Evaluating Large Language Models via Linguistic Profiling

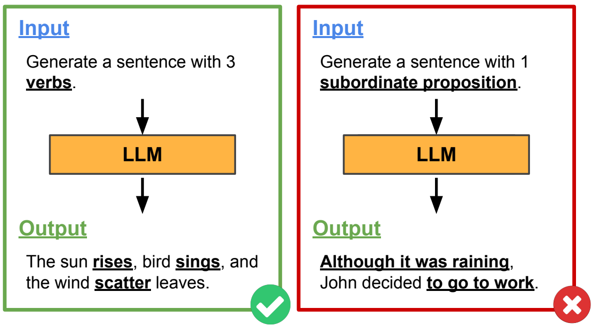
Alessio Miaschi, Felice Dell'Orletta, Giulia Venturi
Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC)
ItaliaNLP Lab, Pisa
{name.surname}@ilc.cnr.it

Abstract

Large Language Models (LLMs) undergo extensive evaluation against various benchmarks collected in established leaderboards to assess their performance across multiple tasks. However, to the best of our knowledge, there is a lack of comprehensive studies evaluating these models' linguistic abilities independent of specific tasks. In this paper, we introduce a novel evaluation methodology designed to test LLMs' sentence generation abilities under specific linguistic constraints. Drawing on the 'linguistic profiling' approach, we rigorously investigate the extent to which five LLMs of varying sizes, tested in both zero- and few-shot scenarios, effectively adhere to (morpho)syntactic constraints. Our findings shed light on the linguistic proficiency of LLMs, revealing both their capabilities and limitations in generating linguistically-constrained sentences[1].

1 Introduction

Recent advancements in Natural Language Processing (NLP) have been significantly shaped by the emergence and refinement of Large-scale Language Models (LLMs) (Achiam et al., 2023; Touvron et al., 2023; Jiang et al., 2023). These models demonstrated remarkable capabilities in solving multiple tasks and in generating coherent and contextually relevant texts, underscoring their potential for capturing complex linguistic structures with high precision and accuracy (Contreras Kallens et al., 2023). Such capabilities have been extensively evaluated against several benchmarks, as evidenced by the success of platforms such as the OpenLLM Leaderboard (Beeching et al., 2023), in a task-oriented scenario covering a wide range of NLP tasks such as commonsense reasoning (Zellers et al., 2019; Hendrycks et al., 2021a), mathematical problem-solving (Hendrycks et al., 2021b), etc. Nevertheless, to the best of our knowledge, a comprehensive evaluation of LLMs' linguistic abilities, independent of specific tasks and possibly cross-cutting across them, is still missing. From a different perspective, studies on Controllable Text Generation (CTG) indirectly tested these abilities by evaluating LLMs in the resolution of specific generative tasks, such as text simplification (Li and Shardlow, 2024) or paraphrase generation (Sun et al., 2021), when conditioned for targeted linguistic constraints. Our hypothesis is that, while earlier investigations have demonstrated the implicit encoding of many linguistic phenomena within the representations of smaller models (Jawahar et al., 2019; Tenney et al., 2019; Rogers et al., 2020; Ettinger, 2020; Ramnath et al., 2020), there is no guarantee that generative LLMs can comply with such properties in generating texts.



Generate a sentence with 3 **verbs**.

The sun **rises**, bird **sings**, and the wind **scatter** leaves.

Generate a sentence with 1 **subordinate proposition**.

Although it was raining, John decided **to go to work**.

Figure 1: Illustrated examples of the evaluation methodology. An LLM is prompted to generate a sentence while adhering to a targeted linguistic constraint (e.g. use of verbs and subordinate propositions).

Building upon these premises, in this work we present the results of an extensive evaluation designed to test LLMs' linguistic abilities to generate sentences while adhering to targeted linguistic constraints representative of various morpho-syntactic and syntactic phenomena. For this purpose, we devised a methodology inspired by the 'linguistic profiling' approach proposed by van Halteren (2004),

[1] Data is available at the following repository: https://github.com/alemiaschi/LLM_profiling.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2835–2848
November 12-16, 2024 ©2024 Association for Computational Linguistics

"in which large numbers of counts of linguistic features are used as a text profile". Thus, it allows computing the distribution of constrained linguistic properties within generated sentences (see Figure 1). The approach was tested on the English language and against five LLMs of different sizes, both in zero- and few-shot scenarios. The approach we devised aims to provide several insights into the linguistic proficiency of LLMs, shedding light on their capabilities and limitations in producing text that adheres to targeted linguistic constraints.

Contributions Our main contributions are:

- We propose a new evaluation framework for assessing the linguistic abilities of LLMs in sentence generation based on the 'linguistic profiling' approach;
- We perform an extensive evaluation of the abilities of the models to generate sentences adhering to specific linguistic constraints;
- We investigate how models, when linguistically constrained for increasing values of a given linguistic property, shape all other properties of the generated sentences accordingly.

2 Related Work

The remarkable and unprecedented performance of LLMs across diverse tasks has significantly increased awareness regarding the importance of evaluating these models. This awareness has prompted a comprehensive reflection on the multifaceted nature of evaluation protocols (Chang et al., 2024). Additionally, the rapid emergence of open leaderboards has become pivotal for comparatively assessing the capabilities and limitations of various models. A well-known example is the OpenLLM Leaderboard platform[2], which provides official rankings for evaluating the performance of models, or the Italian LLM-Leaderboard (Bacciu et al., 2024), specifically developed to evaluate Italian models. Existing leaderboards report LLMs' performance across a spectrum of text-understanding and generation tasks. However, to the best of our knowledge, evaluation methodologies specifically designed to quantitatively assess the multilevel linguistic abilities of LLMs in text generation, independent of specific tasks, are lacking.

Insights in this direction have emerged from the definition of prompting methodologies that have been used to assess the linguistic competence of the models in the resolution of diverse NLP tasks. Li et al. (2022) proposed for the first time a study devoted to evaluating GPT-2 (Radford et al., 2019) in the resolution of 5 tasks, e.g. Part-Of-Speech tagging, showing that the tested properties are indeed encoded in the pre-trained model. Following a similar approach, Blevins et al. (2023) proposed a structured prompting to evaluate the abilities of GPT-neo (Black et al., 2021) and GPT-3 (Brown et al., 2020) models in the resolution of sequence tagging tasks, e.g. Named Entity Recognition. Di Marco et al. (2023) tested multilingual BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) in encoding morphological (e.g. the number of nouns) and syntactic (e.g. distinction between subject and object) properties.

On a different side, studies focused on CTG assessed the linguistic capabilities of LLMs through direct examinations of their adherence to linguistic constraints following an instructional tuning phase (Zhou et al., 2023) or indirectly through the analysis of their performance in diverse tasks. An example of the latter is the study by Sun et al. (2023), which examined the controllability of LLMs across 5 generation tasks. Specifically, they showed that in the syntactically-controlled paraphrase generation task, ChatGPT is good at mimicking syntactic structures from sentences, but it struggles when the syntactic parses are directly incorporated into the prompt. Alhafni et al. (2024), instead, investigated the abilities of LLMs in adhering to lexical and morpho-syntactic constraints, for the task of personalized text generation.

3 Approach

We systematically evaluate the ability of several LLMs to generate sentences with targeted linguistic constraints corresponding to a set of morpho-syntactic and syntactic properties of a sentence, denoted as $P = \{p_1, p_2, ..., p_n\}$. Specifically, we prompted the models to generate sentences containing these properties within a fixed prompt structure. In particular, for each property p_i, we asked the models to generate a fixed number of sentences having a precise value v_{p_i}, as drawn from a set of possible values $Vp = \{v_{p_1}, v_{p_2}, ..., v_{p_n}\}$. For instance, a prompt asking the model to generate a sentence with two verbs will have the following

[2] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

structure: *"Generate a sentence with 2 verbs"*[3].

Given the well-known difficulty of LLMs in producing texts with precise numerical constraints (Sun et al., 2023), we decided to constrain the models on increasing values of linguistic properties Vp_i, in order 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.

We tested our approach across both zero- and few-shot scenarios. Specifically, for the few-shot evaluation, we augmented each prompt with five exemplar sentences for each linguistic constraint.

3.1 Linguistic Properties

The set of linguistic properties P we used as constraints during the generation process encompasses diverse morpho-syntactic and syntactic phenomena of a sentence. They 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 a pre-trained Transformer-based model (Miaschi et al., 2020; Sarti et al., 2021).

Morpho-syntactic properties. They include the set of Part-Of-Speech (POS) defined by the Universal Dependency project (de Marneffe et al., 2021) and can be grouped into two main subsets: content (*PROPN*, *NOUN*, *VERB*, *ADJ*, *ADV*, *PRON*) and functional (*NUM*, *CCONJ*, *AUX*, *ADP*, *DET*, *SCONJ*, *PUNCT*).

Syntactic properties. They include three sub-sets encompassing a range of linguistic phenomena, specifically chosen to test the diverse syntactic abilities of linguistically constrained LLMs. Namely, they are meant to profile the generated sentences according to: i) the global and local syntactic tree structure, including the depth of the syntactic tree (hereafter referred to as *max_depth*) and length of the longest dependency link (*max_link*); ii) word order phenomena, circumscribed to the distribution of the English canonical order of the two main elements of the sentence, i.e. pre-verbal subjects (*subj_pre*) and post-verbal objects (*obj_post*); iii) use of subordination, including the proportion of subordinative clauses in a sentence (*subord_prop*) and their relative order with respect to the principal clause, i.e. in pre- (*subord_pre*) and post-verbal (*subord_post*) position.

3.2 Values Selection

To ensure the selection of authentic values of linguistic properties, we relied on the largest English Universal Dependency (UD) treebank, i.e. English Universal Dependency (EWT) (Silveira et al., 2014), version 2.5 (Zeman et al., 2019). To avoid dealing with excessively short or long sentences, possibly containing non-standard values, we filtered the treebank to retain only sentences containing a minimum of 5 and a maximum of 40 tokens. The resulting dataset contains 19,282 sentences.

The values were extracted from EWT sentences with ProfilingUD tool (Brunato et al., 2020), which allows the extraction of more than 130 properties representative of the linguistic structure underlying a sentence and derived from raw, morpho-syntactic and syntactic levels of annotation based on the UD formalism.

In the few-shot configuration, we used 5 exemplar sentences extracted from EWT. This was done to guarantee the comparability across the experiments, since, for certain values of properties v_{p_i}, EWT contains no more than 5 sentences.

Sets of values. As outlined in Sec. 3, we asked each model to generate a fixed number of sentences following a set of increasing values $Vp = \{v_{p_1}, v_{p_2}, ..., v_{p_n}\}$ for each linguistic property. In particular, we selected five increasing values for each linguistic property[4]. Specifically, we asked each model to generate 50 sentences for every value v_{p_i} within the set of five values V_p, thus obtaining a total of 250 sentences per property.

3.3 Models

We tested the abilities of LLMs of different sizes, ranging from 2 to 13 billion parameters. We opted to leverage the instruction-tuned variants of these models since our aim was to assess the abilities of LLMs specifically tailored to adhere more closely to prompts with detailed instructions. In particular, we utilized Gemma in both 2B and 7B parameters variants (Team et al., 2024), LLaMA-2 (7B and

[3] The list of the prompts used for the experiments is available in the Appendix A.

[4] The set of properties values are reported in Appendix A.

13B parameters) (Touvron et al., 2023) and Mistral (7B parameters) (Jiang et al., 2023)[5].

3.4 Evaluation

The constrained generation abilities of the models were evaluated according to two metrics. First, we computed the Success Rate (SR) for each of the five values v_{p_i} in the set of values V_p. This was measured by the fraction of times the model generated a sentence whose property value exactly corresponds to the one provided. Moreover, for monitoring the ability of the model to adhere to the constraints in terms of increasing property values in the set V_p, we computed the Spearman correlation coefficients (ρ) between the increasing property values extracted from EWT and those extracted from the sentences generated by the models. This metric offers an overview of how well the models are capable of following the constraints at a macro-level, whether increasing, decreasing, or removing a specific property when asked. Given the acknowledged complexity of evaluating LLMs, we opted for two different evaluation metrics, since they offer two distinct perspectives on models' behavior. To extract the set of linguistic properties P from the models' outputs we utilized ProfilingUD.

4 Results

In the following sections, first we present the results obtained by the models in following the linguistic constraints (Sec. 4.1 and 4.2). Subsequently, we analyze how increasing the linguistic property values shapes all the other sentence properties accordingly (Sec. 4.3). Table 1 reports the results in terms of SR and Spearman correlation (ρ) between the linguistic properties extracted from the *EWT* treebank and the generated sentences.

4.1 How Precisely do LLMs Follow Constraints?

As expected, and in line with our initial hypothesis, we generally obtained lower SR values compared with ρ scores (see column *Success Rate* in Table 1). By examining the average scores (*Avg* row), it is evident that Mistral is the most accurate model, both in a zero- and few-shot scenario. This suggests that it is the most proficient model in mastering the (morpho)syntactic knowledge we considered. Conversely, Gemma2 exhibits on average the poorest accuracy. As expected, the larger model variants (Gemma-7 and LLaMA-13) outperform the smaller variants (Gemma-2 and LLaMA-7), with a more notable difference between the two Gemma models. However, it appears that the type of models' architecture also plays a main role. In fact, Mistral, despite having fewer parameters, outperforms a model with almost double of its parameters (LLaMA-13). As expected, all models tend to become more accurate after the few-shot learning, particularly the two versions of Gemma, which become even more accurate than Mistral, which tends to lose accuracy. This result is quite surprising, given that the examples provided in this experimental scenario are the same for all models. However, Sun et al. (2023) found a similar trend when testing the numerical planning abilities of LLMs, showing deteriorated performances after a few-shot phase.

Morpho-syntactic constraints. First, we observed that, on average, **all models tend to adhere slightly more accurately to these constraints rather than syntactic ones**. Focusing on the differences between content and functional POS, we did not observe notable differences among the models in the zero-shot scenario. However, diverse trends emerge in the few-shot scenario: **the majority of models** (except the two versions of LLaMA) **are more accurate in generating sentences with an exact number of functional words**. This is especially the case of Gemma7, which becomes the most accurate model for the majority of POS. On the contrary, Mistral's ability to generate sentences with a precise number of functional words, particularly adpositions (*ADP*) and subordinate conjunctions (*SCONJ*), deteriorates.

Syntactic constraints. For all models, the constraints most challenging to adhere to in the zero-shot scenario are the depth of the syntactic tree of the sentence (*max_depth*) and the length of the longest dependency link (*max_link*), both assuming the knowledge of either global or local structure of the sentence. Quite interestingly, the SR for these two properties remains quite low also after the few-shot learning, even with some differences among the models. Specifically, Gemma-2 becomes the most accurate model in generating sentences with a precise *max_depth* and Gemma-7 the most accurate one in generating sentences with a precise *max_link*. On the contrary, in the few-shot scenario, Mistral's SR remains on average stable, but it deteriorates significantly in generating sentences with a controlled number of pre-verbal subjects (*subj_pre*)

[5] For details about compute parameters and computational budget see Appendix B.

Ling. properties	**Gemma2**	**Gemma7**	**LLaMA7**	**LLaMA13**	**Mistral**	**Gemma2**	**Gemma7**	**LLaMA7**	**LLaMA13**	**Mistral**
			Success Rate					**Spearman**		
Morphosyntax					**0-shot**					
ADJ	25.2	36.8	33.6	42	50	0.59	0.73	0.74	0.79	0.92
ADV	28.8	70.8	34.4	38.8	74	##	0.88	0.52	0.65	0.95
NOUN	8.8	26	23.2	29.6	12.4	0.63	0.72	0.62	0.66	0.93
PRON	19.6	22.8	36.4	34	41.6	0.26	0.35	0.58	0.80	0.91
PROPN	25.6	29.2	28	22	22	##	0.66	0.60	0.67	0.88
VERB	25.2	50.8	46.8	37.2	57.6	0.56	0.83	0.78	0.71	0.76
ADP	23.6	54.4	31.2	31.6	64.4	0.55	0.89	0.48	0.64	0.96
AUX	21.6	23.6	35.2	37.2	29.2	##	0.29	0.32	0.56	0.96
CCONJ	24	33.2	35.6	35.2	33.2	0.27	0.33	0.35	0.33	0.42
DET	14.8	15.6	14.8	25.6	32	0.28	0.36	##	0.28	0.79
NUM	37.6	48	43.2	40.8	65.2	0.49	0.74	0.60	0.62	0.94
PUNCT	14.8	19.2	26	23.6	29.2	0.24	0.54	0.63	0.61	0.78
SCONJ	23.2	27.6	27.6	42.4	68.8	##	0.44	0.40	0.62	0.92
Avg	22.52	35.23	32	33.85	44.58	0.30	0.60	0.51	0.61	0.86
Syntax					**0-shot**					
max_depth	13.6	17.6	16.4	20.4	29.2	##	0.18	##	##	0.76
max_link	9.2	7.2	5.2	6.8	3.6	##	0.44	0.57	0.43	0.75
obj_post	25.2	36.4	35.2	36.4	40.8	0.21	0.47	0.37	0.38	0.59
subj_pre	20.4	21.2	22.8	26.4	63.6	##	##	0.37	0.13	0.84
subord_post	20	36.8	29.2	29.6	32.8	0.13	0.65	0.44	0.58	0.59
subord_pre	22	23.2	24	32.8	48.8	##	0.33	0.13	0.34	0.72
subord_prop	23.6	37.6	33.2	37.2	41.6	0.28	0.60	0.45	0.67	0.83
Avg	19.14	25.71	23.71	27.09	37.2	0.08	0.38	0.33	0.36	0.73
Morphosyntax					**5-shot**					
ADJ	28	47.6	34.4	42.8	45.6	0.19	0.78	0.76	0.79	0.86
ADV	33.2	47.2	34.8	41.2	51.6	0.43	0.62	0.52	0.71	0.80
NOUN	43.6	20.4	34.4	28.4	18.8	0.87	0.76	0.77	0.75	0.90
PRON	38.4	45.6	34	39.2	39.6	0.63	0.65	0.78	0.85	0.81
PROPN	30.4	40.4	28.4	29.6	29.2	0.25	0.87	0.76	0.81	0.81
VERB	29.2	51.6	38.4	37.6	52	0.42	0.77	0.77	0.72	0.87
ADP	44.8	47.2	28.8	26	42	0.46	0.81	0.53	0.61	0.77
AUX	31.6	45.6	27.6	38.4	35.6	0.37	0.70	0.53	0.59	0.60
CCONJ	38	63.6	34	33.2	34.4	0.53	0.56	0.52	0.52	0.60
DET	41.2	37.6	31.6	30	28.4	0.49	0.77	0.65	0.65	0.65
NUM	34	71.6	44.8	43.2	57.6	##	0.63	0.72	0.74	0.77
PUNCT	42	40	34	34.8	31.6	0.60	0.70	0.73	0.79	0.69
SCONJ	30.8	43.2	31.2	40.8	50.4	0.26	0.66	0.62	0.71	0.74
Avg	35.78	46.28	33.57	35.78	39.75	0.42	0.71	0.67	0.71	0.76
Syntax					**5-shot**					
max_depth	52	24.4	30.4	22.4	38.8	0.80	0.56	0.39	0.40	0.78
max_link	22.8	47.2	10	10.8	15.6	0.40	0.86	0.64	0.52	0.70
obj_post	31.6	67.6	32	43.6	44.8	0.42	0.84	0.51	0.62	0.72
subj_pre	51.2	42.4	41.6	36.8	50	0.59	0.52	0.55	0.47	0.74
subord_post	33.2	34	26.4	27.6	34	0.58	0.59	0.53	0.54	0.77
subord_pre	47.6	33.6	34	31.6	45.6	0.12	0.24	0.33	0.35	0.56
subord_prop	33.6	50.4	34.8	32.8	34	0.39	0.79	0.68	0.66	0.74
Avg	38.86	42.8	29.89	29.37	37.54	0.47	0.63	0.52	0.51	0.71

Table 1: Success rate (%) and Spearman correlation coefficients between morpho-syntactic and syntactic properties extracted from the gold and the generated sentences. The best and worst scores for each property and each metric are highlighted in 🟩 and 🟥 respectively. Non-statistically significant correlation scores are reported with *##*.

and subordinate clauses (*subord_prop*).

Sets of constraint values. 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 English language possibilities. Focusing on the 0-shot scenario, we can notice that lower scores are generally associated with the last set, which corresponds to the highest value of each property. This suggests that **models generally encounter more difficulty in generating sentences with higher values (potentially less frequent in the English language) of a property**. However, this trend does not hold for all features: e.g. *subj_pre* exhibits increasing SR scores as the value of the property increases from 2 to 4, and all models struggle to generate sentences without nouns. Conversely, scores tend to be higher when the value is 0, which corresponds to generating a sentence lacking a given property. These two opposite trends serve as evidence that, on average, **models are capable of distinguishing when they are asked to generate a sentence with or without a given feature**. Focusing instead on the few-shot scenario, we can observe a reverse trend: the SR of the last group of values, slightly increases, although remaining comparatively lower on average than the others. This suggests that the models are specializing their linguistic abilities according to the characteristics of the provided EWT samples.

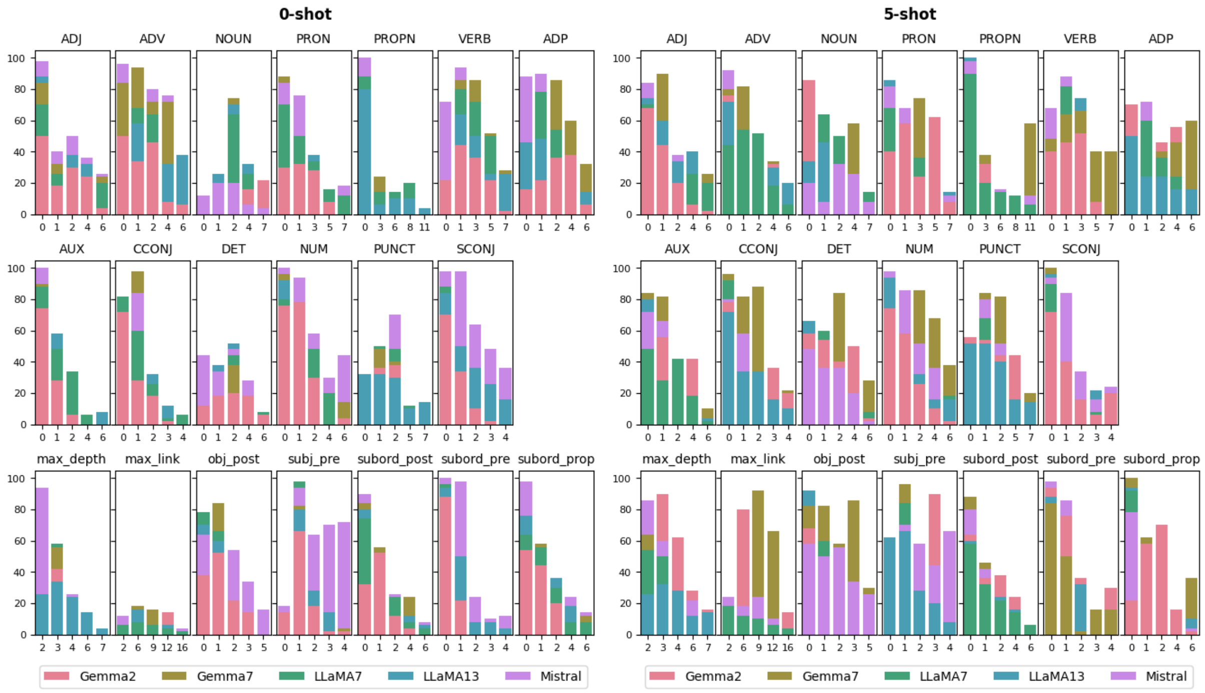


Figure 2: Success rate (%) for each linguistic property and each model in the 0- and 5-shot scenarios. Scores are reported for each group of feature values.

4.2 How Well Do LLMs Adhere to Increasing Constraints?

As discussed previously, although a model may have difficulty producing sentences with an exact v_{p_i}, it may be able to understand the difference between ranges of values. The *Spearman* column in Table 1 shows the models' abilities (in terms of ρ scores) to generate sentences characterized by a distribution of linguistic properties that vary according to the relative increasing p_{v_i}. Although there are general trends similar to those observed earlier, we note some significant differences. First, it is evident that the performance surpasses that of the SR evaluation. This confirms our initial hypothesis, indicating that **while a model may struggle in adhering to a specific value, it shows greater sensitivity to the variations of values within the provided prompts**. Another notable difference is that i) although Mistral's scores tend to be on average lower in the few-shot scenarios compared to the zero-shot ones, the model always remains the best-performing one (*Avg* row of Table 1) and ii) the difference between variants of the same model (i.e. Gemma and LLaMA) is more pronounced than in the previous evaluation. Additionally, the results of this evaluation emphasize the models' stronger adherence to morpho-syntactic rather than syntactic constraints. This suggests that the models are better able to master categorical knowledge, which is simpler compared to the relational competence required to adhere to syntactic constraints.

Morpho-syntactic constraints. On average all models demonstrate higher correlation scores when constrained for content than for functional POS, both in zero- and few-shot scenarios. This suggests that **for all models, it is easier to generate new sentences containing an increasing number of open word classes than closed ones**. This result indicates a further distinction compared to the previous evaluation method. In addition, similar to what was previously observed, Mistral's ability to adhere to POS constraints appears to diminish in the few-shot setting. An in-depth analysis reveals that, according to the present evaluation scenario, this is mostly due to a significant drop in its ability to create sentences with an increasing number of functional words.

Syntactic constraints. Differently to what was observed in Section 4.1, in zero-shot not all models show the lowest correlation scores for *max_depth* and *max_link*, among all syntactic constraints. Thus, even if all models consistently struggle to adhere to specific values of these two syntactic properties correctly, each model has a different sensitivity to adhere to their incremental values. However, similarly, after the few-shot learning, the

Model	Constraint		n_tokens	
	0-shot	5-shot	0-shot	5-shot
Gemma2	3	**10**	13	0
Gemma7	3	**10**	10	4
LLaMA7	6	5	13	12
LLaMA13	7	**10**	8	6
Mistral	**11**	7	4	11

Table 2: Counts of how many times the ρ correlation with the requested linguistic constraint or with sentence length (*n_tokens*) is the highest w.r.t the correlations with the other properties.

majority of the models (except Mistral) seem to acquire this capacity and in particular, Gemma-2, which became the most proficient model in generating sentences controlled for the depth of the syntactic tree, even if we have obtained a non-significant correlation in the zero-shot scenario. As observed for the morpho-syntactic constraints, Mistral in the few-shot diminishes its syntactic abilities. This is mostly the case of the model's capacity to generate sentences with an increasing *max_link* value and amount of subordinate clauses, specifically preceding the main clause (*subord_pre*).

4.3 How Do Linguistic Constraints Shape Sentence Properties?

In the previous sections, we analysed how well LLMs can generate sentences following a linguistic constraint. However, generating new sentences for increasing values p_{v_i} of a given linguistic constraint may lead to correlated changes in the distribution of other linguistic characteristics of the sentence. To explore this, for each controlled property we computed the correlation between the increasing property values in V_p (hereafter referred to as *controlled* values) and the values of all the linguistic properties as generated by the LLMs (referred to as *predicted* values), including also the correlation with the length of the sentence (*n_tokens*).

As can be seen in Figure 3, the correlation values are mostly positive, confirming that **constraining the models to increase values of a specific linguistic property implies a corresponding increase in the others**. However, some sparse cases of negative correlations can be observed, which significantly grow overall across models in the few-shot matrices. This is particularly evident in the case of Mistral when constrained to generate sentences with an increasing amount of numbers (*NUM*) and proper nouns (*PROPN*), and Gemma-7 constrained for determiners (*DET*). This suggests that the models tend to specialize according to the provided samples, adjusting the (morpho)syntactic structure of the generated sentences accordingly.

The diagonal scores of each matrix indicate the correlation between the increase of the controlled and predicted values of the same linguistic property. Consequently, the scores tend to be higher for models with higher Spearman values, as shown in Table 1. However, upon examining how frequently the ρ score between the same controlled/predicted constraint is the highest among all other correlations, we note some notable results outlined in Table 2. As indicated in the *Constraint* column, among the 20 constraints considered, such cases are relatively infrequent, and their frequency increases from the zero- to the few-shot scenario. This trend suggests that **constraining generation for a specific linguistic element does not always primarily enhance that element**; rather, numerous other elements are implicated. This is especially the case of sentence length, a characteristic closely associated with many other (morpho)syntactic properties of the sentence. This relationship is demonstrated by the *n_tokens* column, which indicates that in many cases, increasing the value of a linguistic property results in longer sentences. This holds particularly true for the zero-shot, while after the few-shot phase, the models tend to master the ability to follow the linguistic constraints, **suggesting that they are not simply creating longer sentences, but rather sentences with a varied (morpho)syntactic structure**. The only exception is represented by Mistral. For instance, when the models are constrained to increase the frequency of subordinating conjunctions, there is a high correlation with the distribution of subordinative clauses, indicating the models' proficiency in utilizing subordination. Interestingly, this holds already in the zero-shot configuration. Differently, after the few-shot learning phase, as the controlled values of pronouns increase, the linguistic property of the generated sentence that exhibits the most significant increase is the distribution of pre-verbal subjects. This potentially indicates that the models (excluding Gemma2) are generating personal pronouns to serve as subjects.

Our investigation into the (morpho)syntactic profile of the generated sentences naturally leads us to compare the combinatorial properties specific to the English language with those specific to sentences produced by the LLMs. Therefore, we conducted a comparative analysis between the generated sentences and those from EWT, which are re-

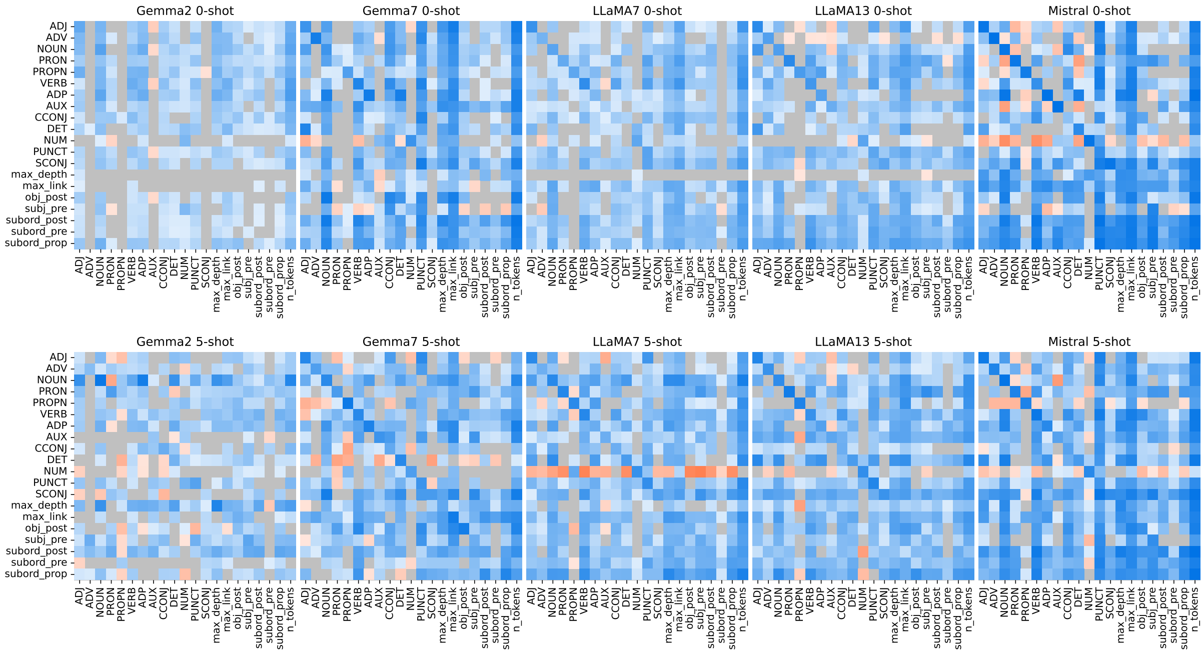


Figure 3: Correlations matrices between *controlled* and *predicted* values. The correlation with the length of the generated sentences (*n_tokens*) is also reported. Gray cells (▨) correspond to non-statistically significant correlations.

garded as representative of English language norms. Our underlying assumption is that despite their inherently different content, the closer the similarities in their (morpho)syntactic properties, the more reliably we can consider the generated sentences as 'naturalistic', i.e. as representative of English language constructs. Thus, we computed the distance between each matrix of Figure 3 and the matrix of the Spearman correlation scores among the values of the 20 considered linguistic properties in EWT sentences[6]. We can consider this second matrix as representative of the authentic combinatorial properties of English since it reports, for each property P, the correlation among their increasing values vp_i with the same property and all the other sentence properties. The average distances among all properties are reported in Table 3[7]. Notably, the ranking of the models by cosine distance mostly resemble the trends previously observed. Specifically, the distance is higher in the 0-shot, especially for Gemma2, which exhibited the lowest correlation score as shown in Table 1. This suggests that all models instructed with only 5 exemplars EWT sentences during the few-shot learning phase tend to generate more naturalistic sentences.

Model	**0-shot**	**5-shot**
Gemma2	0.38	0.31
Gemma7	0.28	0.20
LLaMA7	0.25	0.17
LLaMA13	0.27	0.15
Mistral	0.22	0.17

Table 3: Average cosine distances between the correlation matrix of EWT and predicted correlation matrices for each model. The lowest and highest cosine distances are highlighted in 🟩 and 🟥 respectively.

5 Conclusion

In this paper, we presented the results of an evaluation methodology aimed at testing the linguistic abilities of LLMs of different sizes to generate English sentences according to multiple linguistic constraints. Inspired by the 'linguistic profiling' approach, it is not intended to be a task-oriented evaluation. Given the acknowledged multifaceted nature of evaluating LLMs (Chang et al., 2024), it includes two evaluation metrics showing that while models may struggle in adhering to a specific constraint value, they show greater sensitivity to the value variations. The differences between the scores of the two metrics seem to confirm that they offer two distinct perspectives on models' behaviour, suggesting their suitability for different tasks. Namely, for tasks requiring precise adherence to constraints, e.g. Text Simplification or Question Answering, the

[6] The matrix is reported in Appendix C.

[7] We report in Appendix C the row-wise cosine distances computed per each linguistic property.

Success Rate evaluation is preferred. Conversely, for tasks that rely on the model's ability to follow incremental constraints, e.g. Style Transfer, an evaluation based on Spearman correlation is more appropriate. In addition, our results demonstrated that linguistically-constrained models are capable not only of generating longer sentences but also sentences with varied (morpho)syntactic structures that align with the EWT sentences, which are considered representative of the English language.

Limitations

In this section, we discuss the limitations of our work. 1) **Linguistic properties**: Despite covering multiple aspects of the (morpho)syntactic structure of a sentence, the set of linguistic properties chosen for our experiments are only a possible subset of characteristics that can be used as a testbed in LLM evaluation. Future work should aim to include a broader range of linguistic properties to provide a more comprehensive understanding of the multi-level linguistic competence encoded in LLMs. 2) **Tested LLMs**: Another limitation is the range of LLMs we tested. Ideally, it would be beneficial to include fully open LLMs, to avoid potential biases associated with proprietary models. This would also allow us to anchor the findings not only to the size of the models (in terms of parameters) but also to their pre-training datasets (e.g. textual genres, linguistics characteristics), thereby offering a more robust analysis. 3) **Generated sentences**: In our experiments, we focused on the linguistic structure of the generated sentences without assessing their overall quality. Nevertheless, the quality check of generated sentences is beyond the scope of our study, as it is worth noting that different works have specifically addressed the grammaticality and fluency of LLMs generations (Zhang et al., 2023). However, upon closer inspection, we observed that the vast majority of the generated sentences were fluent, highlighting the remarkable capabilities of these models in producing coherent texts[8]. A further research direction could concern a more comprehensive evaluation where the linguistic abilities of LLMs will be compared with their fluency and grammaticality. 4) **Multilinguality**: Our results are limited to the English language, which constrains the generalizability of our findings to other languages. Since our approach is based on the extraction of linguistic properties (and values) relying on the UD formalism, which is inherently multilingual, it would be relatively straightforward to port this approach to other languages. Therefore, in future work, it would be beneficial to test the approach on a diverse set of languages to evaluate its generalizability and to explore potential cross-linguistic differences in LLM performance. This would provide valuable insights into how well LLMs handle linguistic diversity and the extent to which models trained on languages other than English exhibit different behaviours and capabilities.

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized text generation with fine-grained linguistic control. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 88–101, St. Julians, Malta. Association for Computational Linguistics.

Andrea Bacciu, Cesare Campagnano, Giovanni Trappolini, and Fabrizio Silvestri. 2024. DanteLLM: Let's push Italian LLM research forward! In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4343–4355, Torino, Italia. ELRA and ICCL.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

[8] A sample of the generated sentences by the tested LLMs is reported in Appendix D.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901.

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.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Pablo Contreras Kallens, Ross Deans Kristensen-McLachlan, and Morten H Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marion Di Marco, Katharina Hämmerl, and Alexander Fraser. 2023. A study on accessing linguistic information in pre-trained language models by using prompts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7328–7336, Singapore. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2022. Probing via prompting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1144–1157, Seattle, United States. Association for Computational Linguistics.

Zihao Li and Matthew Shardlow. 2024. How do control tokens affect natural language generation tasks like text simplification. *Natural Language Engineering*, pages 1–28.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sahana Ramnath, Preksha Nema, Deep Sahni, and Mitesh M. Khapra. 2020. Towards interpreting BERT for reading comprehension based QA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3236–3242, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Gabriele Sarti, Dominique Brunato, and Felice Dell'Orletta. 2021. That looks hard: Characterizing linguistic complexity in humans and language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–60, Online. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jiao Sun, Xuezhe Ma, and Nanyun Peng. 2021. AESOP: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5176–5189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the Association for Computational Linguistics*, pages 200–207.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, and et al. 2019. Universal dependencies 2.5. In *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL)*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3).

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *International Conference on Machine Learning*, pages 42602–42613. PMLR.

Features	Prompt
POS	Generate a sentence with v_{p_i} <POS>
max_depth	Generate a sentence with a tree height of v_{p_i} in the syntactic tree
max_link	Generate a sentence with the longest dependency link of v_{p_i} words
obj_post	Generate a sentence with v_{p_i} post-verbal objects
subj_pre	Generate a sentence with v_{p_i} pre-verbal subjects
subord_post	Generate a sentence with v_{p_i} subordinate propositions following the main clause
subord_pre	Generate a sentence with v_{p_i} subordinate propositions preceding the main clause
subord_prop	Generate a sentence with v_{p_i} subordinate propositions

Table 4: Prompts used for the generation of the sentence with the LLMs.

Features	v_{p_1}	v_{p_2}	v_{p_3}	v_{p_4}	v_{p_5}
ADJ	0	1	2	4	6
ADV	0	1	2	4	6
NOUN	0	1	2	4	7
PRON	0	1	3	5	7
PROPN	0	3	6	8	11
VERB	0	1	3	5	7
ADP	0	1	2	4	6
AUX	0	1	2	4	6
CCONJ	0	1	2	3	4
DET	0	1	2	4	6
NUM	0	1	2	4	6
PUNCT	0	1	2	5	7
SCONJ	0	1	2	3	4
max_depth	2	3	4	6	7
max_link	2	6	9	12	16
obj_post	0	1	2	3	5
subj_pre	0	1	2	3	4
subord_post	0	1	2	4	6
subord_pre	0	1	2	3	4
subord_prop	0	1	2	4	6

Table 5: The sets of property values used for the experiments.

A Prompts and Feature Values

As mentioned in Sec. 3, we define a set of fixed prompts for each linguistic property out of the 20 tested in our experiments. Table 4 reports the tested prompts. The values of the linguistic properties $Vp = \{v_{p_1}, v_{p_2}, ..., v_{p_n}\}$ are instead reported in Table 5.

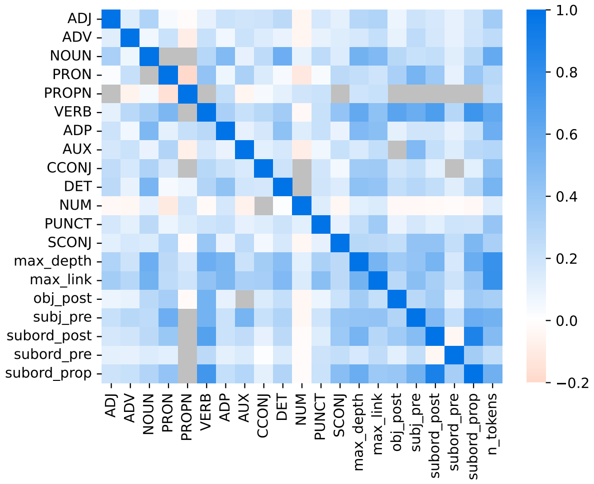


Figure 4: Correlation matrix of the EWT Treebank. Each row in the matrix shows the Spearman correlation scores between the increasing values of a linguistic property vp_i across all sentences in the treebank (*y-axis*) with the same property and all the other sentence properties (*x-axis*). Gray cells (▨) correspond to non-statistically significant correlations.

B LLMs Inference Details

We generated the sentences prompting the models using 8-bit floating points. We used two NVIDIA GeForce RTX 4090 GPUs for the experiments.

C Relationship with EWT Treebank

This appendix contains additional materials useful for deepening the understanding of the relationship between the (morpho)syntactic profiles of sentences generated by LLMs and those specific to the English Treebank.

Specifically, Figure 4 presents the correlation matrix among the values of each linguistic property extracted from the selected sentences in the EWT. This matrix is intended to provide an overview of the combinatorial properties of the English language as represented by the EWT. As expected, the diagonal scores of the matrix are the highest, equal to 1. It follows from the fact that they reflect the correlation between the increasing value of a given linguistic property and itself. Note that these scores are higher than the diagonal scores of the matrices that show the correlations between the controlled and predicted values of the same linguistic property, as illustrated in Figure 3. As discussed in Section 4.3, this indicates that constraining generation for a specific linguistic property does not primarily enhance that property in the generated sentences, but rather affects multiple sentence properties. As a

Features	Gemma2	Gemma7	LLaMA7	LLaMA13	Mistral
0-shot					
ADJ	0.18	0.20	0.13	0.20	0.19
ADV	0.57	0.29	0.27	0.41	0.23
NOUN	0.06	0.09	0.08	0.10	0.17
PRON	0.41	0.43	0.25	0.16	0.16
PROPN	0.71	0.29	0.37	0.43	0.31
VERB	0.18	0.27	0.14	0.20	0.21
ADP	0.10	0.09	0.08	0.12	0.13
AUX	0.58	0.36	0.29	0.26	0.30
CCONJ	0.26	0.28	0.20	0.28	0.26
DET	0.18	0.21	0.34	0.27	0.21
NUM	0.10	0.42	0.21	0.15	0.47
PUNCT	0.29	0.21	0.14	0.20	0.18
SCONJ	0.43	0.24	0.21	0.14	0.14
max_depth	0.85	0.18	0.95	1.11	0.07
max_link	0.63	0.21	0.04	0.07	0.06
obj_post	0.42	0.43	0.29	0.32	0.26
subj_pre	0.54	0.80	0.29	0.38	0.42
subord_post	0.34	0.17	0.14	0.13	0.17
subord_pre	0.53	0.35	0.40	0.34	0.27
subord_prop	0.25	0.16	0.14	0.11	0.12
Avg	0.38	0.28	0.25	0.27	0.22
5-shot					
ADJ	0.43	0.31	0.21	0.18	0.15
ADV	0.26	0.24	0.30	0.22	0.21
NOUN	0.09	0.09	0.08	0.09	0.13
PRON	0.26	0.14	0.08	0.13	0.09
PROPN	0.50	0.24	0.21	0.21	0.25
VERB	0.08	0.08	0.05	0.07	0.11
ADP	0.15	0.08	0.12	0.15	0.11
AUX	0.52	0.25	0.18	0.25	0.23
CCONJ	0.21	0.21	0.16	0.09	0.19
DET	0.32	0.57	0.14	0.13	0.12
NUM	1.12	0.26	0.57	0.17	0.28
PUNCT	0.22	0.18	0.12	0.12	0.16
SCONJ	0.49	0.24	0.13	0.15	0.18
max_depth	0.06	0.09	0.15	0.13	0.04
max_link	0.30	0.07	0.04	0.06	0.06
obj_post	0.32	0.15	0.17	0.19	0.19
subj_pre	0.11	0.21	0.12	0.14	0.31
subord_post	0.08	0.10	0.10	0.12	0.13
subord_pre	0.59	0.33	0.32	0.31	0.28
subord_prop	0.18	0.11	0.09	0.10	0.09
Avg	0.31	0.20	0.17	0.15	0.17

Table 6: Cosine distances between the correlation matrix of EWT and the predicted correlation matrices for each model. The lowest and highest cosine distances are highlighted in 🟩 and 🟥 respectively.

consequence, the correlation between each property's increasing trend and the sentence's length is higher in the generated sentences (Figure 3) than in the EWT sentences, as the length of the sentence is highly related to the increase of any other linguistic structures.

In Table 6 we report instead the cosine distances between the correlation matrix of EWT and the predicted correlation matrices (Figure 3) for each model and each linguistic property. Unlike Table 3, which presents the average distances across all properties, this table provides detailed distances for each individual property. As the distance between the matrices varies significantly depending on the linguistic characteristic considered. This indicates that the ability of LLMs to generate sentences with linguistic structures specific to the English language depends on the particular linguistic phenomenon. Additionally, Mistral is not consistently the model with the smallest distances, suggesting that it is not always the most capable of generating 'naturalistic' sentences.

D Sample of Generated Sentences

Table 7 reports a sample of generated sentences in the zero-shot configuration by Gemma2 and Mistral when constrained on three linguistic properties: *ADJ*, *NOUN* and *subord_prop*. The full list of the generated sentences, along with the values of their (morpho)syntactic properties, are available at the following link: https://github.com/alemiaschi/LLM_profiling.

Property	C-value	G-value	LLM	Sentence
ADJ	1	1	Gemma2	The **beautiful** moon shone brightly that night.
ADJ	4	3	Gemma2	The **vibrant** colors painted upon the **ancient** walls whispered secrets to the **curious** visitor.
ADJ	6	5	Gemma2	The **vast** and **colorful** array of wildflowers presented a **delightful** landscape of **exquisite** beauty and **rare** wonder.
ADJ	0	0	Mistral	John kicks the ball.
ADJ	2	2	Mistral	**Blue** and **elegant** peacocks strut proudly on the lawn.
ADJ	6	6	Mistral	The **magnificent**, **radiant**, **enchanting**, **vibrant**, **exquisite**, **breathtaking** sunset left everyone in awe.
NOUN	1	2	Gemma2	There is only one **thing** in the **world** that can make you forget how to breathe.
NOUN	4	4	Gemma2	The weary **travellers** traversed the desolate **plains**, seeking **respite** from the scorching **heat**.
NOUN	7	7	Gemma2	A **company** that distributes solar **panels** across different **cities** earned the **admiration** of the **community** after their **initiative** to help the **environment**.
NOUN	0	0	Mistral	It's raining.
NOUN	1	1	Mistral	The **clock** chimes.
NOUN	7	7	Mistral	The **monkey** grabbed the red **umbrella**, the yellow **balloon**, the wooden **spoon**, the green **apple**, the blue **bottle**, and the white **tablecloth**.
subord_prop	0	0	Gemma2	She made a wise and insightful decision.
subord_prop	2	2	Gemma2	Despite the storm, the lighthouse remained steadfast, **projecting unwavering beams of hope** that **guided sailors to safety**.
subord_prop	4	3	Gemma2	The tiny spider, despite the large cage, managed to **weave a beautiful web** that **was admired by all** who **saw it**.
subord_prop	1	1	Mistral	**After finishing dinner**, we'll watch a movie.
subord_prop	1	2	Mistral	**Although it was raining**, we decided to **go for a walk**.

Table 7: Samples of sentences generated in the zero-shot configuration by two different LLMs, each constrained for a subset of linguistic properties: adjectives (*ADJ*), nouns (*NOUN*) and subordinate clauses (*subord_prop*). The controlled 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 controlled property are highlighted in bold within the generated sentences.