturi G. (2024). Controllable Text Generation To Evaluate Linguistic Abilities of Italian LLMs. In *Proceedings of the Tenth Italian Conference on Computational Linguistics* (CLiC-it 2024, Pisa)
- Hupkes, Dieuwke, et al. "A taxonomy and review of generalization research in NLP." *Nature Machine Intelligence* 5.10 (2023): 1161-1174
- Elhage, Nelson, et al. "A mathematical framework for transformer circuits." *Transformer Circuits Thread* 1.1 (2021): 12
- Olsson, Catherine, et al. "In-context learning and induction heads." *arXiv preprint arXiv:2209.11895* (2022)