The OuLiBench Benchmark: Formal Constraints as a Lens into LLM Linguistic Competence

Silvio Calderaro[1], Alessio Miaschi[2] and Felice Dell'Orletta[2]

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

Abstract

Recent progress in Large Language Models (LLMs) has led to impressive capabilities in Natural Language Generation (NLG). However, standard evaluation benchmarks often focus on surface-level performance and are predominantly English-centric, limiting insights into models' deeper linguistic competences, especially in other languages. In this paper, we introduce OuLiBench, a novel benchmark inspired by the literary movement OuLiPo, designed to evaluate LLMs' ability to generate Italian text under explicit linguistic constraints, ranging from morpho-syntactic requirements to creative and structural challenges. Our goal is to assess the extent to which LLMs can understand and manipulate language when guided by specific, sometimes artificial constraints. We evaluate a range of state-of-the-art models in both zero- and few-shot settings, comparing performance across constraint types and difficulty levels. Our results highlight significant variability across models and tasks, shedding light on the limits of controllable text generation and offering a new lens for probing LLMs' generative and linguistic competence beyond traditional benchmarks.

Keywords

Large Language Models, Benchmark, Evaluation, Controllable Text Generation

1. Introduction and Background

The recent and rapid advancements in Large Language Models (LLMs) development has profoundly reshaped the landscape of Natural Language Processing (NLP) [1, 2, 3, 4]. These models exhibit remarkable proficiency across a wide range of tasks, particularly excelling in the generation of coherent and contextually appropriate text. They demonstrate a sophisticated grasp of complex linguistic structures with high accuracy. Such capabilities have been extensively evaluated through a variety of benchmarks, many of which are aggregated on platforms like the Open LLM Leaderboard [5] to facilitate cross-model comparisons.

However, despite the value of these benchmarks as reference frameworks, a significant gap remains in the comprehensive assessment of LLMs' intrinsic linguistic competencies, independently of specific task formulations and with a cross-cutting perspective [6, 7]. Standard evaluation metrics often emphasize surface-level features (e.g., n-gram overlap using BLEU or ROUGE), which may fail to capture deep semantic understanding or robust syntactic flexibility.

Another critical issue, often underestimated in current evaluation methodologies, is the overwhelming predominance of benchmarks developed and validated primarily for the English language [8]. This bias significantly limits the accurate assessment of multilingual systems or models tailored for other languages, such as Italian. Moreover, it impedes the identification and study of culturally specific linguistic phenomena, which are inherently tied to the socio-cultural characteristics of individual linguistic communities.

Concurrently, Controllable Text Generation (CTG) is emerging as a pivotal research area within the LLM domain [9, 10, 11, 12, 13]. CTG focuses on developing and analyzing techniques that guide text generation to conform to explicit constraints, such as style (e.g., formal vs. informal), emotional tone, desired length, structural complexity (e.g., number of subordinate clauses), and predefined semantic content. By leveraging strategies such as prompt conditioning, targeted fine-tuning on annotated datasets, and the implementation of dedicated control mechanisms, CTG research aims to produce generative systems capable of generating outputs that precisely satisfy specified criteria. Intrinsically, this field not only provides methodologies for evaluations better aligned with practical and real-world communicative needs but also emphasizes the models' ability to manipulate language in response to explicit conditions.

This focus on controlled generation naturally raises the question of how far such control can be extended, particularly when constraints become highly specific or even deliberately artificial, designed not merely to produce functional output but to probe the very limits of linguis-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
s.calderaro1@studenti.unipi.it (S. Calderaro); alessio.miaschi@ilc.cnr.it (A. Miaschi); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta)
https://alemiaschi.github.io/ (A. Miaschi); http://www.italianlp.it/people/felice-dellorletta/ (F. Dell'Orletta)
0000-0002-0736-5411 (A. Miaschi); 0000-0003-3454-9387 (F. Dell'Orletta)

tic manipulation and computational creativity. In this regard, there exists a compelling parallel with the principles of the literary group OuLiPo (Ouvroir de Littérature Potentielle), which has long explored the generative potential of formal constraints. By imposing stringent rules on literary creation, OuLiPo demonstrates how limitations can paradoxically unlock new expressive forms and reveal deeper structural properties of language. We hypothesize that such intricate, often playful linguistic challenges, when adapted as evaluation tasks, can yield valuable insights into the degree of fine-grained control an LLM can exert and its implicit understanding of linguistic structure, moving beyond mere fluency to assess true generative competence.

Building on these insights, in this paper we introduce OuLiBench, a novel benchmark, and present an extensive evaluation of LLMs' ability to generate Italian text under targeted linguistic constraints, ranging from morpho-syntactic to stylistic-formal phenomena.

By prompting language models to generate sentences that adhere to specific linguistic constraints (e.g., "Generate a sentence with exactly five words" or "Generate a sentence without the letter 'e'") and, where applicable, evaluating their ability to reflect on these constraints or on properties of the generated text, we aim to address the following research questions: i) To what extent can LLMs produce text that satisfies explicit linguistic constraints defined in OuLiBench, including quantitative, structural, and creative constraints? ii) What differences emerge among various LLMs in their ability to meet complex linguistic constraints, and which types of constraints pose the greatest challenges? iii) How does the nature of the constraint (e.g., syntactic vs. creative) affect the quality and coherence of the generated text?

Contributions. Our main contributions are:

- We propose a framework, based on the OuliBench benchmark, for evaluating the linguistic abilities of state-of-the-art Italian LLMs when generating text.
- We conduct extensive evaluations across different open- and closed-source models and linguistic constraints.
- We evaluated models' abilities across several configurations, testing their performance in zero- and few-shot settings.

2. Our Approach

We systematically evaluate the ability of several LLMs to generate Italian sentences under a range of explicitly defined linguistic constraints. These constraints are formalized as a set of properties $P = p_1, p_2, ..., p_n$, where each property p_i corresponds to a specific quantitative, morpho-syntactic or creative linguistic phenomenon. The goal is to assess to what extent models can control these properties during text generation, and how robustly they generalize across different types of constraints.

For each property p_i, we define a corresponding set of possible target values $V_p = v_{p1}, v_{p2}, ..., v_{pn}$. We prompt the models to generate a fixed number of sentences conditioned on each value v_{pi} using a consistent prompt format. For example, for the property "number of words" a representative prompt would be:

> *Genera 50 frasi composte esattamente da 5 parole ciascuna, escludi dal conto la punteggiatura e gli spazi. [transl. Generate 50 sentences consisting of exactly 5 words each, excluding punctuation and spaces from the count.]*

Considering the difficulty that LLMs show in meeting strict numerical specifications, such as generating sentences with an exact length in terms of words or characters, we intentionally structured the evaluation around increasing values of each property. This approach allows us to examine whether the models are sensitive to the relative ordering and magnitude of constraints, even when exact conformity is difficult to achieve. The underlying hypothesis is that although a model may not reliably produce a sentence with exactly 5 words, it may still exhibit a monotonic tendency, generating progressively longer sentences as the required number increases.

For syntactic constraints, such as those related to the syntactic order of the elements (e.g. SVO, SOV, VSO), the analysis focused on the model's ability to adapt the syntactic structure of the sentence to predetermined patterns. Here, the aim is to assess the structural flexibility of the model and its ability to model the output according to specific grammatical configurations. Finally, concerning OuLiPo-inspired linguistic constraints, such as lipograms (texts that deliberately omit a particular letter) and tautograms (texts in which all words start with the same letter), the evaluation was structured around specific letters of the alphabet, testing the model's ability to inhibit or concentrate the use of certain letters within the generated sentences. This allows us to examine the controllability of the models in more creative and stylistic contexts, where the constraints are not numerical but qualitative and symbolic.

The linguistic constraints span both formal properties (e.g. sentence length in words or characters, permutations of sentence elements in the context of linguistic typology) and creative phenomena (e.g., lipograms, tautograms, acrostics), enabling a comprehensive evaluation of controllability across structural and stylistic dimensions. In all cases, the evaluation assesses whether the generated sentence not only satisfies the target constraint but

also maintains syntactic correctness, semantic coherence, and linguistic appropriateness in Italian.

3. OuLiBench

To address the need for more granular evaluation tools for the Italian language, we developed **OuLiBench**. This novel benchmark is specifically designed to thoroughly analyze the capability of LLMs to generate text while adhering to a diverse and progressively complex set of explicit linguistic constraints, thereby moving beyond assessments based on mere surface-level fluency.

3.1. Conceptual Framework and Task Taxonomy

The conceptual foundation of OuLiBench integrates principles from **Controllable Text Generation (CTG)** [10], which focuses on guided generation according to predefined attributes, with the creative, constraint-based methodologies of the **OuLiPo (Ouvroir de Littérature Potentielle)** literary group. Founded in 1960 by writer Raymond Queneau and mathematician François Le Lionnais, OuLiPo emerged as a revolutionary literary movement that sought to explore the potential of literature through the systematic application of formal constraints. In their Premier Manifeste (First Manifesto) [14] of 1961, Le Lionnais articulated the group's foundational philosophy Littérature potentielle, defining littérature potentielle as "the search for new structures and patterns that can be used" to create literary works. The group used the restrictions of literary forms to spark creativity, developing techniques such as lipograms (texts excluding specific letters), tautograms, anagrams and palindromes. This approach demonstrated that systematic limitations could paradoxically expand rather than restrict creative possibilities, generating what the group termed "potential literature". OuLiBench adapts these philosophies into a suite of computationally evaluable tasks, entirely formulated and contextualized for the Italian language.

OuliBench is organized according to a taxonomy that reflects different levels and types of linguistic control:

1. **Quantitative Constraints:** This category assesses the precision of dimensional control over the textual output. Tasks require models to generate sentences adhering to an **exact word count** or an **exact character count** (net of punctuation and spaces). These constraints challenge models to balance numerical restrictions with semantic coherence and grammatical correctness.
2. **Syntactic Constraints:** These tasks evaluate the models' competence in manipulating fundamental Italian grammatical structures. They include **verbal diathesis control** (requiring generation in active, passive, or reflexive/medium voice) and **constituent order permutations** (Subject-Verb-Object), testing flexibility in generating canonical and non-canonical sentence structures.
3. **Stylistic-Formal (OuLiPo-inspired) Constraints:** Representing the most elaborate challenges, this category implements OuLiPian *contraintes*. It includes tasks such as the **Lipogram** (omission of specific letters), **Inverse Lipogram** (mandatory inclusion of specific letters), **Tautogram** (all words starting with the same letter), **Anagram** (at both word and phrasal levels), **Palindrome** (symmetrical text), and **Acrostic** (initial letters of words forming a target word). These tasks demand advanced linguistic planning and sophisticated sub-lexical and structural manipulation.

For each task, specific prompts were formulated in Italian. Table 1 provides a comprehensive overview of the tasks included in OuLiBench, as well as the prompts used for generating the sentences.

4. Experimental Setting

We evaluate a pool of Italian LLMs by testing their ability to follow the linguistic constraints defined in OuliBench. We conduct our experiments in both zero-shot and few-shot settings. In the zero-shot condition, the model receives only the instruction formulated in natural language. In the few-shot configuration, the prompt is augmented with five, ten, and fifteen exemplar sentences corresponding to the same constraint. This setup is intended to investigate whether LLMs improve in constraint-following behaviour when exposed to in-context demonstrations. In the following, we describe the set of tested models and the evaluation strategy adopted to assess the extent to which generated outputs satisfy the defined constraints.

4.1. Models

The landscape of Italian large language models (LLMs) is evolving rapidly, with notable differences in development strategies. Some models have been pre-trained from scratch with intrinsic emphasis on the Italian language, while others have been fine-tuned for Italian starting from well-established architectures. For this study, we selected models with comparable parameter scales: Minerva-7B-instruct-v1.0 (SapienzaNLP) [15], Italia-9B-Instruct-v0.1 (iGeniusAI) [16], Velvet-14B (Almawave) [17], Maestrale-chat-v0.4-beta (mii-llm) [18], and LLaMAntino-3-ANITA-8B-Inst-DPO-ITA (SWAP-UNIBA) [19]. The first group includes three models pre-trained from scratch. Minerva-7B-instruct-v1.0 is

Category	Task Name	Constraint Description	Example Target Sentence (Italian)
Quantitative	Length by Words	Generate Italian sentences with an exact word count.	"Il gatto dorme sul divano." (5 words)
	Length by Characters	Generate Italian sentences with an exact character count (no punct/space).	"Mangio la pizza" (13 chars)
Syntactic	Diathesis Control	Generate Italian sentences in specified voice (active, passive, reflexive).	"La lettera è scritta da Marco." (passiva)
	Word Order Permutations	Generate Italian sentences using specific SVO permutations (SOV, VSO, etc.).	"Mangia la mela Luca" (VOS)
Stylistic-Formal (OuLiPo-inspired)	Lipogram	Generate Italian text excluding a specific letter.	"Oggi vado in montagna" (without 'e')
	Inverse Lipogram	Generate Italian sentences where a specific letter appears min. once for each words.	"Questo esercizio contiene molte esse." ('e')
	Tautogram	Generate Italian text where all words start with the same letter.	"Maria mangia mele morbide" ('m')
	Word Anagram	Generate a valid Italian anagram for a given Italian word.	"Noce" $\rightarrow$ "Ceno"
	Phrasal Anagram	Reorder sentence letters into a new meaningful Italian sentence.	"Amo Roma" $\rightarrow$ "Moro ama"
	Palindrome	Generate Italian text reading the same forwards and backwards.	"Aceto nell'enoteca"
	Acrostic	Generate Italian text where initial word letters form a target word.	"Viva V.E.R.D.I."

Table 1
OuLiBench Task Summary (Evaluated on Italian).

a 7-billion-parameter Transformer pre-trained on 2.5 trillion tokens, balancing Italian, English, and code, and later refined through supervised fine-tuning (SFT) and direct preference optimization (DPO). Italia-9B-Instruct-v0.1 is a 9-billion-parameter Transformer trained from scratch on trillions of tokens, with a strong focus on Italian and domain-specific content. Velvet-14B is a dense 14-billion-parameter Transformer trained from scratch on the Leonardo HPC system using 4 trillion multilingual tokens, approximately 23% of which are in Italian, achieving competitive scores on Italian-language benchmarks. These models integrate Italian language knowledge from the earliest stages of training. The second group is based on existing architectures. LLaMAntino-3-ANITA-8B-Inst-DPO-ITA is derived from Meta-LLaMA-3-8B-Instruct and specializes in Italian through super-fine-tuning (QLoRA SFT) on mixed datasets and DPO optimization. Maestrale-chat-v0.4-beta, based on Mistral-7B, underwent continued pre-training on an Italian corpus and "Occiglot," followed by conversational SFT and DPO alignment aimed at improving factuality and mathematical reasoning. Although these models build upon pre-trained foundations, they have invested significantly in adapting and optimizing for the specific characteristics of the Italian language. To achieve a comprehensive and diversified evaluation of LLM capabilities across the tasks proposed by the benchmark, it was essential to extend the comparison to include larger proprietary models that currently represent the state of the art in the field. This strategic choice enabled assessment of the selected Italian open-source models in relation to the highest standards achieved by global research and development. Specifically, the comparison included Claude Sonnet 4 [20], DeepSeek [21], Gemini 2.5 Flash, and GPT-4o mini [2].

4.2. Prompting Optimization

The effectiveness of text generation using advanced Language Models is critically dependent on the calibration and formulation of prompts. Our research has systematically analyzed the interaction between prompt structure and output quality for each model, defining optimized strategies to maximize compliance with experimental requirements. Generally, precision in criteria definition was found to be critical: for text length control, making explicit the exclusion of non-linguistic elements (such as punctuation and spaces) significantly improved the precision of some models (Maestrale and Anita). Similarly, for the handling of verbal diathesis in particular middle (or reflexive) diathesis, explicit formulations reduced interpretive ambiguities, increasing the adherence of outputs. In the context of OuLiPo constraints, whenever possible we avoided specific terminology in the prompt (Lipograms, Inverse Lipograms, Tautograms, and Palindromes), describing the task directly and using quotation marks to highlight restricted letters.

A crucial aspect of our methodology was the implementation of few-shot learning, exploring its configurations with 0, 5, 10 and 15 examples. The tasks that employed few-shot were: quantitative constraints, diathesis, Lipograms, Palindromes. The examples were collected from the Italian Universal Dependency dataset, a corpus consisting of 34,383 sentences derived from the main Italian treebanks included in the Universal Dependencies project, including ISDT[22] VIT[23], PARTUT[24], PoSTWITTA[25] and TWITTIRO [26].

During few-shot experimentation, it emerged that the Minerva and Velvet models tended to slavishly reproduce the examples provided in the prompt, generating outputs identical or nearly identical to the initial examples, regardless of the variation required by the task. This

behavior compromised the evaluability of the outputs, as it did not allow verification of the model's ability to generalize or adapt to the specific constraint. Consequently, these models were excluded from the tables related to few-shot configurations.

4.3. Evaluation Strategy

The assessment of model performance within OuLiBench employs an integrated approach, combining quantitative metrics for formal adherence with qualitative analyses for more nuanced aspects of generation.

The primary quantitative metrics are:

- **Success Rate (SR):** Calculated as the percentage of generated outputs that *perfectly* satisfy the linguistic constraint imposed by the specific task. This metric provides a direct measure of the model's precision.
- **Spearman's Rank Correlation Coefficient (ρ):** Used to determine the models' sensitivity to incremental or decremental variations in constraints (e.g., whether models produce longer sentences when requested to increase word count), even when exact adherence is not achieved. This metric was only computed for the evaluation of the quantitative constraints.

To apply these metrics, particularly for SR on constraints involving specific lexical or syntactic features, model outputs were pre-processed and analyzed, partly with the support of linguistic analysis tools. In particular, we employed ProfilingUD [27], a tool that 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. ProfilingUD was specifically applied to the sentences generated by the tested models to extract linguistic features used to evaluate model performance (e.g. sentence length, in terms of tokens or characters, diathesis control, etc.).

The qualitative analysis was carried out manually on the responses that had passed the automatic evaluation, meaning those that met the formal constraints required by the task. The aim was to examine more closely the linguistic quality of the sentences produced, considering three main aspects: grammatical correctness, semantic coherence, and linguistic appropriateness. These criteria were not applied according to a strict hierarchy, although semantic coherence often played a central role, as it is crucial for the comprehensibility and meaning of the sentence. In the presence of particularly strong constraints, such as in the case of tautograms or anagrams, the evaluation was conducted with greater flexibility. The rigidity of the structure required by these constraints can compromise the naturalness of the sentences, making it necessary to allow some tolerance in assessing the other qualitative aspects.

5. Results

The results obtained from the application of the OuLiBench benchmark highlight substantial differences among the tested models, both in terms of absolute capabilities and sensitivity to various types of linguistic constraints. The analysis was conducted considering both quantitative metrics (**Success Rate** and **Spearman's correlation**) and qualitative evaluations of semantic coherence and grammatical correctness.

5.1. Overall Performance

Table 2 reports the results obtained by the Italian open-source models, which highlight a significant variability in models' linguistic control capabilities. **LLaMAntino-3-ANITA-8B-Inst-DPO-ITA (Anita) stands out as the best-performing Italian model, achieving an average SR of 53% in the zero-shot setting**, clearly outperforming the others. Velvet-14B reaches an average of 29%, while Maestrale-chat-v0.4-beta and Minerva-7B-instruct-v1.0 show more limited performance, with 19% and 12% respectively.

To better contextualize these results, Table 3 reports the performance of larger proprietary models, which can be considered as an upper bound relative to the Italian ones. Within this group, **Gemini 2.5 Flash** achieves the highest performance with an overall average of 70%, followed by **GPT-4o mini** (66%) and **DeepSeek R1** (65%). **Claude Sonnet 4**, while competitive across several tasks, records an overall average of 61.5%.

5.2. Analysis by Constraint Categories

5.2.1. Quantitative Constraints

Length control tasks proved to be the most challenging for all tested models. In word-count control, Gemini performed best (34%), followed by DeepSeek (30%) and GPT-4o mini (17%), while Claude obtained the worst performance (9%). Among open-source models, Anita achieved 27% in zero-shot, significantly outperforming Maestrale (9%), Velvet (5%), and Minerva (3%). Spearman correlations were consistently high for proprietary models (94%–100%), thus indicating strong ordinal sensitivity despite difficulties in precise control.

Character-count control was even more demanding: Gemini led (14%), trailed by GPT-4o mini (13%), DeepSeek (05%) while Claude struggled severely (0.03%). Anita remained competitive (15%) among open-source models,

Task	Anita				Maestrale				Minerva				Velvet				Task Avg
	0	5	10	15	0	5	10	15	0	5	10	15	0	5	10	15	
Word Length	**.27/.90**	.25/.95	.24/.95	.13/.93	.09/.88	.06/.64	.04/.68	*.02/.89*	.03/##	-	-	-	.05/.66	-	-	-	**.11**
Char. Length	**.15/##**	.13/.46	.02/.31	.13/.60	.006/.71	.03/.92	.03/.88	.03/.96	*0/-.64*	-	-	-	.01/.60	-	-	-	**.04**
Diathesis	**.99**	.89	.93	.90	*.59*	1	.90	**1**	.72	-	-	-	.67	-	-	-	**.74**
Permutations	**.16**	-	-	-	.12	-	-	-	*.08*	-	-	-	.16	-	-	-	**.13**
Lipograms	.59	.56	.62	**.64**	.32	.37	.35	.34	*.28*	-	-	-	.47	-	-	-	**.41**
Inverse Lipogr.	**.55**	-		.	.26	-	-	-	*.04*	-	-	-	.23	-	-	-	**.41**
Word Anagrams	**.92**	-	-	-	.18	-	-	-	*0*	-	-	-	.10	-	-	-	**.30**
Sent. Anagrams	.88	-	-	-	*0*	-	-	-	*0*	-	-	-	**.90**	-	-	-	**.44**
Tautograms	**.73**	-	-	-	.55	-	-	-	*.007*	-	-	-	.08	-	-	-	**.34**
Palindromes	**.54**	.12	.40	.20	.006	.04	.02	*0*	*0*	-	-	-	*0*	-	-	-	**.13**
Acrostics	**.10**	-	-	-	.02	-	-	-	*0*	-	-	-	*0*	-	-	-	**.03**
Model Avg	**.53**	**.39**	**.26**	**.40**	**.19**	**.21**	**.32**	**.35**	**.12**	-	-	-	**.29**	-	-	-	-

Table 2
Performance of the models on OuliBench in both zero- and few-shot configurations according to Success Rate (SR) and Spearman correlation coefficient (rho) (only for quantitative constraints). **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 ##. Tasks for which the models were unable to generate meaningful outputs are marked with -. The task avg. are measured on 0-shot values.

Task	Claude	Deep Seek	Gemini	GPT	Task Avg
Word Length	*.09/.94*	.30/1	**.34/1**	.17/.99	**.22/.98**
Char. Length	*.003/1*	.05/.96	**.14/1**	.13/.82	**.08/.69**
Diathesis	*.89*	**1**	.99	.97	**.96**
Permutations	*.86*	1	.95	**.99**	**.95**
Lipograms	.77	.79	*.73*	**.89**	**.79**
Inverse Lipogr.	*.55*	**.93**	.89	.67	**.76**
Word Anagrams	*.54*	.76	.58	**1**	**.72**
Sent. Anagrams	.56	*.30*	**.94**	.50	**.57**
Tautograms	**.99**	*.91*	.94	.98	**.95**
Palindromes	**.74**	*.18*	.20	.26	**.35**
Acrostics	.78	.98	**1**	*.74*	**.87**
Model Avg	**.62**	**.65**	**.70**	**.66**	

Table 3
Performance of the closed-source models on OuliBench both in zero- and few-shot configurations (SR/ρ). **Best** and *worst* results for each property and metric are in **bold** and *italic*, respectively. ## indicates non-significant correlations.

whereas Velvet (1%) and Maestrale (0.06%) showed major limitations. Minerva failed entirely (0).

5.2.2. Syntactic Constraints

Diathesis control revealed in general a clear advantage for proprietary models: DeepSeek and Anita achieved near-perfect scores (100% and 99%, respectively), followed by GPT-4o mini (97%) and Gemini (99%). Claude trailed slightly (89%), while Italian open-source models—Minerva (72%), Velvet (67%), and Maestrale (59%)—struggled more.

Constituent order permutations highlighted a stark divide: GPT-4o mini excelled (99%), with DeepSeek (100%), Gemini (95%), and Claude (86%) close behind. Open-source models performed uniformly worse: Anita and Velvet (both 16%), Maestrale (12%), and Minerva (8%), suggesting architectural limitations in complex syntactic manipulation.

5.2.3. Stylistic-Formal Constraints

This category showed the widest performance gaps. For lipograms, GPT-4o mini achieved the best results (89%), ahead of DeepSeek (79%), Claude (77%), and Gemini (73%). Anita remained competitive (59%), while other open-source models obtained significantly lower scores: Velvet (47%), Maestrale (32%), and Minerva (28%).

Tautograms revealed polarizing results: Claude led (0.99), followed by GPT-4o mini (98%), Gemini (94%), and DeepSeek (91%). Among open-source models, Anita (0.73) vastly outperformed Maestrale (55%), with Velvet (8%) and Minerva (0.07%) failing critically.

Word anagrams exhibited extreme variability: GPT-4o mini scored perfectly (1.0), while Anita surprised with 92%, surpassing DeepSeek (76%), Gemini (58%), and Claude (54%). Other open-source models failed completely: Maestrale (18%), Velvet (10%), and Minerva (0).

Palindromes were universally the hardest task. Claude led (74%), with GPT-4o mini (26%), Gemini (20%), and DeepSeek (18%) far behind. Anita achieved 54% in zero-shot, while all other open-source models scored zero.

5.3. Effects of Few-Shot Learning

The few-shot learning analysis reveals non-uniform patterns across models and tasks. **Anita** shows a general degradation of performance with an increase in examples (from 53% in zero-shot to 26-40% in few-shot configurations), particularly evident in quantitative tasks where word control decreases from 27% to 13% with 15 examples, and voice control degrades from 99% to 90%. This trend suggests possible contextual overfitting phenomena.

Maestrale, on the other hand, exhibits a pattern of gradual improvement (from 19% in zero-shot to 35% with 15 examples), with clear benefits in quantitative tasks: character control improves from 0.006 to 0.03, and voice

Table 4
Example outputs for selected tasks across four Italian LLMs.

Model	Lipogram (no 'e')	Permutation (VSO)	Tautogram (letter 'm')
Anita	*"Un topo salta sopra un muro alto."* *Eng. "A mouse jumps over a high wall."* Correct lipogram	*"Suona Giulia il pianoforte."* *Eng. "Giulia plays the piano."* Correct	*"Molti mangiano molta minestra."* *Eng. "Many people eat a lot of soup."* Correct
Maestrale	*"I gatti miagolano sonnacchiosi"* *Eng. "Cats meow sleepily"* Correct lipogram	*"Mangia Luca una mela rossa"* *Eng. "Luca eats a red apple"* Correct permutation	*"Mosse machiavelliane macchiano magnanimi mandarini."* *Eng " Machiavellian moves tarnish magnanimous mandarins."* Respects the constraint but the sentence sounds contrived
Velvet	*"Il bosco era fitto e misterioso."* *Eng. "The forest was dense and mysterious."* Violates the constraint with the letter 'e'	*"Luca studia matematica."* *Eng. "Luca studies math.* Does not respect the constraint	*"Meraviglioso momento musicale regalato dall'artista magistrale della serata."* *Eng. "A wonderful musical moment provided by the masterful artist of the evening."* Does not respect the constraint in the whole sentence
Minerva	*"Io non voglio andare al mare oggi."* *Eng. "I don't want to go to the beach today."* Violates the constraint with the letter 'e'	*"Mangio un panino al salame per pranzo."* *Eng. "I eat a salami sandwich for lunch."* The subject is implied, so the sentence does not respect the correct order	*"Marito mio, mia moglie."* *Eng. "My husband, my wife."* Constraint satisfied, slightly artificial

control reaches perfection (1.0) with 5 and 15 examples. A slight improvement from 0.32 to 0.37 with 5 examples is also observed in lipograms, indicating more robust in-context learning capabilities.

It is noteworthy that **Minerva** and **Velvet** systematically tend to reproduce the few-shot examples almost verbatim, particularly in quantitative tasks and lipograms. This behavior made their outputs effectively unassessable in few-shot settings. A plausible explanation is that the high complexity of the tasks, combined with the explicit presence of in-context examples, may lead these models to default to copying strategies rather than genuine generalization. This tendency ultimately compromises output quality and originality, suggesting limitations in their ability to adapt constraints creatively beyond provided exemplars.

6. Discussion

The OuLiBench results provide valuable insights into the linguistic competence of Large Language Models (LLMs), particularly in their ability to generate text under various formal constraints. One of the most striking findings is the performance gap between tasks involving quantitative constraints and those requiring more structural or stylistic control. This disparity suggests that while LLMs exhibit a robust implicit grasp of linguistic structure, they struggle with fine-grained numerical control, a limitation likely rooted in the statistical nature of transformer architectures.

Comparing open-source and closed-source models, the latter generally outperform the former, particularly in tasks involving stylistic-formal constraints. However, this advantage is not consistent across all task types. Notably, even closed-source models, despite their overall superiority, struggle with specific tasks such as palindromes, which require strict character-level control. Similarly, tasks involving quantitative constraints pose significant challenges for both model categories, as they demand precise control over features like length or repetition, capabilities that are difficult to enforce within transformer-based architectures relying on statistical patterns rather than explicit rule-based mechanisms. These limitations further corroborate the value of OuLiBench as a benchmark for evaluating LLMs' ability to generate text while adhering to complex and diverse constraints. Finally, models from both categories perform well on syntactic constraints, suggesting that such structural aspects are relatively well captured by current architectures.

Focusing instead on smaller open-source models, we noticed that their linguistic production frequently suffered, primarily in stylistic-formal tasks, from an inability to generate truly well-structured sentences in Italian, of-

ten producing ungrammatical or semantically incoherent outputs. This degradation of linguistic quality under complex constraints highlights the trade-off between adherence to the constraint and maintenance of basic linguistic competence. A particularly notable pattern emerged in the palindrome tasks: smaller models frequently abandoned Italian and began generating sentences in English. This involuntary code-switching suggests a tendency to revert to the predominant language in the training data when the task deviates from standard generation patterns.

From a more qualitative point of view, the generated outputs of the models reveal systematic behavioral patterns, particularly evident in smaller models but also observable in larger ones. A recurring phenomenon is the tendency for thematic and lexical repetition with superficial word order variations across most tasks, suggesting limitations in creative diversification under constraints.

In the specific case of anagrammatic tasks, Anita and Velvet showed a simplified resolution strategy, limiting themselves to swapping word order within phrases rather than performing true letter-level permutations (as shown in the examples below). This behavior indicates a superficial understanding of the anagrammatic constraint and the adoption of simplified heuristics.

Examples from Anita:

> *Original:* "Tre gatti in casa fanno rumore strepito"
> *Anagram:* "Strepito in casa fanno gatti tre rumore"
> *English:* "Three cats in the house make noise and uproar" → "Uproar in the house make cats three noise"

> *Original:* "Tre per cento in banca stanno"
> *Anagram:* "Stanno in banca trecento per"
> *English:* "Three percent are in the bank" → "Are in the bank threehundred percent"

Examples from Velvet:

> *Original:* "Il sole splende."
> *Anagram:* "Splende il sole."
> *English:* "The sun shines." → "Shines the sun."

> *Original:* "La luna brilla."
> *Anagram:* "Brilla la luna."
> *English:* "The moon shines." → "Shines the moon."

> *Original:* "Il gatto mangia."
> *Anagram:* "Mangia il gatto."
> *English:* "The cat eats." → "Eats the cat."

In summary, these results highlight **the difficulty of models in reflecting and producing according to meta-linguistic principles**, a fundamental feature of human linguistic creativity, thus highlighting **the limitations of multi-objective planning mechanisms with respect to controllability and performance in complex linguistic tasks**.

7. Conclusion and Future Works

In this study, we presented OuLiBench, a novel benchmark designed to rigorously assess the linguistic capabilities of Large Language Models (LLMs) through the generation of Italian texts governed by explicit formal constraints. Drawing inspiration from the Oulipo literary tradition, our benchmark diverges from conventional evaluation methodologies that typically emphasize task performance on downstream applications. Instead, OuLiBench centers its evaluation on the model's proficiency in adhering to a diverse array of linguistic constraints, encompassing structural, quantitative, syntactic, and stylistic dimensions. This shift of focus allows for a more nuanced understanding of a model's fine-grained control over language generation processes. Our empirical evaluation involved both open-source and commercial LLMs tested in zero-shot and few-shot scenarios. The results revealed substantial variability in their ability to meet the prescribed constraints. Quantitative constraints, such as specific letter counts or palindromic structures, posed significant difficulties across the board, underscoring persistent limitations in current architectures for handling sub-lexical control. Conversely, syntactic and stylistic constraints were more successfully navigated by larger models, suggesting that model scale and complexity contribute positively to managing higher-level linguistic features. Notably, Italian-focused LLMs, including Anita, demonstrated competitive performance, highlighting the benefits of dedicated linguistic resources and targeted training on specific languages, which can partially offset the advantages conferred by sheer model size. These findings emphasize the persistent challenges in controllable text generation, especially under intersecting and mutually interacting constraints and demand simultaneous fulfillment without compromising linguistic naturalness and coherence. The results indicate a pressing need for innovative generation frameworks capable of embedding meta-linguistic reasoning and constraint-aware planning mechanisms throughout the text production pipeline. Looking forward, OuLiBench lays the groundwork for several promising directions in computational linguistics and AI research. Extending the benchmark to other languages would facilitate cross-linguistic investigations into the controllability of multilingual LLMs, while the integration of multimodal or pragmatic constraints could

broaden the scope of evaluation beyond purely textual parameters. Additionally, developing refined qualitative and creativity-focused metrics will be critical to advancing our understanding of deep linguistic competence, ultimately guiding the design of next-generation models with enhanced flexibility, expressiveness, and adherence to formal language structures. Ultimately, OuLiBench not only enriches the evaluation toolkit for Italian NLP but also serves as a conceptual bridge between computational linguistics and literary formalism, pushing the boundaries of what LLMs can achieve under constraint.

Acknowledgments

This work has been supported by the FAIR - Future AI Research (PE00000013) project under the NRRP MUR program funded by the NextGenerationEU. Partial support was also received by the project *"Understanding and Enhancing Preference Alignment in Large Language Models Through Controlled Text Generation"* (IsCc8_ALIGNLLM), funded by CINECA under the ISCRA initiative, for the availability of HPC resources and support.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 30, 2017. URL: https://arxiv.org/abs/1706.03762.

[2] OpenAI, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[3] DeepSeek-AI, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.

[4] A. G. et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[5] E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, T. Wolf, Open llm leaderboard, https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

[6] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic evaluation of language models, 2023. URL: https://arxiv.org/abs/2211.09110. arXiv:2211.09110.

[7] S. Tedeschi, J. Bos, T. Declerck, J. Hajic, D. Hershcovich, E. H. Hovy, A. Koller, S. Krek, S. Schockaert, R. Sennrich, E. Shutova, R. Navigli, What's the meaning of superhuman performance in today's nlu?, 2023. URL: https://arxiv.org/abs/2305.08414. arXiv:2305.08414.

[8] S. Levy, N. John, L. Liu, Y. Vyas, J. Ma, Y. Fujinuma, M. Ballesteros, V. Castelli, D. Roth, Comparing biases and the impact of multilingual training across multiple languages, 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. 10260–10280. URL: https://aclanthology.org/2023.emnlp-main.634/. doi:10.18653/v1/2023.emnlp-main.634.

[9] H. Zhang, H. Song, S. Li, M. Zhou, D. Song, A survey of controllable text generation using transformer-based pre-trained language models, 2023. URL: https://arxiv.org/abs/2201.05337. arXiv:2201.05337.

[10] X. Liang, H. Wang, Y. Wang, S. Song, J. Yang, S. Niu, J. Hu, D. Liu, S. Yao, F. Xiong, Z. Li, Controllable text generation for large language models: A survey, 2024. URL: https://arxiv.org/abs/2408.12599. arXiv:2408.12599.

[11] 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: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2024. Equal contribution: *.

[12] A. Miaschi, F. Dell'Orletta, G. Venturi, Evaluating large language models via linguistic profiling, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 2835–2848. URL: https://aclanthology.org/2024.emnlp-main.166/. doi:10.18653/v1/2024.emnlp-main.166.

[13] C. Ciaccio, F. Dell'orletta, A. Miaschi, G. Venturi, Controllable text generation to evaluate linguistic abilities of Italian LLMs, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 221–232. URL: https://aclanthology.org/2024.clicit-1.27/.

[14] F. Le Lionnais, La lipo (le premier manifeste), in: Oulipo (Ed.), La Littérature potentielle, Gallimard,

1973, pp. 19–22. Pubblicato originariamente in *Les Dossiers du Collège de 'Pataphysique*, n. 17 (dicembre 1961).

[15] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 707–719. URL: https://aclanthology.org/2024.clicit-1.77/.

[16] iGeniusAI, Italia-9b-instruct-v0.1, HuggingFace model card, 2024. URL: https://huggingface.co/iGeniusAI/Italia-9B-Instruct-v0.1, 9B parametri, training su testo italiano con supercomputer Leonardo (CINECA).

[17] Almawave, Velvet-14b, HuggingFace & azienda Almawave, 2025. URL: https://huggingface.co/Almawave/Velvet-14B, 14B parametri, multilingue (it, en, es, pt-BR, de, fr), training su HPC Leonardo.

[18] mii-llm, Maestrale-chat-v0.4-beta, HuggingFace model card, 2025. URL: https://huggingface.co/mii-llm/maestrale-chat-v0.4-beta, 7.2B parametri, built with Axolotl, safe chat beta.

[19] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita-8b-inst-dpo-ita, arXiv preprint, 2024. arXiv:2405.07101.

[20] Anthropic, Claude 3.7 Sonnet System Card, https://www.anthropic.com/claude-3-7-sonnet-system-card, 2025. System card for the hybrid-reasoning model Claude 3.7 Sonnet.

[21] DeepSeek-AI, A. Liu, B. Feng, B. Xue, . et al., Deepseek-v3 technical report, arXiv preprint, 2024. arXiv:2412.19437.

[22] C. Bosco, S. Montemagni, M. Simi, Converting italian treebanks: Towards an italian stanford dependency treebank, in: A. Pareja-Lora, M. Liakata, S. Dipper (Eds.), Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 61–69.

[23] L. Alfieri, F. Tamburini, (almost) automatic conversion of the venice italian treebank into the merged italian dependency treebank format, in: Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-IT 2016), CEUR Workshop Proceedings, Napoli, Italy, 2016, pp. 19–23.

[24] M. Sanguinetti, C. Bosco, Parttut: The turin university parallel treebank, in: R. Basili, C. Bosco, R. Delmonte, A. Moschitti, M. Simi (Eds.), Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project, Lecture Notes in Computer Science, Springer Verlag, Heidelberg, 2014.

[25] M. Sanguinetti, C. Bosco, A. Lavelli, A. Mazzei, O. Antonelli, F. Tamburini, Postwita-ud: an italian twitter treebank in universal dependencies, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, p. ?–? URL: https://aclanthology.org/L18-1279/.

[26] A. T. Cignarella, C. Bosco, V. Patti, M. Lai, Application and analysis of a multi-layered scheme for irony on the italian twitter corpus twittirÒ, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 4204–4211.

[27] 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/.