Structural sensitivity does not entail grammaticality: assessing LLMs against the Universal Functional Hierarchy

Tommaso Sgrizzi[1,2,*,†], Asya Zanollo[1,2,†] and Cristiano Chesi[1,2,†]

[1]University School for Advanced Studies IUSS Pavia
[2]Laboratory for Neurocognition, Epistemology, and Theoretical Syntax - NeTS-IUSS Pavia

Abstract

This paper investigates whether large language models (LLMs) generalize core syntactic distinctions associated with restructuring verbs in Italian, a domain tied to the universal hierarchy of functional heads proposed by Cinque [1, 2]. Specifically, we examine whether LLMs distinguish between restructuring and *control* verbs based on canonical syntactic diagnostics: verb ordering, clitic climbing, and auxiliary selection. We also probe how models interpret novel infinitive-selecting pseudoverbs, testing whether they default to *restructuring*- or *control*-like behavior. Using controlled minimal pairs, we evaluate five models of different sizes: Minerva-7B-base-v1.0 [3], GPT2-medium-italian-embeddings [4], Bert-base-italian-xxl-uncased [5], GPT2-small-italian [4], , and GePpeTto [6]. Our findings reveal that none of the models internalize the restructuring hierarchy, nor do they systematically block clitic climbing or select auxiliaries in line with adult grammar. These results highlight fundamental limitations in the syntactic generalization abilities of current LLMs, particularly in domains where structural contrasts are not overtly marked.

Keywords

Large language models (LLMs), Cognitive plausibility, Syntactic evaluation, Universal hierarchy of functional heads, Restructuring verbs

1. Introduction

Large language models (LLMs) have achieved remarkable success across a wide range of natural language understanding tasks, reigniting interest in their syntactic abilities and sparking a vigorous debate regarding the cognitive plausibility of the linguistic generalizations they acquire from data ([7], a.o.). Recent research has begun to probe the extent to which LLMs implicitly encode hierarchical syntactic structure [8, 9, 10], examining their sensitivity to phenomena such as long-distance dependencies and subject-verb agreement. This paper contributes to this growing body of work by investigating whether LLMs are sensitive to a crosslinguistically robust constraint governing the hierarchical distribution of functional verbs in Italian ([1, 11]). Given the broad crosslinguistic relevance of this phenomenon ([12, 13]), our investigation directly addresses the question of the coherence of linguistic structural representations in LLMs: can these models learn and represent aspects of Cinque's hierarchy from the data they are trained on? We considered two aspects: model's size and the training language, in order to observe whether, keeping the size constant, a model trained in Italian would perform better in a task specific for Italian. In terms of size, we compared larger, medium and smaller models — Minerva-7B-base-v1.0 [3], GPT2-medium-italian-embeddings [4], Bert-base-italian-xxl-uncased [5], GPT2-small-italian [4] and GePpeTto [6], to see if a greater number of parameters and training data leads to better generalization in terms of abstracting linguistic rules. The research questions (RQs) that guide this study can be framed as:

- **RQ1:** To what extent do LLMs generalize the verb ordering hierarchy proposed by Cinque (2006) for restructuring verbs?
- **RQ2:** Can LLMs differentiate the underlying structural ambiguity inherent in restructuring versus control verb constructions?
- **RQ3:** What is the syntactic structure assigned by LLMs to novel verbs which introduce non-finite complements?

For instance, as far as RQ1 is concerned, the following contrast shows that the incorrect hierarchical order — which, in this case, directly affects the linear order — of *provare* 'try' ($\text{Asp}_{\text{Conative}}$) and *volere* 'want' ($\text{Mod}_{\text{Volition}}$) leads to ungrammaticality.

(1) a. *Gianni lo vuole provare a riparare.*
Gianni it.CL wants to try to fix

'Gianni wants to try to fix it.'

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
†These authors contributed equally.
tommaso.sgrizzi@iusspavia.it (T. Sgrizzi); asya.zanollo@iusspavia.it (A. Zanollo); cristiano.chesi@iusspavia.it (C. Chesi)
https://tomsgrizzi.github.io/ (T. Sgrizzi)
0000-0003-1375-1359 (T. Sgrizzi); 0009-0001-3987-4843 (A. Zanollo); 0000-0003-1935-1348 (C. Chesi)

b. *Gianni lo prova a voler riparare.
Gianni it.CL tries to wants to fix

Intended: 'Gianni tries to want to fix it.'

Regarding RQ2, only restructuring verbs allow clitic climbing (2) and auxiliary switch (3), as shown in the example below.

(2) a. Gianni lo comincia a riparare.
Gianni it.CL begins to fix

'Gianni begins to fix it.'

b. *Gianni lo corre a riparare.
Gianni it.CL runs to fix

'Gianni runs to fix it.'

(3) a. Gianni ha/è voluto partire.
Gianni has/is wanted to

'Gianni wanted to leave'

b. *Gianni ha/*è preferito partire.
Gianni has/*is preferred to

'Gianni preferred to leave'

Finally, RQ3 can be investigated through the syntactic ingredients laid out above, using both clitic climbing and auxiliary switch as diagnostics for a restructuring-like, or a control-like representation of infinitive-taking verbs. Consider a pseudo-verb like *grabbare*, if models have clear the difference between restructuring and control, they would either block or allow clitic climbing across it, and either block or allow auxiliary switch.

In the next section, we will introduce the empirical domain of restructuring and the relevance of the cartographic enterprise as valid heuristics to test the cognitive plausibility of syntactic generalizations.

2. Universal Functional Hierarchy

In formal linguistics, the cartographic approach refers to the effort to systematically map out the functional structure of the clause. Much like a geographical map reveals detailed topography, syntactic cartography seeks to uncover the fine-grained architecture of language, identifying a universal and richly articulated hierarchy of functional projections that determine the order of constituents in natural language [14]. This enterprise, developed over the past three decades, has shown striking cross-linguistic consistency: while surface word orders vary dramatically across languages, the underlying structural relations often conform to highly constrained and universal hierarchies. For instance, across typologically diverse languages, adverbs and verbal morphology appear in a constrained order that reflects an underlying sequence of functional heads encoding modality, aspect, tense, and voice [2]. A well-known example involves the relative positions of epistemic and aspectual adverbs. Consider the following contrast.

(4) a. *John **probably** has **again** read the book.*
b. **John **again** has **probably** read the book.*

This contrast reflects a deeper generalization: epistemic adverbs like *probably* structurally precede aspectual adverbs like *again* in the functional hierarchy [2]. This ordering is also mirrored in other languages, such as Italian (*Giovanni probabilmente ha di nuovo letto il libro* vs. ?*Giovanni di nuovo ha probabilmente letto il libro*), and even when surface word orders vary, (constrained) movement analyses do preserve the underlying hierarchy. In fact, attested orders tend to be derivable from the base sequence via movement operations constrained by Universal Grammar, while unattested orders — such as stacking adverbs in reverse (*again* > *probably*) are rarely, if ever, observed without resulting in degraded acceptability (see also [15, 16, 17] for a different view on ordering constraints yet still rooted in cognitive principles).

Similarly, in the nominal domain, elements such as demonstrations, numerals, adjectives, and nouns tend to conform to the base order Demonstrative > Numeral > Adjective > Noun [18]. Using English again for illustration, the sequence *those three books* is allowed, but not *red three those books*. These generalizations suggest that natural languages are not arbitrarily diverse but instantiate a shared blueprint with tightly delimited variation, a claim supported by decades of comparative research [19, 20, 21, 22].

Crucially, these cartographic universals are not merely typological observations; they reflect deep structural constraints on human language, likely rooted in cognitive and interface-driven pressures such as learnability, interpretability, and communicative efficiency (see a.o. [23, 24, 25]). As such, they offer a highly structured benchmark for evaluating whether LLMs reflect the underlying principles of natural language cognition or simply reproduce surface-level statistical patterns. Assessing cartographic generalizations in LLMs thus becomes another valuable diagnostic tool for determining whether their internal representations exhibit the kind of compositional and hierarchical structure found in human language.

Importantly, the utility of cartographic diagnostics does not presuppose that LLMs use the same mechanisms as human language acquisition. Instead, it positions cartographic constraints as a structural target: a gold standard against which to assess the depth of linguistic generalization in artificial systems. If LLMs are to be considered cognitively plausible models of language

([26], a.o.), they should, at a minimum, capture the universal constraints that human learners internalize from fragmented, language-specific input. Testing for cartographic effects in LLMs therefore offers a window into the extent to which their representations are not only successful at surface prediction but aligned with the hidden universals that define natural language competence. In this sense, cartography closes the gap between linguistically informed evaluation and cognitively grounded modeling. By operationalizing syntactic universals as testable hypotheses in LLMs, we move closer to understanding not just whether these models can generate human-like language, but whether they have abstracted the kinds of structure that make human language what it is.

2.1. The empirical domain: the case of restructuring verbs in Italian

A particularly revealing case study for testing structural representations from a cartographic perspective in LLMs comes from the domain of restructuring verbs in Italian, as discussed in [1, 11]. Restructuring verbs — such as *potere* 'can', *dovere* 'must', *volere* 'want', *continuare* 'continue', *cominciare* 'begin', are verbs that, despite selecting an infinitival complement, do not behave as if they embed a full clause (cf. [13, 12, 27], a.o.). Instead, they participate in a monoclausal structure, lacking the full complement of functional projections found in fully embedded (i.e., biclausal) contexts. This has observable syntactic consequences: only restructuring verbs permit movement of the object clitic from the complement position of the infinitive up to the matrix verb (e.g., *Marco lo vuole mangiare* 'Marco wants to eat it'), while control verbs, which are superficially similar, do not (e.g., **Marco lo decide di mangiare* 'Marco decides to eat it'). Clitic placement (Clitic Climbing; CC) thus offers a fruitful diagnostic for the underlying syntactic structure of a restructuring configuration.

More specifically, the working hypothesis that we are adopting here ([1, 11]) views restructuring verbs as functional heads occupying a fixed hierarchy (e.g., from lower to higher, Aspectual > Modal > Temporal), with each verb spelling out a specific functional projection (Fig. 1) rooted in the cartographic representation of the inflectional domain.

Restructuring verbs obey in fact strict ordering constraints within sequences: for example, *Marco lo suole voler mangiare spesso* 'Marco usually wants to eat it often' is grammatical, while reversing the restructuring verbs blocks clitic climbing (**Marco lo vuole soler mangiare spesso*) as it is a violation of the hierarchical sequence of functional heads (*$\text{Mod}_{\text{Volition}} > \text{Asp}_{\text{Frequentative}}$). Unlike linear word orders of adjectives or adverbs, which LLMs might learn through surface-level statistical regularities,

$\text{MoodP}_{\text{speech act}} > \text{MoodP}_{\text{evaluative}} > \text{MoodP}_{\text{evidential}} > \text{MoodP}_{\text{epistemic}} > \text{TP(Past)} > \text{TP(Future)} > \text{MoodP}_{\text{irrealis}} > \text{ModP}_{\text{aletic}} > \text{AspP}_{\text{habitual}} > \text{AspP}_{\text{repetitive(I)}} > \text{AspP}_{\text{frequentative(I)}} > \text{ModP}_{\text{volitional}} \; \text{AspP}_{\text{celerative(I)}} > \text{TP(Anterior)} > \text{AspP}_{\text{terminative}} > \text{AspP}_{\text{continuative}} > \text{AspP}_{\text{retrospective}} \; \text{AspP}_{\text{proximate}} > \text{AspP}_{\text{durative}} > \text{AspP}_{\text{generic/progressive}} > \text{AspP}_{\text{prospective}} > \text{ModP}_{\text{obligation}} \; \text{ModP}_{\text{permission/ability}} > \text{AspP}_{\text{completive}} > \text{VoiceP} > \text{AspP}_{\text{celerative(II)}} > \text{AspP}_{\text{repetitive(II)}} > \text{AspP}_{\text{frequentative(II)}}$

Figure 1: [1]:12

these restructuring configurations constrain deeper syntactic dependencies. Besides CC, restructuring verbs like *potere* 'can', *volere* 'want', and *dovere* 'must', can in fact optionally allow the infinitival verb to pick the auxiliary (*essere* 'be', or *avere* 'have'), as in the case of unaccusative verbs.

(5) *Marco ha/è dovuto partire.*
Marco has/is must.PSTPRT leave.INF
Marco had to leave.

Restructuring verbs then present an ideal testing ground for evaluating whether LLMs encode abstract syntactic structures from cartographic generalizations, or merely track co-occurrence frequencies. While *Marco lo finisce di mangiare in fretta* ('Marco finishes eating it quickly') is structurally monoclausal and allows clitic climbing, its control verb counterpart **Marco lo decide di mangiare in fretta* is ungrammatical precisely because the clitic cannot climb out of a true embedded clause. These subtle distinctions, masked by similar surface forms, reflect two different structural representations, underscoring the need to go beyond linearity when assessing syntactic competence in artificial models. Furthermore, evidence from language development [28] shows that the distinction between restructuring and control syntax, and the fixed ordering constrain of restructuring verbs, are acquired very early on. This suggests that children have a clear representation of the difference between control and restructuring verbs, and when encountering a novel infinitive-taking verb, some preliminary corpus data suggest they tend to prefer a restructuring interpretation over a control one [29]. A natural question, then, is whether LLMs also encode such a clear distinction when processing previously unseen infinitive-taking verbs. In summary, we can use at least three solid tests to probe linguistic competence when comparing restructuring and control verbs: (i) the first (restructuring), but not the second (control), allows Clitic Climbing (CC); (ii) the order of predicates lexicalizing positions in the functional hierarchy is rigid; and (iii) restructuring predicates can take both *be* and *have* as auxiliaries.

3. Generalization in LLMs

Despite the impressive performance of state-of-the-art LLMs, it remains an open question whether their enhanced predictive capabilities reflect genuine syntactic knowledge. LLMs are said to exhibit syntactic generalization insofar as they can abstract structural rules from data and apply them to novel grammatical contexts beyond their training input. Wilson et al. (2023) [30] theorize three forms of generalization, differentiating the ability to learn word distributions and the distributions in contexts from the ability to abstract generalization independently of training data. The findings highlight that, while excelling in transferring distributions across syntactically similar context, LLMs struggle in extracting structural hierarchical rules, relying primarily on linear order instead. Accordingly their linguistic knowledge appears to be of a semantic and probabilistic nature and the emergence of human-like abstraction correlates with the increase of training data, radically differentiating from human linguistic competence. The issue of LLMs's grammatical knowledge is approached in the linguistic community through different approaches relying on controlled experimental settings, probing LLMs' performances on minimal pair sentences, and evaluating the internalization of deep hierarchical dependencies of the underlying linguistic structures. Blimp [31] evaluate LLMs with minimal pairs finding that while learning basic dependencies, surface-level patterns, models still cannot encode universal constraints like the argument structure, even a high-resource language like English. Training models on larger corpora leads to better performances suggesting that data play a major role compared to the architecture.

The very same result is obtained in another benchmark, BIG-bench [32], comprising 204 tasks designed to assess linguistic, reasoning, and knowledge-based abilities. Even if larger models show an improvement in syntactic generalization, this can be explained in terms of memorization rather than grammatical abstraction. Deep-structure constraints still represent a challenge.

In a recent study, [33] confirms the relevance of training data size in improving generalization, taking the case of a syntactic universal as the Final-over-Final Condition (FOFC) - the rule governing the structural organization crosslinguistically. They tested models with low-resource languages and found that models fail to learn this constraint when dealing with languages like Basque. A superhuman amount of training examples improves syntactic generalization, but models do not acquire abstract rules of grammar.

Taken together, these studies point to the necessity of incorporating more structured training methodologies and inductive biases, especially in light of the fact that human language acquisition occurs with far less data. Current models remain fundamentally data-dependent rather than rule-based, and simply increasing the scale of training does not really improve the possibility of true syntactic generalizations.

In this context, the empirical domain of restructuring verbs provides an ideal testing ground for disentangling linear generalizations from structural rules. On the one hand, restructuring verbs follow specific linear orderings that could, in principle, be learned from surface patterns in the training data. On the other hand, their ordering can either permit or block syntactic phenomena such as clitic climbing (CC), making linear order a surface reflex of deeper structural constraints. Capturing the relevant syntactic generalizations in this domain therefore requires more than sensitivity to word order — it demands an understanding of the underlying hierarchical structure.

4. Methods

We designed 13 minimal pairs experiments targeting various grammatical contrasts involving clitic placement, auxiliary selection, and verb-verb complementation. In these experiments, we manipulated the presence or absence of restructuring environments, the type of matrix verb (restructuring verbs, *control* verbs, and pseudo-verbs), and the structural distance between multiple occurrences of restructuring verbs, allowing us to probe the models' syntactic representations under different conditions. First, we coded 14 restructuring verbs and 14 infinitive-taking verbs (which we name according to the syntactic literature as *control* verbs, cf. [34]). While the coding of *control* verbs is arbitrary, the numbering of restructuring verbs reflects their position in the functional hierarchy of [1], with *andare* 'to go' assigned code 1 as the lowest verb, and *solere* 'to be used to' assigned code 14 as the highest (see Table. 1). Verbs higher in the hierarchy occur linearly left of lower verbs.

In addition to the verbs above, we also created three pseudo-verbs (i.e., non-existent words in Italian) to test whether LLMs assign them a restructuring-like or *control*-like syntactic representation when they take a non-finite complement. One, *grabbare*, is a bare verb resembling modals (verbs 6, 7, and 12 in Table 1) as well as *solere* 'to be used to' and other control verbs. The other two pseudo-verbs, *drommare a* and *trellare di*, take the prepositions *a* and *di*, respectively: a feature shared with the remaining restructuring and control verbs.

To address RQ1 (introduced in Section §1), we constructed minimal pairs of verb sequences that either respect or violate Cinque's (2006) functional hierarchy. Each item in Exp. 1 presents a grammatical (hierarchy-respecting) sentence alongside a minimally different ungrammatical counterpart, with the two verbs separated

Table 1
List of restructuring and *control* verbs used across conditions, *Functional Projection* refers to restructuring verbs

Code	Restructuring verb	Functional Projection	*Control* verb
1	*andare a* 'to go'	$\text{Asp}_{\text{Andative}}$	*correre a* 'to run'
2	*cominciare a* 'to begin'	$\text{Asp}_{\text{Inceptive}}$	*salire a* 'to go up'
3	*finire di* 'to finish'	$\text{Asp}_{\text{Completive}}$	*dire di* 'to say'
4	*provare a* 'to try'	$\text{Asp}_{\text{Conative}}$	*scendere a* 'to go down'
5	*riuscire a* 'to succeed'	$\text{Asp}_{\text{Success}}$	*osare* 'to dare'
6	*potere* 'can'	$\text{Mod}_{\text{Ability}}$	*preferire di* 'to prefer'
7	*dovere* 'must'	$\text{Mod}_{\text{Obligation}}$	*desiderare* 'to wish'
8	*stare per* 'to be about'	$\text{Asp}_{\text{Prospective}}$	*promettere di* 'to promise'
9	*continuare a* 'to continue'	$\text{Asp}_{\text{Continuative}}$	*decidere di* 'to decide'
10	*smettere di* 'to stop'	$\text{Asp}_{\text{Terminative}}$	*chiedere di* 'to ask'
11	*volere* 'want'	$\text{Mod}_{\text{Volition}}$	*pensare di* 'to think'
12	*tornare a* 'to come back'	$\text{Asp}_{\text{Iterative}}$	*credere di* 'to believe'
13	*tendere a* 'to tend	$\text{Asp}_{\text{Predisp}}$	*sperare di* 'to hope'
14	*solere* 'to be used to'	$\text{Asp}_{\text{Habitual}}$	*scegliere di* 'to chose'

by varying degrees of hierarchical distance. This experiment tests whether LLMs prefer the option adhering to the hierarchy, and whether their preferences correlate with the hierarchical distance between verbs.

A second experiment (Exp. 2) uses the same verb pairs as in Exp. 1, but includes a proclitic clitic in each sentence. This introduces an explicit syntactic cue for restructuring, allowing us to evaluate whether clitic placement influences the model's preference for the grammatical, hierarchy-respecting variant..

To address RQ2, Exp. 3 and Exp. 4 pair *control* verbs with restructuring verbs, testing them in both possible orders: restructuring+control (Exp. 3) and control+restructuring (Exp. 4). Each minimal pair includes clitics, with the grammatical variant displaying enclisis on the infinitival verb and the ungrammatical one displaying proclisis onto the matrix verb. The latter is ruled out because in both cases the *control* verb introduces a clausal boundary that blocks clitic climbing.

To investigate RQ3, we conducted a series of experiments pairing restructuring and *control* verbs with the three pseudo-verbs introduced earlier. Exp. 5 combines each of the three pseudo-verbs (*grabbare, drommare a, trellare di*) with all 15 restructuring verbs, presenting two variants per item: one with proclisis onto the matrix verb (suggesting restructuring), and one with enclisis on the infinitival verb. Exp. 6 reverses the order (restructuring + pseudo-verb) but otherwise follows the same design. Since proclisis requires a monoclausal analysis, these experiments test whether the model treats novel verbs as compatible with restructuring. A systematic preference for the proclitic variant would suggest that the model generalizes restructuring behavior to unseen verbs.

Exp. 7 and Exp. 8 approach the same question from the opposite angle, pairing pseudo-verbs with *control* verbs. In Exp. 7, the order is *control* + pseudo-verb, while in Exp. 8, it is pseudo-verb + *control*. In both cases, only the enclitic variant is grammatical because *control* verbs block clitic climbing, even if the model assumes the pseudo-verb to be restructuring-compatible. This design offers a strong test of whether the model robustly distinguishes restructuring from *control* verbs. If the model is sensitive to this contrast, it should reject the proclitic variant in favor of enclisis, indicating a fine-grained syntactic representation of clitic domain boundaries.

Exp. 9 further probes the syntactic status of pseudo-verbs by pairing them with each other and testing proclitic vs. enclitic placement. This experiment asks whether the model classifies pseudo-verbs as restructuring-like or *control*-like when they co-occur, shedding light on whether it generalizes clitic behavior within novel verb classes.

In Exp. 10, we tested pseudo-verbs in isolation, assessing model preferences for auxiliary selection (*have* vs. *be*) — another syntactic hallmark of restructuring (see §2.1). For comparison, Exp. 13 and Exp. 14 extend this test to restructuring (modals) and *control* verbs, respectively.

Exp. 11 tests pseudo-verbs selecting infinitival complements, presenting both proclitic and enclitic variants. This experiment investigates whether the model prefers proclisis (indicating a restructuring representation, along the lines of Exp. 5) or enclisis, and whether this preference is modulated by the presence or absence of the prepositions *di* and *a*.

Finally, in Exp. 12 and 13 we tested modal (restructuring) verbs and *control* verbs with auxiliary selection, respectively (only modals allow both *essere* 'to be' and *avere* 'to have' with unaccusative verbs, while *control* verbs require *avere*). This allows us to see whether the fine-grained syntactic distinctions between restructuring and *control* have been successfully generalized by these models.

Table 2
Minimal pair generated examples

Group	Good sentence	Bad sentence
Exp. 1	il marinaio continua a riuscire a mandare il pesce	il marinaio continua a tendere a inviare il pesce
Exp. 2	l'esploratore lo può riuscire a toccare	l'esploratore lo può stare per toccare
Exp. 3	il ballerino va a salire a guardarlo	il ballerino lo va a salire a guardare
Exp. 4	il golfista chiede di andare a registrarlo	il golfista lo chiede di andare a registrare
Exp. 7	il viaggiatore scende a trellare di odiarlo	il viaggiatore lo scende a trellare di odiare
Exp. 8	il gestore dromma a ordinare di inviarlo	il gestore lo dromma a ordinare di inviare
Exp. 13	il principe ha preferito venire	il principe è preferito venire
	Proclitic option	Enclitic option
Exp. 5	il gioielliere lo grabba andare a vendere	il gioielliere grabba andare a venderlo
Exp. 6	il gestore lo va a drommare a disinfettare	il gestore va a drommare a disinfettarlo
Exp. 9	il sarto lo trella di grabbare rifiutare	il sarto trella di grabbare rifiutarlo
Exp. 11	l'anziano lo grabba lavare	l'anziano grabba lavarlo
	HAVE auxiliary	BE auxiliary
Exp. 10	il pugile ha grabbato discendere	il pugile è grabbato discendere
Exp. 12	il golfista ha dovuto crescere	il golfista è dovuto crescere

4.1. Materials: Minimal Pairs

The minimal contrasts exemplified in Table 2 have been considered. For each condition internal to each experiment, we generated 100 structurally irrelevant variants displaying different lexical items as subjects, infinitival verbs, and objects (when present). Although some of the items across the experiments were semantically odd, the generalizations are nonetheless still strong, and the contrast within the pairs remains sharp, as in the example below, from Exp. 2.

4. il calciatore lo sta riuscendo a finire di ideare
   the soccer player it.CL is about to be able to finish to design
5. *il calciatore lo riesce a star finendo di ideare
   the soccer player it.CL is able to be about to finish to design

The script responsible for the generation of the minimal pairs is available on GitHub.

4.2. Experiments

Five LLMs have been employed for the evaluation of syntactic generalization with minimal pair sentences. The selection was driven by two key factors for the evaluation: model size and language of training.

Correspondingly we included large, medium and small models - Minerva-7B-base-v1.0, GPT-2 medium and Bert-base-italian-xxl-uncased, GPT2-small and GePpeTto. All models are trained on Italian corpora, hence they allow us to assess whether exposure to Italian during training enhances syntactic generalization in a typologically relevant domain. This setup enables a direct comparison between different models' size, in the ability to internalize the structural dependencies necessary to abstract the relevant generalizations. All models are available on Hugging Face [35, 36, 5, 37, 38].

Minerva-7B-base-v1.0 [3] is a causal LLMs with 7 billion parameters, based on Mistral architecture (32 layers, hidden size 4096, 32 attention heads, context window of 4096 tokens) trained on $\tilde{2}$.48 trillion tokens (1.14T Italian, 1.14T English, 200B code) and a 51200-token vocabulary.

Bert-base-italian-xxl-uncased is the Italian version of BERT base model (uncased), a masked LLMs trained on next sentence prediction. The models has 111M parameters and training data consist of OPUS corpus (https://opus.nlpl.eu/) extended with additional content from the Italian portion of the OSCAR corpus, for a final training corpus of 81GB and 13,138,379,147 tokens.

GroNLP/GPT2-medium-italian-embeddings [4] is built on GPT-2 medium architecture, with 359M parameters with the lexical layer retrained to support Italian.

GroNLP/GPT2-small-italian [4] is a smaller causal Transformer with 121 million parameters, built on GPT-2 small architecture and retrained in Italian.

GePpeTto [6] has a GPT2-small configuration ($\tilde{1}$17 million parameters) and has been trained in Italian corpora - OSCAR (https://huggingface.co/datasets/oscar-corpus/oscar?utm_source=chatgpt.com), PAISA (https://www.corpusitaliano.it/en/?utm_source=chatgpt.com), Wikipedia. GePpeTto, similarly based on the GPT2-small architecture, employs a BPE tokenizer with a reduced vocabulary of 30,000 tokens, specifically adapted for Italian linguistic data.

4.2.1. LLMs Evaluation

The *LM-eval* platform [39] was adopted to perform minimal pair tests. A total of 610,500 minimal pairs were generated and divided into 13 groups, as described in §4.1, and assessed by all the selected models. For each experiment we computed the mean accuracy and standard deviation (3), leaving further statistical analyses for the future. For unknown reasons, some models failed to complete certain evaluation tasks without producing any intelligible error messages.

5. Results

We organize our results around the three core research questions that reflect different dimensions of the models' syntactic generalizations with respect to restructuring verbs, *control* verbs, and infinitive-selecting pseudoverbs. For each question, we present the relevant experimental conditions and summarize the performance of all tested LLMs in terms of mean accuracy and standard deviation. To assess whether models internalize the syntactic hierarchy of restructuring verbs proposed by [1] (RQ1), Exp. 1 and 2 tested sequences of two restructuring verbs in the correct vs. incorrect hierarchical order, with and without clitic pronouns. Mean accuracies in these experiments were consistently low (Minerva: 36–37%, GePpeTto: 36–38%, GPT2: 46–48%), with SD close to 0.5. BERT, however, performed moderately above chance (Exp. 1: 64.6%, Exp. 2: 56.9%), suggesting that it may encode some sensitivity to hierarchical ordering, although not robustly.

Mean accuracies in these experiments were consistently low (Minerva: 36–37%, GePpeTto: 36–38%, GPT2-Small: 46–48%, GPT2-Medium: 57-55%, BERT: 64-56%), with SD close to 0.5, indicating a near-random distribution of choices. Consistent with prior models, BERT's mean accuracy on these tasks was moderate but above chance (Exp. 1: 64.7%, Exp. 2: 57.0%) with substantial variability (Std Devs 0.48–0.50). This suggests some sensitivity to hierarchical ordering, though performance remains far from ceiling and does not indicate robust hierarchical knowledge. The presence of clitics in Exp. 2 did not alter model behavior compared to Exp 1. Models show no evidence of having acquired the hierarchical layout of restructuring verbs, besides BERT's results. However, their responses may correlate with verb distance or hierarchical ordering, which we leave for further research.

To evaluate whether models distinguish between restructuring and control verbs based on syntactic diagnostics (RQ2) we considered two diagnostics: clitic climbing (Exp. 3, 4, 5, 6, 7, 8, 9, 11), and auxiliary switch (Exp. 10, 12, 13). In Exp. 3 and 4, which tested restructuring–*control* verb sequences with clitics, models consistently failed to block clitic climbing where it was expected to be ungrammatical. GePpeTto and Minerva almost systematically chose the ungrammatical option (18–28%), while GPT2-small showed slightly better performance (42–46%) but with high variability, BERT performed near floor (5–6%). A model that shows a bias over 75/80% can be in fact considered structurally coherent, even though it picks the ungrammatical option [40].

In Exp. 7 and 8, which paired *control* verbs with pseudoverbs, models again failed to systematically block clitic climbing. GPT2 reached 48–57% accuracy, while GePpeTto and Minerva remained well below chance (12–18%).

In Exp 9 and 11, which included only pseudoverbs, GePpeTto consistently preferred proclitic constructions (low accuracy = proclisis favored), while Minerva and GPT2-small showed no clear preferences, again reflecting indecision or inconsistency.

As for auxiliary selection, the results reveal further lack of syntactic differentiation: in Exp 10, GePpeTto systematically selected *essere* (7% accuracy), suggesting it interpreted pseudoverbs as restructuring verbs. GPT2-small showed more balanced choices (47%), compatible with the ambiguity characteristic to some restructuring verbs which allow both *avere* and *essere*.

In Exp 12, in fact, testing modal auxiliaries, models should ideally show 50% accuracy, given the optionality of auxiliary selection; instead, both GPT2-small and GePpeTto showed categorical but divergent choices, with accuracies around 5%.

In Exp. 13 (*control* verbs), only Minerva performed above chance (57%), while GePpeTto and GPT2-small selected the incorrect auxiliary (*essere*) almost categorically (1% accuracy), and BERT was the only model to outperform Minerva (63%).

As a result, models largely fail to generalize the syntactic constraints of restructuring and *control* verbs. Clitic climbing is not consistently blocked by *control* verbs, and auxiliary selection does not reliably reflect the transparency effects typical of restructuring verbs nor the ambiguity intrinsic to them. Only GPT2-small shows partial sensitivity in some control constructions, while GePpeTto tends toward an overgeneralization of restructuring syntax (e.g. by overselecting *essere* as an auxiliary).

Finally, a central question of this study addresses how models categorize pseudoverbs — novel verbs not seen during training but constructed to select infinitival complements, and whether they are interpreted as control or restructuring verbs.

In Exp. 5 and 6, pseudoverbs appeared in sequences with restructuring verbs, with proclitic vs. enclitic alternations. Minerva showed a slight preference for the enclitic form (23–29% accuracy), suggesting a bias toward *control*-like syntax. GePpeTto strongly preferred the proclitic form (17% accuracy = 83% proclisis), indi-

Table 3
Mean and standard deviation of model accuracy across experiments, – indicates that the model failed to complete the subtask.

Experiment	UID	Minerva		GePpeTto		GPT2-Small		GPT2-Medium		BERT	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Exp. 1	sequence of two restructuring verbs testing only linear order	0.3646	0.4813	0.3593	0.4798	0.4584	0.4983	0.5682	0.4953	0.6465	0.4781
Exp. 2	sequence_pairs_with_clitics	0.3762	0.4844	0.3825	0.4860	0.4813	0.4997	0.5531	0.4972	0.5699	0.4951
Exp. 3	restructuring_and_control_plus_clitics	0.2854	0.4516	0.1867	0.3897	0.4261	0.4945	0.6831	0.4653	0.0553	0.2286
Exp. 4	control_and_restructuring_plus_clitics	0.2253	0.4178	0.1893	0.3917	0.4659	0.4988	0.6344	0.4816	0.0578	0.2333
Exp. 5	pseudo_and_restructuring_plus_clitics	0.2336	0.4231	—	—	—	—	0.5871	0.4924	0.1371	0.3440
Exp. 6	restructuring_and_pseudo_plus_clitics	0.2857	0.4518	0.1686	0.3744	0.4188	0.4934	0.6562	0.4750	0.1140	0.3179
Exp. 7	control_and_pseudo_plus_clitics	0.1569	0.3637	0.1250	0.3307	0.4798	0.4996	0.6210	0.4852	0.1081	0.3105
Exp. 8	pseudo_and_control_plus_clitics	0.1810	0.3850	0.1267	0.3326	0.5752	0.4943	0.5952	0.4909	0.1643	0.3706
Exp. 9	pairs_of_pseudo_verbs_plus_clitics	0.5583	0.4966	0.1533	0.3603	0.5300	0.4991	0.5083	0.5003	0.1817	0.3859
Exp. 10	auxiliary_switch_with_pseudoverbs	—	—	0.0700	0.2551	0.4700	0.4991	0.3300	0.4710	0.5367	0.4995
Exp. 11	pseudo_verbs_plus_clitics	0.2267	0.4187	0.1100	0.3129	0.5000	0.5000	0.5533	0.4980	0.1433	0.3510
Exp. 12	auxiliary_switch_with_modals	—	—	0.0533	0.2247	0.0500	0.2179	0.0100	0.0997	0.4833	0.5006
Exp. 13	auxiliary_switch_with_control_verbs	0.5700	0.4951	0.0100	0.0995	0.0100	0.0995	0.0000	0.0000	0.6267	0.4845

cating a restructuring-like interpretation. GPT2-small was ambivalent. Since the three pseudoverbs differ in whether they select a preposition, mirroring the variation found among restructuring verbs, further analyses will investigate this property as a potential factor.

Exp. 9 and 11, which tested proclitic/enclitic preferences with pseudoverb–pseudoverb sequences, reinforced these trends: GePpeTto showed a consistent preference for proclitic constructions (11–15% accuracy), while GPT2-small and Minerva again showed no strong preference.

In Exp. 10, which tested auxiliary selection with pseudoverbs, GePpeTto again opted overwhelmingly for *essere*, consistent with restructuring behavior, while GPT2-small distributed responses more evenly. BERT distributed its choices roughly evenly (53.7% accuracy), suggesting some awareness of optionality, though this may be an artifact of random choice.

These results suggest that GePpeTto interprets novel infinitive-selecting verbs as restructuring verbs by default (although without expressing the available optionality with *avere*), consistently favoring proclisis and auxiliary *essere*. In contrast, GPT2-small and Minerva exhibit uncertainty or mixed behavior, with no consistent syntactic categorization of pseudoverbs.

6. Discussion

Overall, the findings reveal that the models' behavior does not align with the predictions raised by the framework of [1], nor with the grammatical requirements characteristic of the syntax of non-finite complements in Italian. Instead, their choices are often inconsistent, insensitive to syntactic structure, or driven by superficial factors. The first research question addressed whether models generalize the hierarchical structure of restructuring verbs as observed in the syntactic literature ([1, 11]). Our results clearly indicate that no such hierarchy is reflected in the models' performance. Accuracies were consistently low, and variability high. These findings echo previous results showing that LLMs often fail to internalize syntactic hierarchies when such structures are not directly observed during training or explicitly encoded [30]. Even BERT, which slightly outperformed other models on restructuring verb order, failed across the board on clitic-related diagnostics. This has implications for how much syntactic theory — especially fine-grained distinctions like cartographic hierarchies — is learnable from surface patterns alone.

In the second set of questions, we tested whether models are able to handle clitic climbing and auxiliary selection, two classical diagnostics that distinguish restructuring from *control*. Across all clitic-related experiments, models consistently failed to block clitic climbing where it should be ungrammatical, especially in the presence of control verbs. This strongly suggests that models do not encode the syntactic opacity of *control* verbs. A potential explanation for these results lies in tokenization artifacts. Unlike proclitic clitics (e.g., *lo ha visto* 'it.OBJ has seen'), enclitics (e.g., *vederlo* 'see-it.OBJ') should be tokenized as subword fragments. If models fail to treat enclitics as distinct morphemes, this may increase their preference for proclitic constructions simply because the latter are tokenized as independent words, easily recognizable as syntactic objects.

Auxiliary selection patterns further support the view that models lack a deep representation of infinitive-taking verb classes. None of the models consistently mapped *control* verbs to *avere*, or correctly captured the optionality of auxiliary selection in modals (with the partial exception of BERT in Exp. 13, having 63% accuracy). GPT2 again performed marginally better than the others in preserving optionality, but even it failed to align with the expected 50% distribution. Surprisingly, both GPT2-small and GePpeTto nearly categorically misassigned *essere* to control verbs, a highly ungrammatical option in Italian.

These findings point to a broader issue: models do not reliably encode the syntactic transparency of restructuring verbs nor the obligatory opacity of control verbs.

Syntactic features that are not overtly marked in surface form — such as whether a verb transmits argument structure or allows clitic climbing — appear to be difficult for models to capture, even when such distinctions are central to grammaticality.

7. Conclusions

This study investigated whether LLMs encode abstract syntactic generalizations by testing their sensitivity to the restructuring verb hierarchy in Italian. Using a suite of controlled minimal pair experiments targeting verb order, clitic placement, and auxiliary selection, we assessed models' ability to capture structural dependencies that go beyond linear surface patterns.

The models tested — GPT2-small-italian, GPT2-medium-italian-embeddings, GePpeTto, Bert-base-italian-xxl-uncased and Minerva-7B-base-v1.0 — showed limited sensitivity to the syntactic hierarchy of restructuring verbs, failed to consistently distinguish restructuring from *control* verbs based on key syntactic diagnostics, and did not consistently categorize novel infinitive-taking verbs based on the non-finite embedding typology available in Italian. These findings highlight fundamental limitations in the syntactic abstraction capacities of current models, particularly in domains where structural contrasts are not overtly marked in surface form.

While none of the models fully internalize the hierarchical structure of restructuring verbs, some results (as BERT's above-chance accuracy in distinguishing hierarchy-respecting sequences in Exp. 1) suggest at least some limited sensitivity to structural cues. However, this sensitivity is neither robust nor consistent across models or conditions, and most importantly does not translate into reliable grammaticality judgments. For example, clitic placement's explicit cues for restructuring failed to improve performance, and models consistently failed to block ungrammatical clitic climbing or the *essere* auxiliary selection in the context of control verbs. These findings indicate that, to the extent models are sensitive to structural hierarchies, in the domain of cartographic generalizations this sensitivity remains shallow and insufficient for capturing the related grammatical distinctions.

Addressing these limitations will require new approaches to model design, training, and evaluation that go beyond surface-level pattern recognition, and may involve encoding linguistic biases into model architectures—much like cartographic hierarchies are hypothesized to be innately hardwired in human cognition.

8. Limitations

The main limitation of the current research lies in the exclusive usage of publicly available pre-trained models as outlined in 4.2. To obtain a fine-grained understanding of models' capacity on syntactic generalization, future works will employ models trained from scratch, with a training regimen reproducing human language acquisition stages (see 2). The alignment between learning trajectories and the implementation of more structured training methodologies and inductive biases (see 3) will hopefully improve models' performance in syntactic tasks [41, 42]

Moreover, we are in the process of designing an acceptability judgment task to present these contrasts to native speakers and properly compare LLM performance with human data.

Further analyses - currently underway - are required to provide a more comprehensive understanding of the syntactic behaviors tested. These will be reported in future work.

Acknowledgments

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title T-GRA2L: Testing GRAdeness and GRAmmaticality in Linguistics – CUP I53D23003900006 - Grant Assignment Decree No. 104 adopted on the 2nd February 2022 by the Italian Ministry of Ministry of University and Research (MUR). PI: CC

References

[1] G. Cinque, Restructuring and functional heads, Cartography of Syntactic Structures (Hardcover), Oxford University Press, Cary, NC, 2006.

[2] G. Cinque, Adverbs and functional heads: A cross-linguistic perspective, Oxford University Press, 1999.

[3] R. Orlando, L. Moroni, P.-L. H. 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: Proceedings of the 10th Italian conference on computational linguistics (CLiC-it 2024), 2024, pp. 707–719.

[4] W. De Vries, M. Nissim, As good as new. how to successfully recycle english gpt-2 to make models for other languages, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP

2021, 2021, pp. 836–846. doi:10.18653/v1/2021.findings-acl.74.
[5] DBMDZ - Bavarian State Library, Bert-base italian xxl uncased, https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased, 2020. Accessed: 2025-08-01.
[6] L. De Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, M. Guerini, Geppetto carves italian into a language model, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-It 2020, Bologna, 2021.
[7] T. Linzen, E. Dupoux, Y. Goldberg, Assessing the ability of lstms to learn syntax-sensitive dependencies, Transactions of the Association for Computational Linguistics 4 (2016) 521–535.
[8] Y. Goldberg, Assessing bert's syntactic abilities, 2019. URL: https://arxiv.org/abs/1901.05287. arXiv:1901.05287.
[9] E. Wilcox, R. Levy, T. Morita, R. Futrell, What do rnn language models learn about filler-gap dependencies?, 2018. URL: https://arxiv.org/abs/1809.00042. arXiv:1809.00042.
[10] J. Hu, J. Gauthier, P. Qian, E. Wilcox, R. Levy, A systematic assessment of syntactic generalization in neural language models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1725–1744. URL: https://www.aclweb.org/anthology/2020.acl-main.158.
[11] T. Grano, Control and Restructuring, Oxford Studies in Theoretical Linguistics, Oxford University Press, London, England, 2015.
[12] S. Wurmbrand, Infinitives, Berlin: De Gruyter Mouton, 2001.
[13] S. Wurmbrand, Restructuring cross-linguistically, LingBuzz (2015). doi:lingbuzz/002514.
[14] G. Cinque, L. Rizzi, The cartography of syntactic structures, CISCL Working Papers on Language and Cognition 2 (2012) 43–59. doi:10.1093/oxfordhb/9780199544004.013.0003.
[15] G. Scontras, J. Degen, N. D. Goodman, Subjectivity predicts adjective ordering preferences, Open Mind 1 (2017) 53–66.
[16] G. Scontras, J. Degen, N. D. Goodman, On the grammatical source of adjective ordering preferences, Semantics and Pragmatics 12 (2019) 7–1.
[17] G. Scontras, Adjective ordering across languages, Annual Review of Linguistics 9 (2023) 357–376.
[18] G. Cinque, Deriving greenberg's universal 20 and its exceptions, Linguistic inquiry 36 (2005) 315–332.
[19] L. Rizzi, The fine structure of the left periphery, Elements of grammar: Handbook in generative syntax (1997) 281–337.
[20] L. Rizzi, G. Bocci, Left periphery of the clause: Primarily illustrated for italian, The Wiley Blackwell Companion to Syntax, Second Edition (2017) 1–30.
[21] R. Kayne, Some notes on comparative syntax, with special reference to english and french, The Oxford Handbook of Comparative Syntax (2012) 3–69. doi:10.1093/oxfordhb/9780195136517.013.0001.
[22] K. Abels, Towards a restrictive theory of (remnant) movement!, Linguistic variation yearbook 7 (2007) 53–120.
[23] G. Ramchand, P. Svenonius, Deriving the functional hierarchy, Language sciences 46 (2014) 152–174.
[24] G. C. Ramchand, Situations and syntactic structures: Rethinking auxiliaries and order in English, volume 77, MIT Press, 2018.
[25] T. Biberauer, Peripheral significance: a phasal perspective on the grammaticalisation of speaker perspective, Jung (2017) 93.
[26] M. Binz, E. Schulz, Turning large language models into cognitive models, arXiv preprint arXiv:2306.03917 (2023).
[27] M. Olivier, C. Sevdali, R. Folli, Clitic Climbing and Restructuring in the History of French, Glossa 8 (2023) 1–45.
[28] T. Sgrizzi, When infinitives are not under control: the growing trees hypothesis and the developmental advantage of restructuring verbs, RGG 46 (2024) 1–39.
[29] T. Sgrizzi, The Acquisition of Restructuring and Control, Master's thesis, University of Siena, Siena, Italy, 2022.
[30] M. Wilson, J. Petty, R. Frank, How abstract is linguistic generalization in large language models? experiments with argument structure, Transactions of the Association for Computational Linguistics 11 (2023) 1377–1395.
[31] A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, S. R. Bowman, Blimp: The benchmark of linguistic minimal pairs for english, Transactions of the Association for Computational Linguistics 8 (2020) 377–392.
[32] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, arXiv preprint arXiv:2206.04615 (2022).
[33] J. Hale, M. Stanojević, Do llms learn a true syntactic universal?, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 17106–17119.
[34] I. Landau, Control (Elements), LingBuzz (2024). doi:lingbuzz/008204.
[35] SapienzaNLP - Sapienza University of Rome, Minerva-7b-base-v1.0, https://huggingface.co/sapienzanlp/Minerva-7B-base-v1.0, 2024. Accessed:

2025-08-01.

[36] GroNLP - University of Groningen, gpt2-medium-italian-embeddings, https://huggingface.co/GroNLP/gpt2-medium-italian-embeddings, 2020. Accessed: 2025-08-01.

[37] GroNLP - University of Groningen, gpt2-small-italian, https://huggingface.co/GroNLP/gpt2-small-italian, 2020. Accessed: 2025-08-01.

[38] L. D. Mattei, Geppetto: Italian gpt-2 model, https://huggingface.co/LorenzoDeMattei/GePpeTto, 2021. Accessed: 2025-08-01.

[39] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, A. Zou, The language model evaluation harness, 2024. URL: https://zenodo.org/records/12608602. doi:10.5281/zenodo.12608602.

[40] C. Chesi, M. Barbini, M. L. P. Bianchessi, V. Bressan, A. Fusco, S. Neri, S. Rossi, T. Sgrizzi, From recursion to incrementality: Return to recurrent neural networks, Linguistic Vanguard (forthcoming).

[41] L. Charpentier, L. Choshen, R. Cotterell, M. O. Gul, M. Hu, J. Jumelet, T. Linzen, J. Liu, A. Mueller, C. Ross, et al., Babylm turns 3: Call for papers for the 2025 babylm workshop, arXiv preprint arXiv:2502.10645 (2025).

[42] A. Fusco, M. Barbini, M. L. P. Bianchessi, V. Bressan, S. Neri, S. Rossi, T. Sgrizzi, C. Chesi, Recurrent networks are (linguistically) better? an (ongoing) experiment on small-lm training on child-directed speech in italian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 382–389.