Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

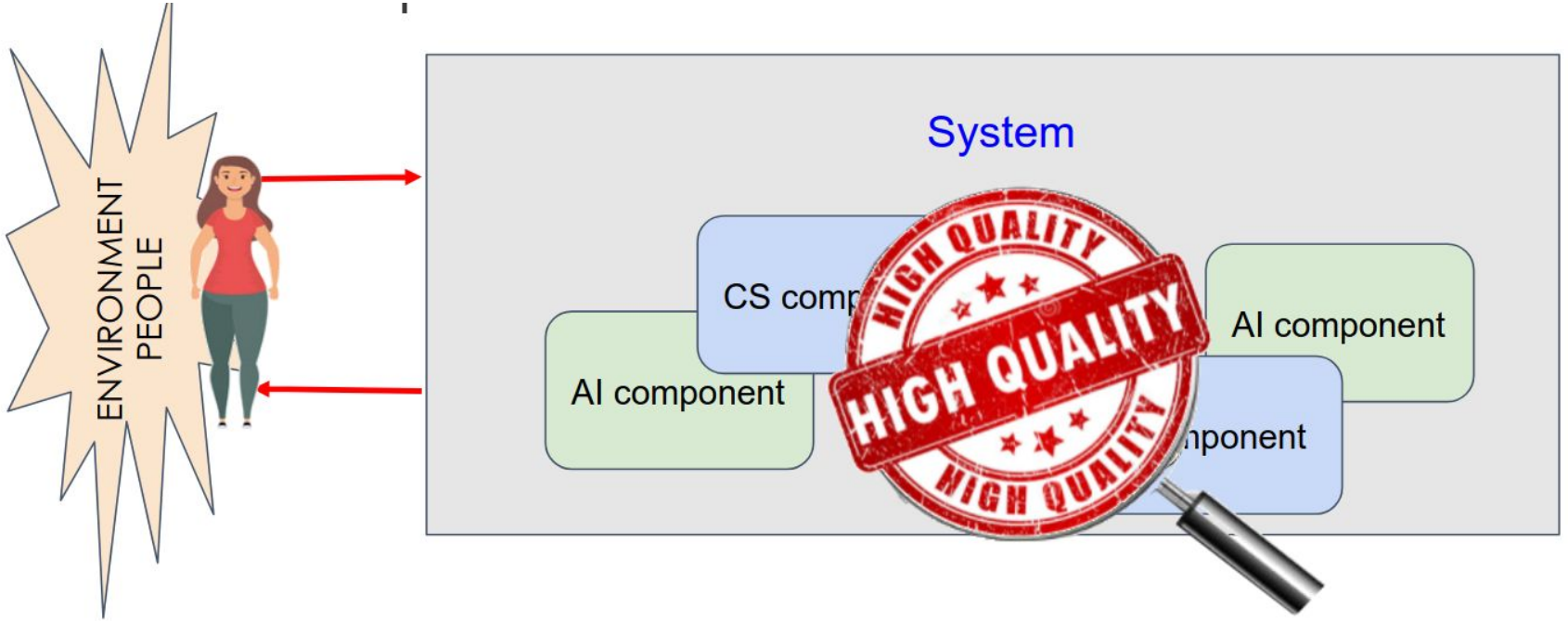
Controllable Text Generation for Evaluating LLMs' Linguistic Competence

Alessio Miaschi, Istituto di Linguistica Computazionale “A. Zampolli” (CNR-ILC), Pisa

Istituto di Linguistica Computazionale “Antonio Zampolli”
Consiglio Nazionale delle Ricerche

Introduction

- Develop **measurable qualities** and assessment/certification techniques
- **WP 5.4**: **Natural Language Generation** and **Text Quality Assessment**



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Introduction

- 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)

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Introduction

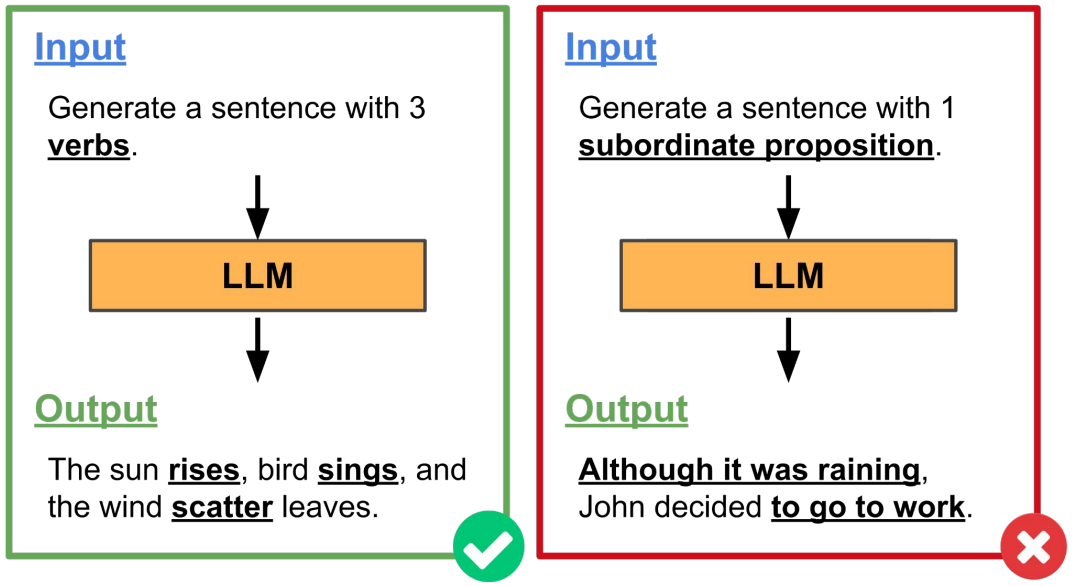
- Motivations:
  - Large Language Models (LLMs) demonstrated remarkable capabilities in solving multiple tasks and in generating coherent and contextually relevant texts
  - Such capabilities have been extensively evaluated against several benchmarks, as evidenced by the success of platforms such as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

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

- 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



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

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.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

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.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

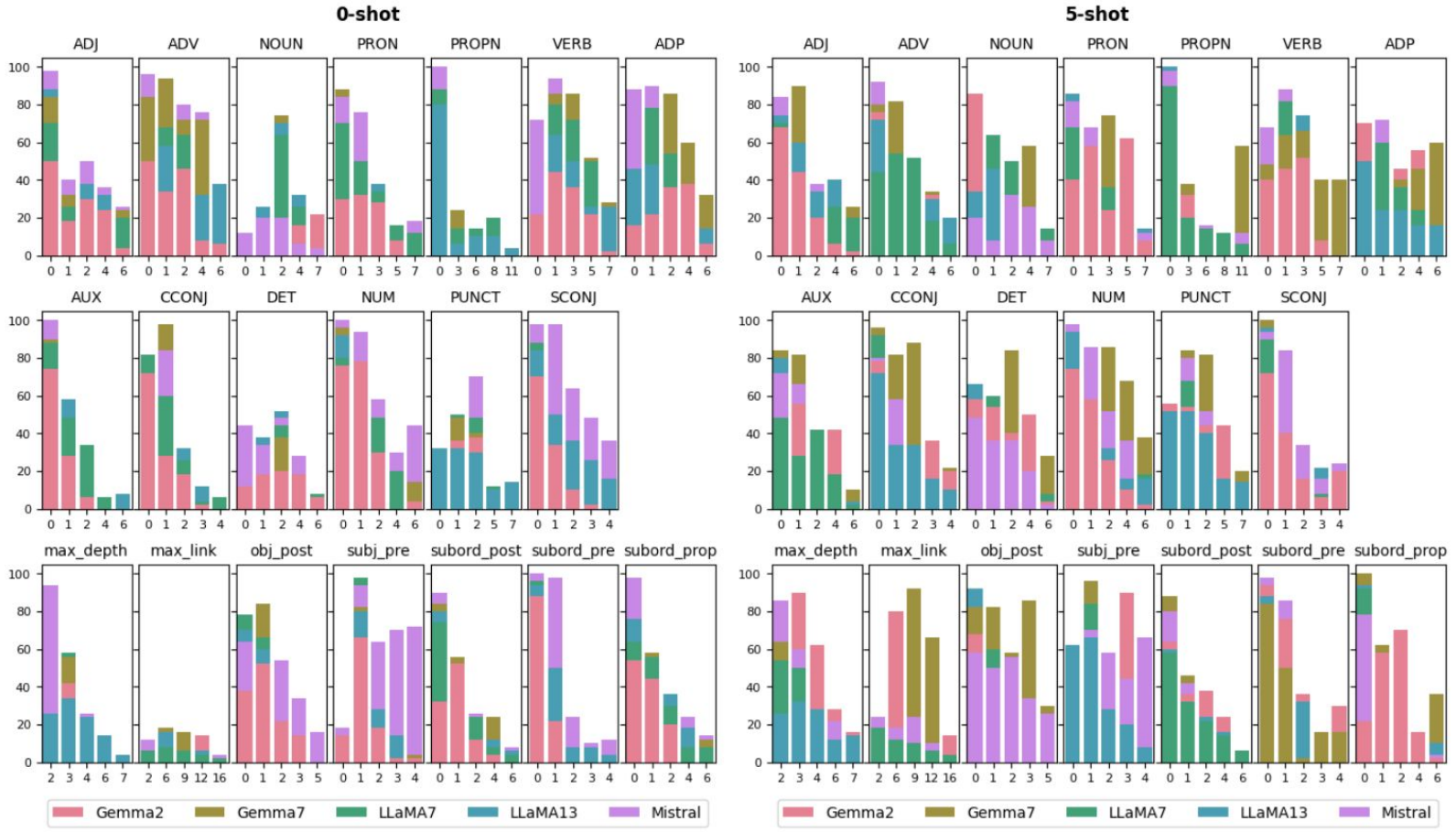
Success Rate Results

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

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

How Do LLMs Follow Constraints Across Values?



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Spearman Results

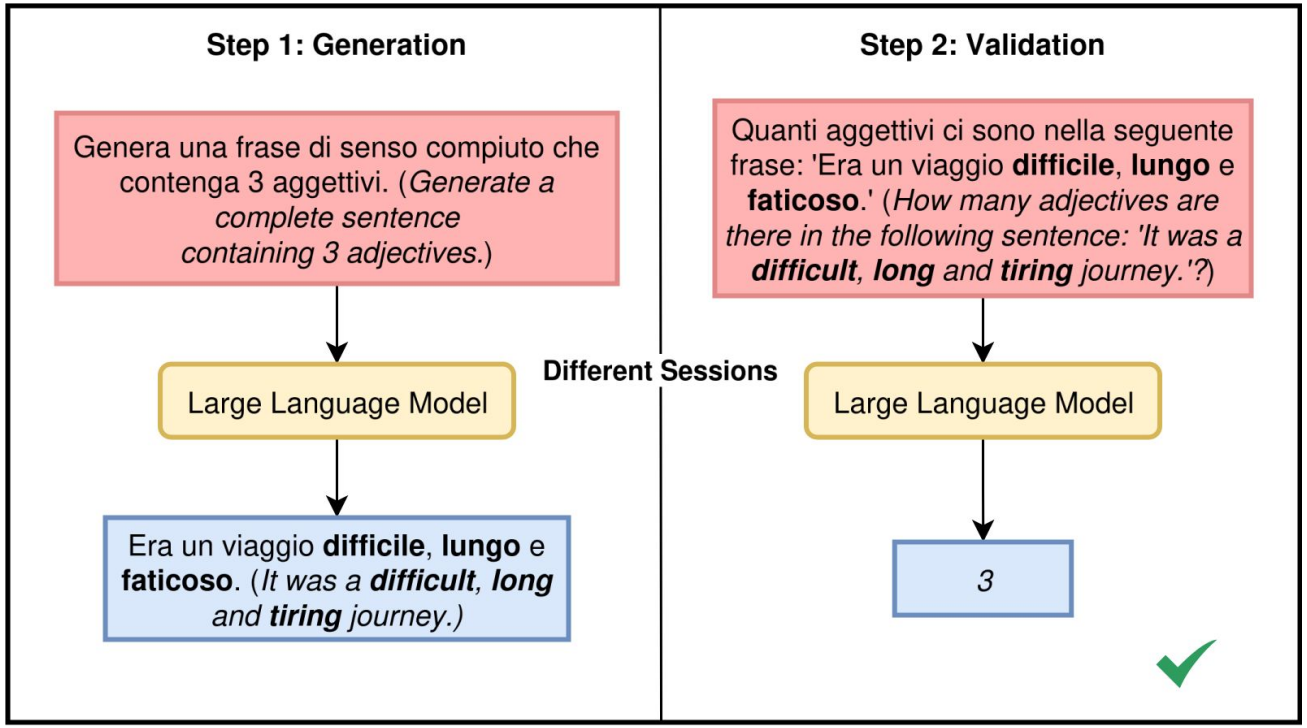
Ling. properties	Gemma2	Gemma7	LLaMA7	LLaMA13	Mistral
			Spearman		
Morphosyntax			0-shot		
ADJ	0.59	0.73	0.74	0.79	0.92
ADV	##	0.88	0.52	0.65	0.95
NOUN	0.63	0.72	0.62	0.66	0.93
PRON	0.26	0.35	0.58	0.80	0.91
PROPN	##	0.66	0.60	0.67	0.88
VERB	0.56	0.83	0.78	0.71	0.76
ADP	0.55	0.89	0.48	0.64	0.96
AUX	##	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

Evaluating Italian LLMs

- Focus on Italian LLMs:
  - ANITA, Camoscio, Cerbero, DanteLLM, Italia, LLaMAntino

- Two-steps evaluation:
  - Generation: e.g. "*Generate a sentence with 2 adjectives*".
  - Validation: e.g. "*How many adjectives does this sentence have?*".



Ciaccio C., Dell'Orletta F., Miaschi A., Venturi G. (2024). Controllable Text Generation To Evaluate Linguistic Abilities of Italian LLMS. In *Proceedings of 10th Italian Conference on Computational Linguistics* (CLiC-it 2024, Pisa)

Generation Results

Model	n_tokens	NOUN	VERB	ADJ	ADV	subj	obj	subord	Avg
ANITA	**.25/.97**	**.47/.97**	**.46/.96**	**.53/.96**	**.45/.91**	.23/.29	**.36/.44**	**.52/.91**	**.41/.80**
Camoscio	.1/*.51*	.14/.44	*.16/.18*	.17/.28	*.16/.17*	.25/*.15*	*.2/##*	*.22/.13*	.18/*.23*
Cerbero	.06/.57	.15/.56	.24/.5	.25/.38	.22/.31	.23/*.15*	.23/*.13*	.26/.33	.21/.37
DanteLLM	.11/.79	.15/.54	.22/.66	.29/.62	.21/.35	**.36/.34**	.31/.3	.32/.51	.25/.51
Italia	*.03*/.62	*.09/.34*	*.16*/.2	*.16/.28*	.18/##	*.22*/.16	.21/.22	*.22/.18*	*.16*/.25
LlaMAntino	.05/.57	.12/.48	.19/.43	.17/.31	.2/.23	.33/.3	.23/.17	.23/.28	.19/.35
Avg	*.1*/**.67**	.19/.56	.24/.49	.26/.47	.24/.33	.27/.23	.26/*.21*	**.29**/.39	

Table 2

Success rate and Spearman correlation coefficients (*SR*/ρ) between the linguistic constraints and the feature values extracted from the generated sentences. The best and worst scores for each property and each metric are highlighted in **bold** and *italic* respectively. Non-statistically significant correlation scores are reported with ##.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Validation Results

	Model	n_tokens	NOUN	VERB	ADJ	ADV	subj	obj	subord	Avg
Cons.	ANITA	.06/.96	.43/.97	.57/.96	.52/.95	.55/.94	.82/.96	.8/.95	.64/.94	**.55/.95**
	Camoscio	.28/.44	.06/.31	.23/.28	.19/.2	.19/.2	.25/.27	.24/.18	.2/##	.2/.23
	Cerbero	.27/.56	.2/.49	.2/.51	.31/.5	.24/.46	.31/.3	.22/.11	.3/.42	.26/.42
	DanteLLM	.21/##	.18/.59	.12/.63	.33/.6	.13/.35	.37/.43	.25/.28	.31/##	.24/.36
	Italia	.26/.54	.04/.27	.16/.31	.02/.14	.02/.11	.28/.39	.21/.23	.25/.28	.15/.28
	LLaMAntino	.06/##	.07/##	.18/##	.2/##	.14/.24	.42/.71	.31/##	.2/.46	.2/.18
	Avg	.19/.42	.16/.44	.24/.45	.26/.4	.21/.38	**.41/.51**	.34/.29	.32/.35	
Cons.+	ANITA	.06/.91	.63/.96	.53/.98	.7/.96	.73/.96	.92/.74	.79/.68	.84/.98	**.65/.9**
	Camoscio	.55/.89	.14/.52	.47/.41	.23/.33	.21/##	.65/.41	.5/.31	.14/##	.36/.36
	Cerbero	.47/.94	.39/.83	.45/.81	.73/.8	.66/.77	.53/.34	.61/.34	.66/.65	.56/.68
	DanteLLM	.38/.94	.36/.8	.39/.82	.63/.85	.32/.44	.56/.45	.51/.36	.63/##	.47/.58
	Italia	.35/.86	.05/.47	.16/.5	.03/##	.08/##	.7/.54	.36/.28	.47/.51	.27/.4
	LLaMAntino	.25/.85	.08/.82	.35/.6	.25/.51	.32/.39	.38/.64	.59/##	.4/.53	.33/.54
	Avg	.34/**.9**	.28/.73	.39/.68	.43/.58	.39/.43	**.62**/.52	.56/.33	.52/.45	

Table 3

Success rate and Spearman correlation coefficients (SR/ρ) between the linguistic constraints asked during sentence generation and the values predicted during the validation step. Consistency results are reported for both the overall sentences (*Cons.*) and a filtered subset of sentences that correctly matched the asked linguistic constraint (*Cons.+*).

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

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**
- When validating each model against their own generated sentences, we noticed that the **generation abilities do not always align with the ability of the models to recognize the linguistic properties of their generated sentences**.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

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)
- Analyzing and controlling the “linguistic profile” of generated texts to develop more robust Machine-Generated Text (MGT) detection systems

Thank you!

alessio.miaschi@ilc.cnr.it