

# Controllable Text Generation To Evaluate Linguistic Abilities of Italian LLMs

Cristiano Ciaccio<sup>1</sup>, Felice Dell’Orletta<sup>1</sup>, Alessio Miaschi<sup>1</sup> and Giulia Venturi<sup>1</sup>

<sup>1</sup>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.

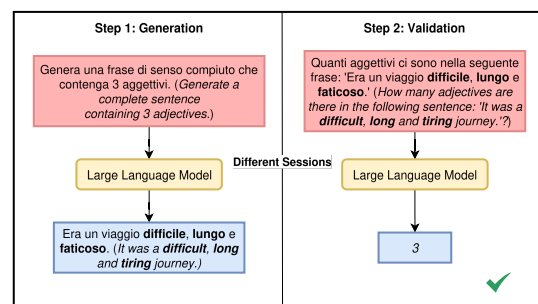
## Keywords

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

## 1. Introduction and Background

Large-scale Language Models (LLMs) [1, 2, 3] have exhibited extraordinary proficiency in a wide range of tasks, from text generation to complex problem-solving, by producing coherent and fluent texts [4]. Their ability to understand context, generate human-like responses, and even engage in creative tasks underscores their potential in various applications. Such capabilities have been extensively evaluated against several benchmarks, as evidenced by the success of platforms such as the OpenLLM Leaderboard [5] or Italian LLM-Leaderboard [6], specifically developed to evaluate Italian models. However, despite their impressive capabilities, the evaluation of LLMs’ linguistic abilities when generating sentences remains an understudied topic. In fact, while earlier works have demonstrated the implicit encoding of many linguistic phenomena within the representations of smaller models [7, 8, 9] or by prompting LLMs to assess their linguistic competence [10, 11, 12], there is no guarantee that generative LLMs can comply with such properties in generating texts.

Studies on Controllable Text Generation (CTG) indirectly assessed models’ capabilities by examining their adherence to linguistic constraints [13]. For instance, [14] studied the abilities of LLMs in adhering to lexical and morpho-syntactic constraints when generating personalized texts. Nevertheless, these works are mainly focused on task-oriented scenarios (e.g. text simplification) and therefore they do not provide systematic evaluations of



**Figure 1:** The diagram shows our evaluation framework composed of two main steps: the first involves the generation of a sentence that adheres to a specific linguistic constraint; while the second consists of asking the model, in a new session, to validate its own generated sentence. The reported example shows a correct case of constrained linguistic generation and validation, indicating a consistent behaviour across tasks.

the linguistic abilities of these models.

From a complementary perspective, in recent years, several works have proposed diverse approaches to assess the consistency of LLMs as an essential component of the models’ evaluation [15], where consistency can be defined as “the requirement that no two statements given by the system are contradictor” [16] or “the invariance of its behaviour under meaning-preserving alternations in its input” [17]. Despite their differences, all these approaches aim to understand the reasoning processes that the models employ in various reasoning tasks [18, 19] while also measuring the predictability and coherence of the models’ generated responses under different conditioning inputs. Among these, [20] studied the consistency between generation (e.g. “what is 7+8?”) and validation (e.g. “7+8=15, True or False?”) of LLMs considering 6 different tasks (e.g. arithmetic reasoning, style

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 – 06, 2024, Pisa, Italy

✉ cristiano.ciaccio@ilc.cnr.it (C. Ciaccio);

felice.dellorletta@ilc.cnr.it (F. Dell’Orletta);

alessio.miaschi@ilc.cnr.it (A. Miaschi); giulia.venturi@ilc.cnr.it

(G. Venturi)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

transfer). [21], instead, employed several consistency checks to measure models’ faithfulness and to understand whether self-explanations truly reflect the model’s behaviour. Importantly, the training procedure of an LM does not explicitly target consistency [17], meaning this ability to produce non-contradictory statements eventually emerges as a byproduct of pre-training and fine-tuning. Therefore, studying models under such conditions serves as a valuable proxy for evaluating their capacity to handle different but complementary tasks, such as generation vs. validation.

In this paper, we bring together the two perspectives and propose an evaluation approach to thoroughly test the linguistic abilities of several Italian LLMs. Specifically, by instructing a model to generate sentences that adhere to a set of targeted linguistic constraints (e.g. “*Generate a sentence with 2 adjectives*”) and then asking to validate its own sentences (“*How many adjectives does this sentence have: <s>?*”), we seek to answer the following research questions: i) To what extent is an Italian LLM capable of generating sentences that adhere to specific linguistic constraints? ii) How consistent are LLM’s responses to the validation questions w.r.t. the specified linguistic constraints? iii) How well can Italian LLMs recognize the linguistic features present in their own generated sentences?

**Contributions.** Our main contributions are:

- We propose a framework for evaluating the linguistic abilities of state-of-the-art Italian LLMs when generating text.
- We conduct extensive evaluations across different models and linguistic constraints.
- We assess models’ consistency with the requested constraints and their ability to validate their own generated content.

## 2. Approach

For the purpose of this paper, we devised a two-step approach aimed at *i*) assessing LLMs’ ability to follow a set of linguistic constraints, and *ii*) validating their ability to recognize the presence of linguistic constraints in generated sentences.

To achieve the first goal, we asked the models to generate sentences with targeted linguistic constraints corresponding to a set of morpho-syntactic and syntactic properties of a sentence, denoted as  $P = \{p_1, p_2, \dots, p_n\}$ . In particular, for each property, we prompted each LLM to produce a fixed number of sentences having a precise value  $v_{p_i}$ , as drawn from a set of possible values  $Vp = \{v_{p_1}, v_{p_2}, \dots, v_{p_n}\}$ . For instance, a prompt asking the model to generate a sentence with two verbs will have the following structure:

*Genera una frase di senso compiuto che contenga  
2 verbi.  
(trad. Generate a complete sentence containing  
2 verbs.)*

Given the well-known difficulty of LLMs in producing texts with precise numerical constraints [13], we decided to constrain the models on increasing values of linguistic properties  $Vp_i$ , to evaluate their ability also to generate sentences following incremental constraints. Our premise lies in the fact that while an LLM may struggle to precisely generate a sentence with an exact value of a particular linguistic property, it is likely to be sensitive to incremental values, i.e. it can generate a sentence characterized by either the absence or the frequent occurrence of a linguistic property.

As a second step, we validate each model against their own samples:

*Quanti verbi ci sono nella seguente frase: <s>?  
(trad. How many verbs does this sentence have:  
<s>?)*

where  $\langle s \rangle$  corresponds to the sentence that the same LLM generated in the previous step. This validation process was conducted by evaluating the models’ responses against the requested linguistic constraints’ values and the actual property values generated by the models. Here the goal is twofold: first, to measure the linguistic consistency of a model, that is if the requested features in the generation step align with the ones found by the model in their own samples; secondly, to assess the models’ ability to correctly recognize the actual properties of their generated sentences.

Due to some models struggling to produce reliable responses in a zero-shot scenario, we experimented with a few-shot scenario<sup>1</sup> to ensure more comparable results.

### 2.1. Linguistic Constraints

The linguistic properties  $P$  we employed as constraints in the generation process include raw, morpho-syntactic, and syntactic properties of a sentence. In particular, we tested the following ones: the length of the sentence in terms of tokens ( $n\_tokens$ ); a subset of Part-Of-Speech (POS) as defined by the Universal Dependency (UD) project [22], i.e. noun (*NOUN*), verb (*VERB*), adjective (*ADJ*) and adverb (*ADV*); the number of subjects and objects in a sentence (*subj* and *obj*), and the number of subordinative clauses in a sentence (*subord*) still as defined by the UD framework. These properties have been shown to play a highly predictive role when leveraged by traditional learning models on various classification problems and can also be effectively used to profile the knowledge encoded in the internal representations of

<sup>1</sup>See Appendix B.1 for details.

Model	Params	Pre-train	SFT/IT	CPT
ANITA	8B	✗	✗	✓
Camoscio	7B	✗	✓	✗
Cerbero	7B	✗	✓	✗
DanteLLM	7B	✗	✓	✗
Italia	9B	✓	✓	✗
LLaMAntino	7B	✗	✓	✗

**Table 1**

Details of the LLMs used in our experiments. The *Pre-train* column indicates if the model was pre-trained exclusively on Italian, the *SFT/IT* column shows whether the model underwent a supervised fine-tuning (SFT) or instruction-tuning (IT) phase for adaptation to the Italian language, and *CPT* (*Continual Pre-training*) indicates whether the model underwent a continual pre-training phase on the Italian language.

a pre-trained Transformer-based model and to enhance their linguistic abilities [23, 24].

### Constraints Selection.

To ensure the selection of authentic property values, we relied on different sections of the Italian Universal Dependency Treebank (IUDT) version 2.5 [25], namely ParTUT [26], VIT [27], ISDT [28], PoSTWITA [29] and TWITTIRÒ [30]. To avoid dealing with excessively short or long sentences, possibly containing non-standard values, we filtered the treebanks to retain only sentences containing a minimum of 5 and a maximum of 40 tokens. The resulting dataset contains 26,744 sentences. Starting from this subset, we selected five increasing values for each linguistic property<sup>2</sup>. Specifically, we asked each model to generate 100 sentences for every value  $v_{p_i}$  within the set of five values  $V_p$ , thus obtaining a total of 500 sentences per property.

Moreover, since we performed our experiments in a few-shot scenario, we used 5 exemplar sentences for each linguistic property extracted from IUDT.

## 2.2. Models

We evaluated several Italian LLMs, with parameter counts ranging from 7 to 9 billion. We specifically leveraged the instruction-tuned variants of these models to assess their ability to adhere more closely to prompts containing detailed instructions. Importantly, we selected models that differ across several factors (architecture, the amount of pre-training and instruction tuning data, the language adaptation strategy, etc.) in order to investigate how these characteristics impact performance. The overall models used in our experiments are: ANITA [31], Camoscio [32], Cerbero [33], DanteLLM [6], Italia<sup>3</sup> and LLaMAntino [34]<sup>4</sup>.

<sup>2</sup>The set of properties values are reported in Appendix B.2.

<sup>3</sup><https://huggingface.co/iGeniusAI/Italia-9B-Instruct-v0.1>.

<sup>4</sup>See Appendix A for more information about the models.

## 2.3. Evaluation

Both steps of analysis were evaluated using two metrics. First, we computed the Success Rate (SR) for each model and linguistic property. Specifically, for the generation of sentences with linguistic constraints, we measured the SR as the fraction of times the model generated a sentence whose property value exactly matched the requested value. For the validation step, we computed the SR as the fraction of times the model’s response accurately matched *i*) the requested linguistic constraint (consistency) and *ii*) the property value of the generated sentence.

As previously mentioned, given the difficulty LLMs have in following precise numerical constraints, we relied also on a metric that measures the models’ abilities to comply with increasing values rather than precise ones. For the evaluation of the generation step, we calculated the Spearman correlation coefficients ( $\rho$ ) between the increasing property values we requested and those extracted from the generated sentences. This metric provides an overall picture of the models’ ability to follow constraints at a macro level, including increasing, decreasing, or removing a specific property when asked. For the validation step, the  $\rho$  correlation was computed between the responses produced by the model and *i*) the requested linguistic constraints, and *ii*) the property values of the generated sentences.

Models’ generated sentences were linguistically annotated with Stanza [35] and further analyzed using Profiling-UD [36], a web-based application that captures multiple aspects of sentence structure. The tool extracts around 130 properties representative of the underlying linguistic structure of a sentence, derived from raw, morphosyntactic, and syntactic levels of sentence annotation, all based on the Universal Dependencies (UD) formalism [37]. Thus, it allows computing the distribution of the set of constrained linguistic properties  $P$  and their values within generated sentences.

## 3. Results

### 3.1. Sentence Generation

Table 2 reports the results in terms of Success Rate (SR) and Spearman correlation ( $\rho$ ) obtained for each model and each linguistic property. When examining the average scores across all linguistic constraints (*Avg* column), we notice that the model rankings remain consistent across both evaluation metrics. Specifically, ANITA consistently outperforms the other models on average, while Italia (SR) and Camoscio ( $\rho$ ) perform the worst. Interestingly, the scores do not correlate with the models’ parameter sizes; for example, the largest model, Italia, ranks poorly in terms of SR. However, a distinction is

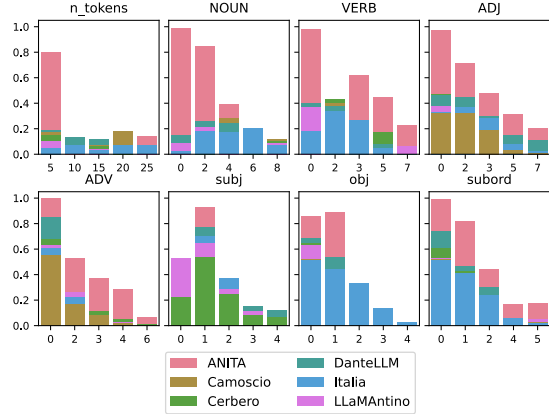
Model	$n\_tokens$	NOUN	VERB	ADJ	ADV	subj	obj	subord	Avg
ANITA	<b>.25/.97</b>	<b>.47/.97</b>	<b>.46/.96</b>	<b>.53/.96</b>	<b>.45/.91</b>	.23/.29	<b>.36/.44</b>	<b>.52/.91</b>	<b>.41/.80</b>
Camoscio	.1/.51	.14/.44	.16/.18	.17/.28	.16/.17	.25/.15	.2/#	.22/.13	.18/.23
Cerbero	.06/.57	.15/.56	.24/.5	.25/.38	.22/.31	.23/.15	.23/.13	.26/.33	.21/.37
DanteLLM	.11/.79	.15/.54	.22/.66	.29/.62	.21/.35	<b>.36/.34</b>	.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
<b>Avg</b>	<b>.1/.67</b>	<b>.19/.56</b>	<b>.24/.49</b>	<b>.26/.47</b>	<b>.24/.33</b>	<b>.27/.23</b>	<b>.26/.21</b>	<b>.29/.39</b>	

**Table 2**

Success rate and Spearman correlation coefficients ( $SR/\rho$ ) 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 **##**.

evident between architectures: models with more recent, higher-performing architectures like ANITA (based on LLaMA 3), DanteLLM, and Cerbero (both based on Mistral) tend to excel. Notably, ANITA stands out with its base model, LLaMA 3, being pre-trained on an impressive dataset of 15 trillion tokens and having already undergone an instruction tuning and alignment phase using both Proximal Policy Optimization (PPO) [38] and Direct Preference Optimization (DPO) [39] in the English language. This suggests that the aforementioned strategy may enhance instruction-following abilities since also DanteLLM was instruction-tuned on Italian starting from the English-instructed version of Mistral. On the contrary, Cerbero, which is based on the non-instruct version of Mistral, obtained lower performance compared to DanteLLM. Given the lack of insight into the models pre-training data and the importance of understanding this phenomenon, further study on the impact of instruction tuning before language adaptation is encouraged.

**Linguistic Properties.** When we analyze which linguistic constraints the models followed the most, we observe notable differences between the two evaluation metrics, highlighting their complementarity and their ability to capture diverse aspects of the models’ constrained sentence generation capabilities. Specifically, the rankings of linguistic properties based on SR and Spearman correlation scores differ significantly. On average (*Avg* row), the top three linguistic characteristics with the highest SR are the use of subordination, subjects and objects (paired with adjectives). In contrast, the top three characteristics with the highest Spearman scores are the length of the generated sentences ( $n\_tokens$ ), the use of adjectives, and verbs. Interestingly, in terms of SR, on average the models struggle with generating sentences featuring a specific length in terms of the number of tokens. One possible explanation for this behaviour could be that, although sentence length can be considered a basic property, its wide range of variation makes it challenging for an LLM to generate sentences with an exact number of tokens compared to other properties. Conversely,  $n\_tokens$  achieves the highest Spearman scores among all models indicating that the models are still capable of following



**Figure 2:** Success rate for each linguistic property and each model. Scores are reported for each group of feature values.

an increasing trend in token constraints.

Figure 2 illustrates, for each model and each property, the SR scores obtained in the generation of sentences with a value  $v_{p_i}$ , reported on the x-axis. This analysis enables us to identify linguistic control elements that models can adhere to more accurately, thereby indicating their proficiency in mastering specific property values within the spectrum of Italian language possibilities. Generally, models achieve lower scores for high property values, while scores tend to be higher when the property value is 0, indicating the absence of the given property. These contrasting trends suggest that models can differentiate between generating sentences with or without a specific property and face greater difficulty with higher property values, which may be less common in Italian. An interesting exception is the *subj* property, where SR scores increase as the property value rises from 0 to 1. This indicates that models are less accurate at generating sentences without a subject.

	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	<b>.55/.95</b>
	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
	<b>Avg</b>	.19/.42	.16/.44	.24/.45	.26/.4	.21/.38	<b>.41/.51</b>	.34/.29	.32/.35	
Cons.+	ANITA	.06/.91	.63/.96	.53/.98	.7/.96	.73/.96	.92/.74	.79/.68	.84/.98	<b>.65/.9</b>
	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
	<b>Avg</b>	.34/.9	.28/.73	.39/.68	.43/.58	.39/.43	<b>.62/.52</b>	.56/.33	.52/.45	

**Table 3**

Success rate and Spearman correlation coefficients ( $SR/\rho$ ) 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.+*).

Model	n_tokens	NOUN	VERB	ADJ	ADV	subj	obj	subord	Avg
ANITA	.07/.95	<b>.47/.97</b>	<b>.32/.96</b>	<b>.46/.95</b>	<b>.44/.92</b>	<b>.35/.29</b>	<b>.31/.41</b>	<b>.49/.9</b>	.36/.79
Camoscio	<b>.15/.75</b>	.25/.53	.28/.29	.18/.29	.19/.17	<b>.63/.17</b>	.4/.17	.17/##	.28/.3
Cerbero	.12/.93	.26/.69	42/.71	.4/.49	.42/.49	.38/##	<b>.55/.19</b>	<b>.49/.45</b>	<b>.38/.49</b>
DanteLLM	.12/##	.26/.64	<b>.51/.75</b>	.42/.72	.35/.23	.49/.2	.44/.2	.46/##	<b>.38/.34</b>
Italia	.04/.8	.18/.52	.2/.38	.28/.16	.28/##	.52/.17	.42/.17	.34/.27	.28/.32
LLaMAntino	.13/##	.18/##	.27/##	.21/##	.33/.27	.26/.3	.36/##	.26/.32	.25/.11
<b>Avg</b>	.11/.57	.27/.56	<b>.33/.51</b>	.33/.43	.34/.36	<b>.44/.19</b>	.41/.19	.37/.32	

**Table 4**

Success rate and Spearman correlation coefficients ( $SR/\rho$ ) between the features extracted from the generated sentences and those predicted during the validation step. 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 ##.

### 3.2. Sentence Validation

As mentioned in Section 2, the validation step of our study is two-fold.

**Consistency.** Table 3 presents the results of the validation of the consistency of the LLMs, evaluated against the requested linguistic constraints’ values. The results are reported for two sets of generated sentences: the entire set (*Cons.* in the table) and the subset including only the sentences generated by correctly following the constraints (*Cons.+*)<sup>5</sup>. A first observation concerns the fact that the scores, both in terms of SR and Spearman, are higher when we consider the *Cons.+* set. This suggests that when the models generate sentences that precisely adhere to the requested values, they tend to answer the validation question more accurately, thus showing greater coherence with the requested constraints. However, we can notice some differences across LLMs, linguistic characteristics and evaluation metrics.

By focusing on the ranking of the LLMs (*Avg* column), we find that ANITA is the most coherent model in terms of both SR and Spearman scores. This aligns with the

<sup>5</sup>Note that for this subset, the number of sentences for each model and linguistic property varies as detailed in Appendix C.

results discussed in Section 3.1: the model that demonstrated the best controlled generation abilities is also the most capable of correctly answering the validation question and the most consistent with the requests. When we focus on the analysis of the linguistic constraints we observe some differences between the two evaluation metrics considered. In terms of SR, both for *Cons.* and *Cons.+*, we notice that the constraints the models are better able to follow (see Table 2) are also those the models can better recognize in the generated sentences. Specifically, these are the three syntactic properties of the sentence we considered (*subj*, *obj*, *subord*). Two main exceptions are ANITA and Camoscio. ANITA, while being the best model in generating sentences with the exact number of requested tokens (*n\_tokens*), is the least able to recognize the length of the generated sentences. On the contrary, for the same constraint, Camoscio, with only a 0.1 SR in sentence generation, is the model most capable of correctly answering the validation question. Such a direct relationship with the generation abilities is less observable for the evaluation in terms of Spearman correlation scores. Namely, the ranking of the Spearman scores in the *Avg* row in Table 3 does not align with the ranking in Table 2. For example, consider the sub-



ject constraint: while it is the constraint that models are, on average, least able to incrementally follow, it is the one with which they are most consistent in terms of the requested values.

**Recognizing linguistic properties.** Table 4 reports the results of the second validation step. A general comparison between the Avg column here and the corresponding column in Table 2, reveals different trends, depending on the evaluation metric. This highlights that our approach effectively distinguishes the models’ varying abilities. Specifically, in terms of SR, most models, except ANITA, show a stronger ability to recognize the linguistic properties of their own generated sentences than to correctly generate sentences with the requested constraint. Conversely, when considering Spearman evaluation, four out of the six models, i.e. ANITA, Camoscio, DanteLLM, and LLaMAntino, demonstrate greater proficiency in generating sentences following incremental constraints than in validating the linguistic properties of those sentences. A final remark concerns the ranking of the linguistic features (Avg row in the table). It generally aligns with the one discussed in Section 3.1 for both evaluation metrics. The main exception is the models’ ability to recognize the exact number of subjects in their own generated sentences. This linguistic characteristic is the best recognized on average across the models in terms of SR (0.44), which is notably higher compared to the average SR of the generation abilities (0.27).

## 4. Conclusion and Future Works

In this paper, we presented the results of a new framework to extensively evaluate the linguistic abilities of Italian LLMs when generating sentences according to multiple linguistic constraints and, subsequently, when validating the linguistic properties of their own outputs. Results showed that models’ architectures and dimensions of pre-training data have an impact on their ability to correctly follow the constraints, with ANITA being the best-performing model across all configurations. When validating each model against their own generated sentences, we noticed that i) LLMs tend to be more consistent with the requested constraints when they correctly followed them during the generation phase, and ii) the generation abilities do not always align with the ability of the models to recognize the linguistic properties of their generated sentences.


Our findings also highlighted that the evaluation metric chosen can significantly affect the results, underscoring the complexity of evaluating LLMs and the necessity for further research in this direction.

Considering that the evaluation of LLMs is an ongoing and multifaceted effort across all languages, we believe that this study opens the way for numerous further

in-depth analyses focused on various aspects of evaluation. Among other aspects, we could evaluate the overall quality of the generated sentences, which we have not accounted for so far. Preliminary investigations revealed that the overall quality of the generations varies across Italian LLMs, with Italia appearing to be the most fluent<sup>6</sup>. Thus, future research should also involve a more comprehensive evaluation that compares the linguistic abilities of LLMs with their fluency and grammaticality.

## Acknowledgments

This work has been supported by:

FAIR - Future AI Research (PE00000013) project under the NRRP MUR program funded by the NextGenerationEU. 

TEAMING-UP - Teaming up with Social Artificial Agents project under the PRIN grant no. 20177FX2A7 funded by the Italian Ministry of University and Research.

## References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
- [2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
- [3] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
- [4] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the power of llms in practice: A survey on chatgpt and beyond, ACM Trans. Knowl. Discov. Data 18 (2024). URL: <https://doi.org/10.1145/3649506>. doi:10.1145/3649506.
- [5] E. Beeching, C. Fourier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, T. Wolf, Open llm leaderboard, [https://huggingface.co/spaces/open-llm-leaderboard/open\\_llm\\_leaderboard](https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard), 2023.
- [6] A. Bacciu, C. Campagnano, G. Trappolini, F. Silvestri, DanteLLM: Let’s push Italian LLM research forward!, in: N. Calzolari, M.-Y. Kan, V. Hoste,

<sup>6</sup>A sample of the generated sentences can be found in Appendix C.

- A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 4343–4355. URL: <https://aclanthology.org/2024.lrec-main.388>.
- [7] G. Jawahar, B. Sagot, D. Seddah, What does BERT learn about the structure of language?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3651–3657. URL: <https://aclanthology.org/P19-1356>. doi:10.18653/v1/P19-1356.
- [8] I. Tenney, D. Das, E. Pavlick, BERT rediscovers the classical NLP pipeline, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4593–4601. URL: <https://aclanthology.org/P19-1452>. doi:10.18653/v1/P19-1452.
- [9] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: What we know about how BERT works, Transactions of the Association for Computational Linguistics 8 (2020) 842–866. URL: <https://aclanthology.org/2020.tacl-1.54>. doi:10.1162/tacl\_a\_00349.
- [10] J. Li, R. Cotterell, M. Sachan, Probing via prompting, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1144–1157. URL: <https://aclanthology.org/2022.naacl-main.84>. doi:10.18653/v1/2022.naacl-main.84.
- [11] T. Blevins, H. Gonen, L. Zettlemoyer, Prompting language models for linguistic structure, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 6649–6663. URL: <https://aclanthology.org/2023.acl-long.367>. doi:10.18653/v1/2023.acl-long.367.
- [12] M. Di Marco, K. Hämmerl, A. Fraser, A study on accessing linguistic information in pre-trained language models by using prompts, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 7328–7336. URL: <https://aclanthology.org/2023.emnlp-main.454>. doi:10.18653/v1/2023.emnlp-main.454.
- [13] J. Sun, Y. Tian, W. Zhou, N. Xu, Q. Hu, R. Gupta, J. Wieting, N. Peng, X. Ma, Evaluating large language models on controlled generation tasks, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 3155–3168. URL: <https://aclanthology.org/2023.emnlp-main.190>. doi:10.18653/v1/2023.emnlp-main.190.
- [14] B. Alhafni, V. Kulkarni, D. Kumar, V. Raheja, Personalized text generation with fine-grained linguistic control, in: A. Deshpande, E. Hwang, V. Murahari, J. S. Park, D. Yang, A. Sabharwal, K. Narasimhan, A. Kalyan (Eds.), Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 88–101. URL: <https://aclanthology.org/2024.personalize-1.8>.
- [15] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: The Eleventh International Conference on Learning Representations, 2023. URL: <https://openreview.net/forum?id=1PL1NIMMrw>.
- [16] A. Chen, J. Phang, A. Parrish, V. Padmakumar, C. Zhao, S. R. Bowman, K. Cho, Two failures of self-consistency in the multi-step reasoning of llms, Transactions on Machine Learning Research (2024).
- [17] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, Y. Goldberg, Measuring and improving consistency in pretrained language models, Transactions of the Association for Computational Linguistics 9 (2021) 1012–1031. URL: <https://aclanthology.org/2021.tacl-1.60>. doi:10.1162/tacl\_a\_00410.
- [18] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al., Language models (mostly) know what they know, arXiv preprint arXiv:2207.05221 (2022).
- [19] L. Parcalabescu, A. Frank, On measuring faithfulness of natural language explanations, arXiv preprint arXiv:2311.07466 (2023).
- [20] X. L. Li, V. Shrivastava, S. Li, T. Hashimoto, P. Liang, Benchmarking and improving generator-validator consistency of language models, in: The Twelfth International Conference on Learning Representations, 2023.
- [21] A. Madsen, S. Chandar, S. Reddy, Are self-explanations from large language models faithful?, ArXiv abs/2401.07927 (2024). URL: <https://api.semanticscholar.org/CorpusID:266999774>.
- [22] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Com-

- putational Linguistics 47 (2021) 255–308. URL: <https://aclanthology.org/2021.cl-2.11>. doi:10.1162/coli\_a\_00402.
- [23] A. Miaschi, D. Brunato, F. Dell’Orletta, G. Venturi, Linguistic profiling of a neural language model, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 745–756. URL: <https://aclanthology.org/2020.coling-main.65>. doi:10.18653/v1/2020.coling-main.65.
- [24] A. Miaschi, F. Dell’Orletta, G. Venturi, Linguistic knowledge can enhance encoder-decoder models (if you let it), in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 10539–10554. URL: <https://aclanthology.org/2024.lrec-main.922>.
- [25] D. Zeman, J. Nivre, M. Abrams, et al., Universal dependencies 2.5, in: LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), 2019. URL: <http://hdl.handle.net/11234/1-3105>.
- [26] M. Sanguinetti, C. Bosco, PartTUT: The turin university parallel treebank, in: R. B. et al. (Ed.), Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project, Springer, 2015, p. 51–69. URL: [https://link.springer.com/chapter/10.1007/978-3-319-14206-7\\_3](https://link.springer.com/chapter/10.1007/978-3-319-14206-7_3).
- [27] R. Delmonte, A. Bristot, S. Tonelli, VIT - Venice Italian Treebank: Syntactic and quantitative features, in: Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories, 2007.
- [28] C. Bosco, S. Montemagni, M. Simi, Converting italian treebanks: Towards an italian stanford dependency treebank, in: Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse, 2013.
- [29] M. Sanguinetti, C. Bosco, A. Lavelli, A. Mazzei, F. Tamburini, PoSTWITA-UD: an Italian Twitter Treebank in universal dependencies, in: Proceedings of the Eleventh Language Resources and Evaluation Conference (LREC 2018), 2018. URL: <https://www.aclweb.org/anthology/L18-1279.pdf>.
- [30] A. T. Cignarella, C. Bosco, P. Rosso, Presenting TWITTIRÒ-UD: An italian twitter treebank in universal dependencies, in: Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019), 2019. URL: <https://www.aclweb.org/anthology/W19-7723.pdf>.
- [31] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, arXiv preprint arXiv:2405.07101 (2024).
- [32] A. Santilli, E. Rodolà, Camoscio: an italian instruction-tuned llama, in: Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR.org, 2023.
- [33] F. A. Galatolo, M. G. Cimino, Cerbero-7b: A leap forward in language-specific llms through enhanced chat corpus generation and evaluation, arXiv preprint arXiv:2311.15698 (2023).
- [34] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, arXiv preprint arXiv:2312.09993 (2023).
- [35] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: A. Celikyilmaz, T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 101–108. URL: <https://aclanthology.org/2020.acl-demos.14>. doi:10.18653/v1/2020.acl-demos.14.
- [36] D. Brunato, A. Cimino, F. Dell’Orletta, G. Venturi, S. Montemagni, Profiling-UD: a tool for linguistic profiling of texts, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7145–7151. URL: <https://aclanthology.org/2020.lrec-1.883>.
- [37] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47 (2021) 255–308. URL: [https://doi.org/10.1162/coli\\_a\\_00402](https://doi.org/10.1162/coli_a_00402). doi:10.1162/coli\_a\_00402.
- [38] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).
- [39] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, C. Finn, Direct preference optimization: Your language model is secretly a reward model, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 53728–53741. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf).
- [40] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettle-



- moyer, Qlora: Efficient finetuning of quantized llms, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), *Advances in Neural Information Processing Systems*, volume 36, Curran Associates, Inc., 2023, pp. 10088–10115. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf).
- [41] G. Sarti, M. Nissim, IT5: Text-to-text pretraining for Italian language understanding and generation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024*, pp. 9422–9433. URL: <https://aclanthology.org/2024.lrec-main.823>.
- [42] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, *arXiv preprint arXiv:2106.09685* (2021).
- [43] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.
- [44] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, in: C. Ghidini, B. Magnini, A. Passerini, P. Traverso (Eds.), *AI\*IA 2018 – Advances in Artificial Intelligence*, Springer International Publishing, Cham, 2018, pp. 389–402.
- [45] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: *Proceedings of Machine Translation Summit X: Papers*, Phuket, Thailand, 2005, pp. 79–86. URL: <https://aclanthology.org/2005.mtsummit-papers.11>.
- [46] C. Xu, D. Guo, N. Duan, J. McAuley, Baize: An open-source chat model with parameter-efficient tuning on self-chat data, *arXiv preprint arXiv:2304.01196* (2023).
- [47] A. Bacciu, G. Trappolini, A. Santilli, E. Rodolà, F. Silvestri, Fauno: The italian large language model that will leave you senza parole!, <https://github.com/andreabac3/Fauno-Italian-LLM>, 2023.
- [48] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration, in: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020. URL: <https://openreview.net/forum?id=rygGQyrFvH>.

## A. Model details

The following section briefly discusses each model’s peculiarities related to the training strategy, data and architecture to show the key differences between the tested models.

**LLaMAntino 7B** [34]<sup>7</sup> is an instruction tuned language model based on Meta’s LLaMA 2 7B [2]: a decoder-only transformer pre-trained on 2 trillion tokens of multilingual texts. The language adaptation phase was performed using QLoRA [40] on the filtered Oscar Dataset for the Italian language released by Sarti et al. [41] (20 billion tokens). The model was further instruction tuned on the Italian translated Dolly dataset<sup>8</sup>.

**ANITA 8B** [31]<sup>9</sup>, is an instruction tuned model based on Meta’s LLaMA 3 8B Instruct, a decoder-only transformer pre-trained on 15 trillion tokens of multilingual texts and further instruction tuned and preference aligned with DPO [39] and PPO [38] using QLoRA. Differently from LLaMAntino, ANITA delays the language adaptation phase by firstly undergoing an instruction tuning and DPO alignment in English on a set of  $\approx 100k$  prompts<sup>10</sup>. Later, the model is adapted to the Italian language by performing SFT on a small sample of 100k examples from the Clean Italian mc4 Corpus [41].

**Camoscio 7B** [32]<sup>11</sup> is an instruction tuned model based on Meta’s LLaMA 7B [2], a decoder-only transformer pre-trained on 1 trillion tokens of English text. Camoscio was developed by performing SFT with LoRA [42] on the translated Alpaca [43] instruction dataset.

**DanteLLM 7B** [6]<sup>12</sup> is an instruction tuned model based on the instruct version of Mistral 7B [3], a transformer decoder-only model pre-trained on internet-scale data (there are no public information on the data used for pre-training). DanteLLM is the result of a LoRA instruction tuning on the Italian SQuAD [44], Europarl dataset [45], Alpaca and Italian Quora [46, 47].

**Cerbero 7B** [33]<sup>13</sup>, is an instruction tuned model based on Mistral 7B. Differently from the other models, Cerbero avoids PEFT (such as LoRA/QLoRA) and directly finetunes Mistral 7B on the Fauno Dataset [47] and a synthetically generated chat dataset.

**Italia 9B**<sup>14</sup> is an instruction-tuned transformer model pre-trained from scratch on trillions of tokens of Italian

Features	$v_{p_1}$	$v_{p_2}$	$v_{p_3}$	$v_{p_4}$	$v_{p_5}$
n_tokens	5	10	15	20	25
NOUN	0	2	4	6	8
VERB	0	2	3	5	7
ADJ	0	2	3	5	7
ADV	0	2	3	4	6
subj	0	1	2	3	4
obj	0	1	2	3	4
subord	0	1	2	4	5

**Table 5**

The sets of property values used for the experiments.

texts. The company behind the model hasn’t released detailed information on the data and architecture.

## B. Further details on the experiments

### B.1. Generation parameters and technical set-up

For the generation of linguistically constrained sentences we set the same parameters across all models: as decoding strategy we used nucleus sampling [48] (top-p = 0.92, top-k = 50, temperature = 0.8); in order to further ensure diversity during generation we randomly sample 1/3 tokens of the last generated sentence and set their probabilities to *-inf* for the next generation step exclusively. In the validation step the decoding is set to be greedy. Due to some models producing explanations and other uninformative textual data we relied on a 5-shot conditioning and on regular expressions to extract the sentences. Given a system prompt  $sys\_p$ , a linguistic feature  $feat$  and a value  $v$ , the linguistically constrained sentence generation task is formatted as follows:

$sys\_p$  + Genera una frase di senso compiuto che contenga +  $v$  +  $feat$ . Non fornire spiegazioni.

(trad.  $sys\_p$  + Generate a complete sentence containing +  $v$  +  $feat$ . Do not give explanations.)

While in the validation step the model is prompted about recognising the linguistic properties of its own sentence  $sent$ :

$sys\_p$  + Quante  $feat$  ci sono nella seguente frase: ' $sent$ '? Non fornire spiegazioni.

(trad.  $sys\_p$  + How many  $feat$  are there in the following sentence: ' $sent$ '? Do not give an explanation.)

For each model we used the author’s recommended chat template and the specified system prompt when available, otherwise we exclude it. All models are loaded

<sup>7</sup><https://huggingface.co/swap-uniba/LLaMAntino-2-chat-7b-hf-UltraChat-ITA>

<sup>8</sup><https://huggingface.co/datasets/basilepp19/dolly-15k-it>

<sup>9</sup><https://huggingface.co/swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA>

<sup>10</sup>[https://huggingface.co/datasets/Chat-Error/wizard\\_alpaca\\_dolly\\_orca](https://huggingface.co/datasets/Chat-Error/wizard_alpaca_dolly_orca)

<sup>11</sup><https://huggingface.co/sag-uniroma2/extremITA-Camoscio-7b>

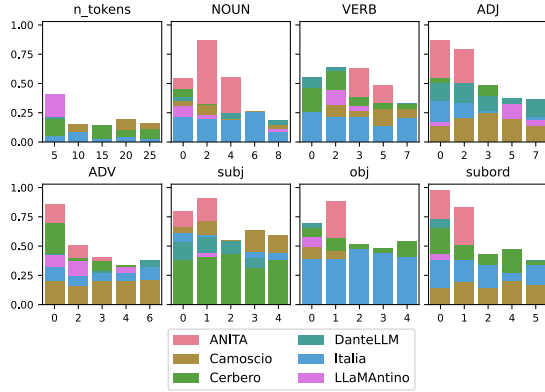
<sup>12</sup><https://huggingface.co/rstless-research/DanteLLM-7B-Instruct-Italian-v0.1>

<sup>13</sup><https://huggingface.co/galatolo/cerbero-7b>

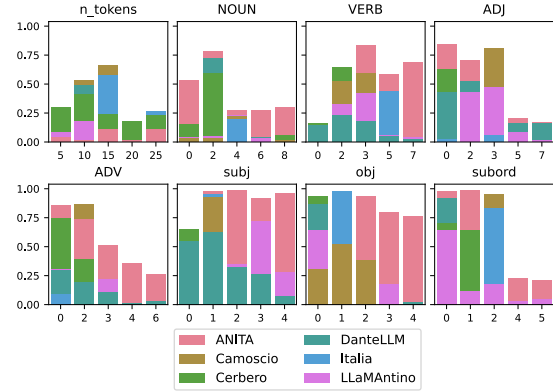
<sup>14</sup><https://huggingface.co/iGeniusAI/Italia-9B-Instruct-v0.1>

Model	n_tokens	NOUN	VERB	ADJ	ADV	subj	obj	subord
ANITA	126	236	230	267	226	113	179	260
Camoscio	49	71	78	87	82	123	101	109
Cerbero	32	76	121	126	110	116	113	128
DanteLLM	55	74	111	147	103	178	154	158
Italia	17	43	82	79	89	110	107	111
LLaMAntino	24	61	95	85	98	164	117	113

**Table 6**  
Number of samples used in the *Cons.+* step.



**Figure 3:** Success rate between the features extracted from the generated sentences and those predicted during the validation step. Scores are reported for each group of feature values.



**Figure 4:** Success rate between the linguistic constraints asked during generation and the values predicted during the validation step. Scores are reported for each group of feature values.

in 16bfloat precision through the Hugging Face API using the *transformer*<sup>15</sup> library. All the experiments are conducted on a single NVIDIA RTX-4090 GPU.

## B.2. Constraint values

Table 5 reports the values of the linguistic properties  $Vp = \{v_{p_1}, v_{p_2}, \dots, v_{p_n}\}$  used in our experiments.

## C. Further details about the results

Table 7 reports a sample of generated sentences by Italia and ANITA when constrained on three linguistic properties: *ADJ*, *NOUN* and *subord*. The full list of the generated sentences, along with the values of their linguistic properties will be freely available upon acceptance.

Table 6 reports the number of samples used in the *Cons.+* subset, i.e. considering only those sentences generated when the models correctly followed the constraints.

Figures 3 and 4 report the SR of the validation step for each group of feature values when computed again

i) the requested linguistic constraints (Fig. 4) and ii) the property values of the generated sentences (Fig. 3).

<sup>15</sup><https://huggingface.co/docs/transformers/index>.

Property	C-value	G-value	LLM	Sentence
ADJ	0	1	Italia	Il suo spirito <b>indomito</b> ha sempre trovato il modo di superare le sfide, senza mai arrendersi, nonostante le avversità ( <i>His indomitable spirit always found ways to overcome challenges, never giving up despite adversity.</i> ).
ADJ	3	3	Italia	Il suo atteggiamento <b>gentile</b> e la sua presenza <b>rassicurante</b> hanno aiutato tutti a sentirsi a proprio agio e ben <b>accolti</b> ( <i>His kind attitude and reassuring presence helped everyone feel comfortable and welcome.</i> ).
ADJ	5	5	Italia	La sua passione per la cucina <b>italiana autentica</b> si manifesta nei suoi piatti <b>ricchi</b> di sapori <b>tradizionali</b> , che richiamano la genuinità della cucina <b>casalinga</b> . ( <i>His passion for authentic Italian cuisine is evident in his dishes rich in traditional flavors that evoke the authenticity of home cooking.</i> )
ADJ	0	0	ANITA	L'uomo va in ufficio ( <i>The man goes to the office.</i> ).
ADJ	2	2	ANITA	Le case <b>vecchie</b> richiedono lavori <b>costosi</b> di restauro ( <i>Old houses require expensive restoration work.</i> ).
ADJ	5	4	ANITA	Il <b>grande</b> ed <b>antico</b> castello si erge sulla cima di un <b>piccolo e ripido</b> colle ( <i>The large and ancient castle stands on the top of a small and steep hill.</i> ).
NOUN	2	4	Italia	Dopo un'attenta <b>analisi</b> , l' <b>esperto</b> ha concluso che la <b>situazione</b> era di una <b>gravità</b> estrema. ( <i>After careful analysis, the expert concluded that the situation was of extreme gravity.</i> )
NOUN	2	2	Italia	La <b>verità</b> è emersa, ma ha lasciato un <b>segno</b> che non potrà mai essere cancellato. ( <i>The truth has emerged, but it has left a mark that can never be erased.</i> )
NOUN	4	4	Italia	La <b>missione</b> umanitaria è stata ostacolata da un'insolita <b>condizione</b> meteorologica che ha causato <b>ritardi</b> nei <b>soccorsi</b> . ( <i>The humanitarian mission was hampered by unusual weather conditions that caused delays in relief efforts.</i> )
NOUN	2	2	ANITA	Il <b>cane</b> corre dietro il <b>gatto</b> . ( <i>The dog runs after the cat.</i> )
NOUN	2	3	ANITA	Il <b>bambino</b> ha in <b>braccio</b> il <b>gatto</b> . ( <i>The child is holding the cat.</i> )
NOUN	4	4	ANITA	I <b>clienti</b> visitano il <b>ristorante</b> con la <b>famiglia</b> e i <b>bambini</b> . ( <i>Customers visit the restaurant with their families and children.</i> )
subord	0	0	Italia	La pioggia porta sollievo alle piante assetate. ( <i>Rain brings relief to thirsty plants.</i> )
subord	1	1	Italia	Il suo viaggio attraverso le montagne è stato reso più agevole dalla presenza di un amico <b>che lo ha accompagnato lungo il percorso</b> . ( <i>His journey through the mountains was made easier by the presence of a friend who accompanied him along the way.</i> )
subord	2	1	Italia	L'impegno di tutti è fondamentale <b>per garantire il successo del progetto</b> . ( <i>Everyone's commitment is essential to ensure the success of the project.</i> )
subord	1	1	ANITA	Ho visitato la città <b>in cui nacque Manzoni</b> . ( <i>I visited the city where Manzoni was born.</i> )
subord	1	0	ANITA	Il concerto inizia solo dopo le nove. ( <i>The concert does not start until after nine o'clock.</i> )
subord	2	2	ANITA	L'uomo <b>che aveva visto il film che era uscito l'anno prima</b> , era rimasto deluso. ( <i>The man who had seen the film that came out the year before was disappointed.</i> )

**Table 7**

Samples of sentences generated by two of the LLMs we considered, each constrained for a subset of linguistic properties: adjectives (*ADJ*), nouns (*NOUN*) and subordinate clauses (*subord*). The constraint value (*C-value*) of each property in the prompt and the actual value (*G-value*) of the property in the generated sentences are provided. Note that we reported samples where the models either correctly or incorrectly follow the constraint. Instances of the constrained property are highlighted in bold within the generated sentences.