Large Language Models Under Evaluation: An Acceptability, Complexity And Coherence Assessment In Italian

Cristiano Chesi*
Scuola Universitaria Superiore IUSS, Pavia

Francesco Vespignani**
Università degli Studi di Padova

Roberto Zamparelli†
Università degli Studi di Trento

This paper discusses the results of various experiments assessing the morphosyntactic and semantic competence in Italian of four very large language models (vLLMs): `davinci` *(GPT-3/ChatGPT),* `davinci-002`, `davinci-003` *(both GPT-3.5 models) and* `gpt-4-1106-preview` *(GPT-4). We evaluated these models on (i) acceptability, (ii) complexity, and (iii) coherence judgments using 7-point Likert scales and on (iv) syntactic development through a forced choice task. The test sets were drawn from shared NLP tasks and standard linguistic assessments. The results suggest that, although fine-tuned transformers outperform all GPT models, GPT-4 represents a significant improvement over third-generation GPT models. According to our tests, even if GPT-4 and fine-tuned transformers cannot be considered descriptively or explanatorily adequate, they nonetheless pose a challenge to the poverty of the stimulus hypothesis. The "theory" expressed by GPT models is not linguistically intelligible in any relevant sense, and their training data is orders of magnitude larger than the primary linguistic input available to children. Nevertheless, GPT-4 captures certain generalizations, such as the constraints blocking the insertion of an overt resumptive clitic in specific gap positions, that are arguably unlearnable from just primary positive data.*

1. Introduction

Very Large Language Models (vLLMs), particularly those based on transformer architectures (Vaswani et al. 2017) such as the GPT series, appear to challenge the poverty of the stimulus hypothesis (Chomsky 1986). This conclusion is supported by various position papers arguing that vLLMs represent the current state of the art in scientific linguistic theories (Piantadosi 2024, a.o.). According to this view, generative grammars, including the current Minimalist Framework (Chomsky 1995), do not seem capable of competing with machine learning-based models when various computational perspectives are considered, ranging from machine translation to grammaticality judgments.

* NeTS Lab - P.zza Vittoria 15, Pavia, Italy. E-mail: cristiano.chesi@iusspavia.it
** Dep. of Developmental Psychology and Socialisation - Via Venezia, 8 - Padova, Italy. E-mail: francesco.vespignani@unipd.it
† Dep. of Psychology and Cognitive Sciences - Corso Bettini, 31, Rovereto, Italy. E-mail: roberto.zamparelli@unitn.it

While it is undeniable that GPT-like models exhibit impressive capabilities across a wide range of natural language processing tasks, it would nevertheless be premature to dismiss explicit linguistic theories as inadequate. These theories may still offer valuable contributions to our understanding of linguistic competence, particularly in terms of descriptive and explanatory adequacy (Chomsky 1965).

This study aims to assess the extent to which the aforementioned position is overly radical and to explore the potential role of explicit linguistic theory in the era dominated by Transformer-based models (2022-2024). To this end, we first evaluate GPT-family vLLMs across four specific, gradable linguistic dimensions: (i) acceptability, (ii) complexity, (iii) coherence, and (iv) syntactic development. We then consider the insights gained from comparing various models, both in terms of their predictions for linguistic development and the characteristics of the underlying theory, particularly its dimension and transparency.

This work is organized as follows. First, we revisit the notion of linguistic competence and the distinctions among observational, descriptive, and explanatory adequacy. We then introduce the test sets used to assess the vLLMs. Specifically, we refer to the dataset from the Accompl-IT shared task at Evalita 2020 for evaluating (i) grammaticality and (ii) complexity (Brunato et al. 2018); the PreTENS task from SemEval 2022 for the (iii) coherence dimension (Chowdhury et al. 2022); and the COnVERSA test set (Chesi et al. 2024) for assessing (iv) linguistic development, focusing on sensitivity to minimal pairs.

The GPT models evaluated are four: the pre-trained `davinci model (GPT-3)`; `davinci-002` and `davinci-003`, both trained on the same dataset with different fine-tuning approaches (representing GPT-3.5); and `gpt-4-1106-preview (GPT-4)`, the first preview of the most recent model available at the time of writing. While testing these models, we evaluate their performance not only in comparison to human benchmarks but also against the best-performing, smaller-scale, fine-tuned transformer models that won the respective shared tasks in three dimensions: (i) acceptability, (ii) complexity, and (iii) coherence.

We conclude that the performance of the three GPT-3.x models is substantially lower than that of the winning fine-tuned transformer models. However, none of these models can be regarded as plausible linguistic theories, as they lack both descriptive and explanatory adequacy. In contrast, GPT-4 demonstrates a notable improvement over its predecessors and performs comparably (though still inferior) to the state-of-the-art fine-tuned models. Importantly, GPT-4 produces morphosyntactically and semantically consistent predictions across all tasks, particularly in the linguistic development task. This lends at least partial support to the challenge posed by the poverty of the stimulus hypothesis, as articulated by Piantadosi (Piantadosi 2024).

2. Theoretical Background: a Definition of Linguistic Competence and Theory Adequacy

Evaluating linguistic competence is a complex and sensitive undertaking. One viable approach, grounded in the Chomskyan tradition, is to adopt a purely set-theoretic perspective, that is, to assess a theory based on its ability to distinguish between grammatical and ungrammatical sentences. A theory that consistently captures this distinction in accordance with native speakers' intuitions is considered observationally adequate (Chomsky 1965). Grammaticality judgments can sometimes be straightforward, as illustrated by example (1a) (grammatical) versus (1b) (ungrammatical, marked with the conventional asterisk *). In other cases, however, the judgments may be less

clear-cut, as in the contrast between (2a) and (2b), where the question mark (?) indicates a less intuitive or marginal status.

(1) a. ChatGPT works well.
b. *ChatGPT work well not.

(2) a. The professor praised the student that created the program that solved the problem.
b. ?*The program that the student that the professor praised created solved the problem.

While in (1b) both clear agreement and word-order violations are detectable, in (2b) we observe the application of a consistent syntactic strategy (namely, relative clause formation) that yields well-formed outcomes in (2a) (subject relative clause formation on direct objects), but creates difficulties in (2b) (object relative clause formation on subjects). No native speaker of English (apart from formally trained linguists) would consider (2b) to be grammatical in any intuitive sense and she will typically reject it as ill-formed.

In psycholinguistics, we often refer to *acceptability* as a nuanced counterpart to *grammaticality*. A naïve sentence judgment may be influenced by factors such as semantic plausibility or the frequency of particular words or collocations. For instance, a sentence may be considered grammatical yet still difficult to process because of non-local dependencies, as in the case of (2b), where the dependency between the head of the relative clause ("program") and its predicate ("solves") spans linearly intervening, potentially compatible noun phrases (i.e., "the student" and "the professor".

A formal competence model that can generate or recognize the structure in (3a) should, by the same principles, also generate or recognize the structure in (3b), which provides a relevant, though simplified, structural description of (2b):

(3) a. [[the student]$_i$ [that [the professor] praised $\__i$] $\__i$ created [the program]]
b. [[[the program]$_j$ that [[the student]$_i$ [that [the professor] praised $\__i$] $\__i$ created $\__j$] $\__j$ solves the problem.

To account for more nuanced predictions, it may be necessary to move beyond a binary, set-theoretic approach and consider whether the cognitive resources required to process certain structures (such as those illustrated in (3b)) exceed those available to human speakers. It is therefore essential to consider two additional dimensions of adequacy: on the one hand, the extent to which a specific structural generalization is compatible with (or imposes) additional theoretical assumptions, given that theoretical elegance increases as the number of assumptions decreases; on the other hand, the extent to which the theory makes precise predictions about the cognitive effort or resources required to retrieve or generate the relevant structure, and to determine whether the intended sentence is well-formed and interpretable.

We refer to *descriptive adequacy* in the first case (Chomsky 1965): a theory that accurately predicts phrase structure and the format of non-local dependencies is considered more economical (that is, "smaller" in terms of requiring fewer assumptions or principles) than a theory that must stipulate exceptions or invoke spurious generalizations. In the second case, we refer to *explanatory adequacy*: a theory that more effectively accounts for the complexity associated with linguistic processing is considered more adequate

and is more likely to generate accurate predictions, for example, in contexts such as language acquisition or the manifestation of language disorders.

Particularly with regard to explanatory adequacy, gradient acceptability judgments tend to offer more informative data for fine-grained structural comparisons (Lau, Clark, and Lappin 2017). However, both gradient and categorical (binary) judgments have been shown to correlate reliably when tested on a sufficiently large set of controlled items and with an adequate number of native-speaker informants (Sprouse and Almeida 2017).

2.1 Competence Vs. Performance

The formal (set-theoretic) approach led to what is commonly referred to as Chomsky's hierarchy (Chomsky 1956): a series of inclusion relations, defined in terms of generative power, among grammars classes defined in terms of different constraints (notably, regular vs. context-free grammars). This framework offers the clear advantage of explicitly addressing abstract linguistic properties—such as counting and cross-serial dependencies (Shieber 1985)—and provides a basis for demonstrating equivalences across grammatical formalisms by showing that each formalism can generate a specific set of linguistically relevant structures. This popular and fruitful method is now overshadowed by a much more widespread approach that simply considers models fit in terms of performance on a particular shared benchmark: shared tasks include not only test-set, on which the performance of each system is calculated (benchmark), but also specific training-sets that models' developers usually employ to train their systems on specific tasks. Considering, for instance, grammaticality judgments, various data-sets have been released, including classic linguistic text-books grammaticality examples, such as *CoLA*, the *Corpus of Linguistic Acceptability* (Warstadt, Singh, and Bowman 2019) and *ItaCoLA* for Italian (Trotta et al. 2021), or large data collections of experimental results, for instance including eye-tracking studies, the Multilingual Eye-movement Corpus, *MECO* (Siegelman et al. 2022) or fMRI scans (Brennan et al. 2016). It is useful to classify the linguistic data contained in these datasets along at least three dimensions: two related to the type of collected measure (binary vs. gradient and explicit vs. implicit) and one related to the degree of experimental control over the stimuli (ranging from fully controlled minimal pairs to naturalistic sentences). At one end of this multidimensional classification, we find CoLA-like datasets, which involve explicit, binary grammaticality judgments on carefully controlled items. At the opposite end lie datasets such as that of Brennan and colleagues, which rely on implicit, continuous neurophysiological measures collected during the processing of naturalistic input—for example, listening to the novel "Alice in Wonderland". For the purposes of our assessment, explicit judgments on controlled items represent a more reliable choice. Since our goal is to compare descriptive adequacy (in terms of theory size) and explanatory adequacy (via multi-level comparisons), an overt competence theory of the generative kind, such as Minimalist Grammar, aims to derive elegant generalizations from a minimal set of necessary and sufficient assumptions. As such, it should be competitive with any vLLM, which, in turn, should be able to generalize accurately across the controlled dimensions (i.e., the restricted set of tested phenomena) without being misled by irrelevant lexical variation. Ultimately, fine-grained comparisons are required to determine which theory ranks higher in terms of explanatory adequacy. Gradient judgments may be necessary to organize linguistic phenomena along a natural scale of complexity, one that is also reflected in human acceptability judgments. It is therefore important to clarify a distinction that often leads to misunderstanding (Chesi

and Moro 2015): on the one hand, we are concerned with a purely abstract description of linguistic competence—that is, the grammatical theory under evaluation; on the other hand, we must test the predictions that this theory yields in real-world tasks, using actual human judgments or other implicit measures that reflect realistic human performance. If linguistic competence is an abstract and inherently unobservable form of knowledge, then linguistic performance becomes the only measurable data source available for inferring the most elegant (descriptively adequate) and robust (explanatorily adequate) theoretical account.

Before turning to test set selection and outlining our assessment procedure, it is necessary to address a persistent source of terminological confusion. First, the term *generative*, when used in the context of artificial intelligence, typically refers to a system's capacity to generate novel content. However, within the domain of grammatical theory, *generative* denotes a formal approach oriented toward a set-theoretic goal: namely, the explicit characterization of a recursive procedure capable of generating—and recognizing—the (infinite) set of well-formed sentences constituting a particular language. A second ambiguity arises with the term *parameters*. In AI models, *parameters* are understood as learned numerical weights or dimensions. In contrast, in the context of generative grammar, *parameters* refer to abstract settings that account for systematic cross-linguistic variation. In the first case, the number of parameters serves as a convenient indicator of the size of an AI-trained model, e.g., GPT-3 uses 175 billion parameters (Brown et al. 2020). In the second case, however, the corresponding number (typically only a few tens) is not a meaningful measure of the size or complexity of a generative competence theory. Despite significant advances in the field (Baker 2001; Biberauer and Roberts 2017; Gianollo, Guardiano, and Longobardi 2008; Guardiano and Longobardi 2016; Roberts 2019), no universal consensus has emerged regarding a definitive list of parameters, nor is there a straightforward formalization capable of coherently integrating all proposed micro-, meso-, and macro-parameters into a unified theory. While it would be valuable to compare competing assumptions concerning the formulation and necessity of specific parameters, such efforts face substantial obstacles. As Fong's seminal dissertation illustrates (Fong 1991), bridging the gap between the abstract principles-and-parameters framework (Chomsky 1981) and a fully operational model capable of generating and recognizing (i.e., parsing) well-formed sentences requires the introduction of numerous additional assumptions (effectively, extra "parameters" in the AI sense). The often informal and computationally underspecified nature of certain linguistic intuitions recalls another important distinction proposed by David Marr (Marr 1982), namely, the divide between the *computational* and *algorithmic* levels of analysis. The former, more closely aligned with a competence-based perspective, is concerned with defining the general domain and the goal of the computation (e.g., distinguishing grammatical from ungrammatical sentences). The latter, by contrast, focuses on specifying the precise procedures required to achieve such outcomes, for example, deriving a specific acceptability judgment. Since the algorithmic level lends itself to quantifiable analysis in terms of computational complexity (e.g., memory requirements, i.e., *space*, and processing steps, i.e., *time*), it represents the natural level at which to develop a performance theory.

2.2 Performance Assessment Task

Overall, controlled items provide a more straightforward way to evaluate specific theoretical assumptions and to assess the fit of explicit models, particularly in contrast to naturalistic test sets, where a range of semantic and pragmatic factors, including

context and world knowledge, are difficult to isolate and quantify. For example, if we aim to determine whether number (singular vs. plural) and person (first/second vs. third) agreement violations elicit comparable effects in native speakers of Italian, we can contrast minimal pairs such as (4a) vs. (4b) and (4a) vs. (4b'). To measure the magnitude of any observed differences, native speakers can be asked to rate each sentence on a 7-point Likert scale, assigning a score close to 1 for clearly ungrammatical sentences and a score close to 7 for sentences perceived as fully grammatical. By repeating this task with a sufficient number of native speakers and multiple lexical irrelevant variations of the same paradigm, we can apply robust statistical methods (Bates et al. 2015) to evaluate the presence and magnitude of the effect, treating both participant and lexical variation as random effects.

(4) a. Gianni mangia un panino
G. eat.3P.SG a sandwich

b. *Gianni mangiano un panino (number agreement error)
G. eat.3P.PL a sandwich

b'. *Gianni mangio un panino (person agreement error)
G. eat.1P.SG a sandwich

'Gianni eats a sandwich'

We consider this the most direct approach for falsifying a linguistic theory: all else being equal, if Theory A does not predict a specific effect that is in fact empirically observed, this theory must be rejected in favor of Theory B, which does predict such an effect. It is important to note that a theory may correctly account for numerous relevant phenomena; however, scientific progress fundamentally depends on the ability to disconfirm a theory in favor of a better alternative. In this regard, controlled minimal pairs are essential, as they accelerate this process by isolating the critical contrasts under investigation, something that is often difficult to achieve using only naturalistic data, such as corpus-based evidence.

2.3 On Model Reliability

The notion of *adequacy* is sometimes at odds with the notion of *robustness*. In computational terms, a *robust* system is one that tolerates errors and can successfully complete a computation even when the input is ill-formed. This is exemplified by search engines such as Google, which often return plausible results even when a keyword is mistyped—typically accompanied by the informative suggestion, "Did you mean ...". From this perspective, both sentences (4b) and (4b') may be understandable to native speakers without requiring substantial processing effort. Nevertheless, they should still be classified as ungrammatical in a strict sense. Statistical models are typically tolerant (or robust) with respect to certain morphosyntactic errors, such as subject-verb agreement violations. These models can often identify ill-formed input by detecting unusually low-frequent sequences, which signal deviations from typical patterns in the training data. Moreover, it is inherently straightforward for statistical models to rank different types of violations based on the same statistical criteria. By contrast, an explicit formal grammar lacks such robustness: it simply rejects any input that is locally or globally ill-formed according to the rules, principles, or assumptions defined by the theory. To rank different violations within formal models, additional

assumptions are required. One option is to count the number of violated constraints, as in Optimality Theory (Kager 1999). Alternatively, one may consider the number of operations needed to derive a particular phrase structure (Jakubowicz 2011), or the length of dependencies involved, especially when these span across intervening, structurally relevant elements (Friedmann, Belletti, and Rizzi 2009; Grillo 2008; Rizzi 1990; Starke 2001). In this paper, we are not concerned with robustness in the computational sense. Rather, we focus on solid generalizations that align with the data—whether categorical or gradient—obtained from controlled experimental tasks. A key prediction to be evaluated alongside grammaticality judgments is that of *complexity*: whenever a sentence is judged grammatical or ungrammatical, its processing may still vary in terms of difficulty. For example, while determining the meaning of (4b) or (4b') may be as straightforward as recognizing their ungrammaticality, understanding sentence (3b) may require substantially more processing effort. In such cases, native speakers may find it more difficult to determine whether the sentence is well-formed or not. A theory that correctly predicts the grammatical status of (3b) and (4b/b'), as contrasted with their grammatical counterparts (3a) and (4a), but fails to rank sentences such as (3b) as more complex than (4b) in precise, numerical terms, is less explanatorily adequate than a theory that succeeds in making such distinctions.

2.4 More on the *Categorical* vs. *Gradual* Perception of (Un)Grammatical Contrasts

Our prediction is that vLLMs based on coherent word embeddings — leveraging a sufficiently powerful attention mechanism and trained on a sufficiently rich dataset — should be able, in zero-shot learning contexts, to perform reasonably well on prediction tasks involving gradient judgments (e.g., on a 1-to-7-point scale) of both local and non-local violations. As an illustrative case, consider a language like Italian, where a non-local dependency can be established between two non-adjacent words. This occurs, for example, in subject-verb agreement when the subject is modified by a prepositional phrase whose lexical material intervenes between the subject and the relevant predicate:

(5) a. The friends of the child *plays/play with Lego bricks

b. The friends of the child that plays/*play with Lego bricks *smiles/smile

In (5a), the structure of the determiner/nominal phrase [the friends [of [the child]]] does not allow the singular number feature associated with the embedded Determiner Phrase (DP) [the child] to percolate and trigger subject-verb agreement on the predicate "play", hence the ungrammaticality of "the friends of the child plays". On the other hand, in (5b), the syntactic ambiguity concerning the attachment site of the relative clause (i.e., whether "that play(s)" is associated with "the friends" or "the child") allows for both singular and plural agreement interpretations, depending on the structural analysis: either [the friends [of [the child [that *plays/*play*]]]] or [the [friends [of [the child]] that **plays/play*]]. Modeling this agreement dependency, which, in the first case, spans a linearly intervening DP as in [[the friends [of *[the child]*]] *plays/play], is possible by adopting a word-by-word prediction task, along with specific assumptions about the structure of working memory, as implemented in architectures such as Simple Recurrent Networks (SRNs) (Elman 1993) or Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber 1997). Transformer models (Vaswani et al. 2017) perform better at resolving both the constraint illustrated in (5a) and those in (5b), where correct agreement depends on the appropriate structural analysis of the full sentence, including, crucially, the final predicate, which may induce a garden-path effect if the structure in

(5a) is incorrectly assumed. These models succeed on such configurations by relying on attention mechanisms trained through token-by-token prediction. The core task across all approaches used to train LLMs has remained essentially unchanged since the introduction of SRNs: predict the next (or masked) word. Likewise, the evaluation procedure has remained largely consistent. A trained vLLM is considered *observationally adequate* if, in a case such as (5a), it assigns a higher probability to "play" than to "plays", that is, it makes the correct prediction in line with the grammatical structure. The opposite pattern should be observed in (5b), where the model should assign a lower probability to "smiles" and a higher probability to "smile" at the end of the sentence, consistently with the intended agreement structure. More generally, we can compute the overall sentence probability — or perplexity, which can be calculated in various, but equivalent, ways (Futrell et al. 2019; Levy 2008) — for both grammatical and ungrammatical configurations. An adequate model should assign significantly lower probabilities to ungrammatical sentences compared to their grammatical counterparts. Because these predictions are numerical, the models produce scalar rather than categorical outputs, allowing them to rank structures along a continuum of well-formedness rather than merely classifying them as grammatical or ungrammatical. This scalar nature is particularly useful for examining how model predictions vary in response to irrelevant lexical variations, and for evaluating how closely they correlate with human gradient judgments collected via 7-point Likert scales, both in terms of *perceived acceptability* and *processing difficulty* (i.e., *complexity*).

3. Materials and Methods

3.1 Dataset

The dataset used to assess Acceptability and Complexity is the Italian test-set developed for the shared task Accomplit@Evalita 2020 (Brunato et al. 2020). The Acceptability section consists of 1.683 full sentences annotated with grammaticality judgments collected via crowdsourcing. Participants were asked by Brunato and colleagues to rate each sentence on a 7-point Likert scale in response to the following prompt: “How acceptable is this sentence from 1 (completely ungrammatical) to 7 (perfectly grammatical)?” (in Italian: “Quanto è accettabile questa frase da 1 (completamente agrammaticale) a 7 (perfettamente grammaticale)?”).

The dataset incorporates results from several prior studies. A total of 128 items are drawn from an acceptability study on the role of person features in cleft sentences (Chesi and Canal 2019), in which noun phrases are introduced either by definite determiners (e.g., "i linguisti", "the.PL linguists") or by second-person pronouns (e.g., "voi linguisti", "you.PL linguists"), as in (6a). An additional 515 sentences come from a study on copular constructions (Greco et al. 2020), as in (6b), which includes various prepositional phrases extractions (e.g., "di quale rivolta le foto del muro sono la causa?", "of which riot the pictures of the wall are the cause" vs. "di quale muro le foto sono la causa della rivolta?", "of which wall the pictures are the cause of the riot"). This subset also includes various control conditions assessing baseline acceptability of subjects in preverbal (6c) and postverbal (6c’) positions with unergative predicates (e.g., "abbaiare", "to bark"), unaccusative predicates (e.g., "cadere", "to fall"), and transitive predicates with an overt object (e.g., "mordere l’osso", "to bite the bone"). A further 300 sentences of varying lengths target subject–verb number and person agreement across different syntactic positions, as in (6d) (Mancini, Canal, and Chesi 2018). Finally, 48 sentences involving

extraction from wh-islands are included (Villata et al. 2015), based on the distinction between bare and discourse-linked (D-linked) wh-phrases, as exemplified in (6e).

(6) a. {Sono | Siete} {i | voi} linguisti che {gli | voi} psicologi
{are.3PL | are.2PL} {the | you} linguists that {the | you} psychologists
{criticano | criticate} sempre
{criticize.3PL | criticize.2PL} always

'it's {the | you} linguists that {the | you} psychologists always criticize

b. Le foto del muro sono la causa della rivolta
the pictures of.the wall are the cause of.the riot

b′. La causa della rivolta sono le foto del muro
the cause of.the riot are the pictures of.the wall

'the cause of the riot is the pictures of the wall'

c. I cani {hanno abbaiato | sono caduti | hanno morso l' osso}
the dogs {have barked | have fallen | have bitten the bone}

c′. {Hanno abbaiato | Sono caduti | Hanno morso} i cani {l' osso}
{have barked | have fallen | have bitten} the dogs {the bone}

d. Qualcuno ha detto che io {scrivo | scriviamo} una lettera
Someone has said that I {write.1SG | write.1PL} a letter

e. {Cosa | Quale edificio}$_i$ ti chiedi {chi | quale ingegnere}
{what | which building}$_i$ cl.2SG.DAT ask {who | which engineer}
abbia costruito $\__i$?
have built

'{what/which building}$_i$ did you ask yourself {who/which engineer} have built $\__i$?'

Further 672 full sentences in the test set were generated using specific syntactic patterns, including: (i) wh-extraction, as in (7a); (ii) topicalization with or without a gap, as in (7b); (iii) interrogative wh- or relative clauses involving across-the-board extraction—occurring in one, none, or both conjuncts, as in (7c); (iv) extraction from wh-islands with various intervening elements, as in (7d); (v) extraction from subject or object positions, as in (7e); and (vi) negative polarity items (NPIs) with or without licensing negation, as in (7f).

(7) a. {Che cosa | quale problema}$_i$ lo studente dovrebbe risolvere {$\__i$ | -lo$_i$}?
{that what | which problem}$_i$ the student must solve {$\__i$ | it$_i$}

'{what | which problem}$_i$ must the student solve {$\__i$ | it$_i$}'

b. Questo problema$_i$, lo studente dovrebbe risolver(e) {$\__i$ | -lo$_i$}.
this problem$_i$, the student must solve {$\__i$ | it$_i$}

c. Chi$_i$ ... Maria vuole chiamar(e) {$\__i$ | -lo$_i$} e Mario medicar(e) {$\__i$ | -lo$_i$}.
Who$_i$ ... M. want to.call {$\__i$ | it$_i$} and M. to.medicate {$\__i$ | it$_i$}

d. Quale provvedimento Maria ha saputo {che | dove | perché | quando} il
Which measure M. has knew {that | where | why | when} the
ministro prenderà?
minister will.take

'Which measure has M. knew {that | where | why | when} the minister will take?'

e. Carlo conosceva bene il compagno$_i$ di classe che {incontrare $\__i$
C. knows well the friend of class that {to.meet $\__i$
divertiva sempre Anna | Anna voleva sempre incontrare $\__i$}.
amused always A. | A. wants always to.meet $\__i$}

'C. knows well the classmate that {to meet $\__i$ always amused A. | A. wants always to meet $\__i$}'

f. {Maria | Nessuno} si aspetta che qualcuno possa avere {già
{M. | nobody} himself expects that someone could have {already
| mai} finito questo esercizio.
| never} completed this exercise

'{Maria | no one} expects anyone to have {ever | never} finished this exercise'

The selection of these phenomena is guided primarily by theoretical considerations: standard datasets for evaluating linguistic competence in large language models rarely include the minimal structural contrasts needed to systematically probe diverse functional domains. For reasons of space, we refer interested readers to the original papers for a comprehensive discussion of the linguistic motivation for each contrast.

The Complexity test set included the 672 sentences constructed from the (7) paradigms, along with 1,858 "ecological sentences." Of these, 1,128 were extracted from the Italian Stanford Dependency Treebank (Bosco, Dell'Orletta, and Montemagni 2014; Brunato et al. 2018), while the remaining sentences were drawn from the PoSTWITA and TWITTIRÒ treebanks (Cignarella et al. 2019; Sanguinetti et al. 2018).

For the Coherence task, the reference test set was the Italian section released for the Pretens@SemEval 2023 shared task (Zamparelli et al. 2022). Although the syntactic structure of these sentences was well-formed, a full acceptability judgment required the reader to rely on world knowledge to evaluate presuppositions. These judgments depended on taxonomic inclusion relations between the DPs involved and the specific connective used.

(8) a. Mi piacciono gli alberi più {#degli abeti | dei palazzi}
I love the trees more.than {#the fir.trees | the buildings}

b. Mi piacciono gli alberi, ad eccezione {degli abeti | #dei
I love the trees, with.the exception {of.the fir.trees | #the
palazzi}
buildings}

Eleven connective types are assessed ("ad eccezione di", "with the exception of"; "ed in particolare", "and in particular"; "in generale", "in general"; "generalmente", "generally"; "un tipo di", "a kind of"; "e anche", "and also"; "più di", "more than"; "piuttosto che", "instead of"; "ma preferisco", "but I prefer"; "invece di", "instead of"; "ma non", "but not").

In the end, for the development task, we considered the contrasts included in the COnVERSA tests (Chesi et al. 2024). Unlike the other test sets, this is a forced-choice task based on minimal morphosyntactic pairs. Originally developed for assessing morphosyntactic competence in deaf children, the test employs sentence pairs that differ by only a single morphosyntactic feature. Four types of dependencies licensing this feature are considered: agreement (9A), thematic structure (9B), pronominal licensing (9C), and question formation/answer appropriateness (9D). These are further divided into 18 distinct groups of phenomena, ranging from local determiner–noun agreement (9A1) to the appropriateness of answers to *wh*-questions (9D5).

(9) A1. Il treno. Vs. *I treno.
the.SG train Vs. *the.PL train

A2. Il piatto è pieno. Vs. *Il piatto è piena.
the dish is full.SG.M Vs. *the dish is full.SG.F

A3. Il maestro corregge i compiti. Vs. *Il maestro correggono i compiti.
the teacher corrects the homeworks Vs. *the teacher correct the homeworks

A4. Il nonno dei bambini cammina. Vs. *Il nonno dei bambini camminano.
the granpa of_the children walks Vs. *the granpa of_the children walk

A5. La mamma è andata in ufficio. Vs. *La mamma è andato in ufficio.
the mom has left.SG.F in office Vs. *the mom has left.SG.F in office

'mom has left for the office'

A6. Ai bambini piace la palla. Vs. *Ai bambini piacciono la palla.
To_the children likes the ball Vs. *To_the children like the ball

'The children like the ball'

A7. La sorella e il fratello salutano sempre. Vs. *La sorella e il fratello saluta sempre.
The sister and the brother say_hi always. Vs. *The sister and the brother says_hi always.

'Sister and brother always say hi'

B1. Il libro cade dal tavolo. Vs. *Il libro cade il tavolo.
The book falls from_the table. Vs. *The book falls the table.

B2. Il gatto ha giocato. Vs. *Il gatto è giocato.
The cat has played. Vs. *The cat is played.

B3. Il cuoco è stato riconosciuto dal ragazzo. Vs. *Il cuoco ha riconosciuto dal ragazzo.
The chief has been recognized by_the boy. Vs. *The chief has recognized by_the boy.

C1. Cosa fai? Mangio. Vs. * Mangi.
What (do you) do? (I) eat.1SG. Vs. *(You) eat.2SG.

C2. Il ragazzo scivola. Vs. *Il ragazzo si scivola.
The boy slips Vs. *The boy himself slips.

C3. La nonna disegna un albero e lo colora. Vs. *La nonna
The grandma draws a tree and it.CL.ACC color Vs. The grandma
disegna un albero e gli colora.
draws a tree and to_him.CL.DAT color

'The grandma draws a tree and colors it'

D1. Dove dorme il ragazzo? In camera. Vs. *Di notte
Where sleeps the boy? In (the) bedroom Vs. *At night

D2. Chi mangia? La mamma. Vs. *La pasta.
Who eats? (The) mom. Vs. The pasta.

D3. La bambina mangia? Sì. Vs. *Una torta.
The child eats? Yes. Vs. *A cake.

'Does the child eat?'

D4. Perché il bambino dorme? Perché è tardi. Vs. *Perché no.
Why the child sleeps? Because (it) is late. Vs. *Because no.

'Why does the child sleep?'

D5. Ci sono due bambine. Una corre, l' altra salta e chiama i
There are two children. One runs, the other jumps and calls the
cugini. Quale bambina salta? Quella che chiama i cugini. Vs.
cousins. Which child jumps? The_one that calls the cousins. Vs.
*Quella che i cugini chiamano.
*The_one that the cousins call.

A total of 344 contrasts were obtained, each including at least eight non-significant lexical variations of the same phenomenon. Based on participants' test performance, we inferred an indirect evaluation of three control measures: *Working memory* or attention level (*WM*) assessed by considering performance on items involving at least two arguments, regardless of sentence structure. *Lexical richness* (*LEX*) evaluated by focusing on items containing non-fundamental vocabulary (De Mauro 2016). Sensitivity to fine structural features (or *featural Relativized Minimality*, *fRM*) measured by analyzing performance on items involving non-local dependencies—either linear or hierarchical—as in certain attraction conditions (e.g., 9A4) (Franck et al. 2006)

3.2 Method

We evaluated the performance of GPT-3 (via ChatGPT, March 2023 version) and GPT-3.5 models using the OpenAI API under the completion pipeline. Two models based on GPT-3.5—as identified in the OpenAI documentation and updated in May 2023—were assessed: `davinci-002`, which underwent supervised fine-tuning, and `davinci-003`, which incorporates reinforcement learning. No detailed documentation is available regarding the training dataset or the specific network architecture;

only an approximate parameter count is provided for GPT-3.x models (175 billion parameters) (Brown et al. 2020). We also included the most recent GPT-4 model (`GPT-4-1106-PREVIEW`) in this study. However, technical documentation for this model remains sparse (OpenAI 2023), and the only supported access method via the API is the chat pipeline.

The implementation of the experiments was straightforward. We used the OpenAI API and set the temperature parameter to 0 (`temperature = 0`), limited the number of generated tokens to one (`max_tokens = 1`), and disabled normalization effects by setting both `frequency_penalty` and `presence_penalty` to `0`. We further configured the model to return only the most likely token and its associated probability (`top_p = 1`, `logprobs = 1`).

For Grammaticality, Acceptability, and Coherence tasks, the prompts used for the models mirrored those administered to human participants:

Grammaticality test: IT: "Dai un giudizio numerico di accettabilità alla seguente frase utilizzando una scala da 1 (completamente agrammaticale) a 7 (perfettamente grammaticale)"; EN: "Give a numerical acceptability judgment to the following sentence using a scale from 1 (completely ungrammatical) to 7 (perfectly grammatical)";

Complexity test: IT: "Dai un giudizio numerico sulla complessità della seguente frase utilizzando una scala da 1 (semplicissima) a 7 (molto difficile)"; EN: "Give a numerical complexity judgment to the following sentence using a scale from 1 (extremly simple) to 7 (very complex)";

Coherence test: IT: "Dai un giudizio numerico di accettabilità alla seguente frase utilizzando una scala da 1 (completamente inaccettabile) a 7 (completamente accettabile)"; EN: "Give a numerical acceptability judgment to the following sentence using a scale from 1 (completely unacceptable) to 7 (completely acceptable)";

Linguistic development test: IT: "Quale delle due frasi seguenti (1. oppure 2.) è grammaticalmente preferibile? 1. … 2. …"; EN: "Which of the two following sentence (1. or 2.) are grammatically preferable? 1. … 2. …").

To obtain only numerical predictions for each prompt, in addition to the parameter `max_tokens = 1`, we can add to the prompt "(without comments)" or "(only answer with a numerical judgment / with the number of the selected sentence)" to the end of the request. All these methods produce equivalent behaviors. We repeated the evaluation task individually for each item to avoid interference among sentences (as in the case of "judge the following sentence: a, b, c … n").

3.3 Results

In Table 1, we report partial results for the GPT-3 model (tested exclusively via the ChatGPT interface in March 2023, which is assumed to correspond to the `davinci` model). Notably, this model did not consistently complete the task when accessed through the API using the completion pipeline, as described in §3.2. We also include results for the `davinci-002` and `davinci-003` models, which were tested via the API following the same procedure. As a reminder, the Acceptability, Complexity, and Coherence tasks all require judgments on a 7-point Likert scale. Model performance

was evaluated by computing the Pearson correlation between predicted scores and the average human judgments for each sentence.

For each task, we provide a baseline as well as the performance of the best fine-tuned model that won the respective shared task competition: UmBERTO-MTSA for the Acceptability and Complexity tasks (Sarti 2020), and DeBERTa-v3 for the Coherence task (Xia et al. 2022). Baselines were established using linear regression models adapted from Support Vector Machines (SVMs) trained on n-gram features—unigrams and bigrams for the Acceptability and Complexity tasks, and trigrams for the Coherence task. We employed a cross-fold validation procedure on the training set, using an 80/20 split (80% training, 20% testing) randomly sampled multiple times. The final baseline reflects the average correlation across these folds.

Table 1
Pearson correlation between (zero-shot) predicted scores and average human judgments for each item successfully evaluated both by humans and models in the Acceptability, Complexity, and Coherence tasks. The number in parentheses after r indicates the degrees of freedom, i.e., the number of successfully evaluated items (*** = p<0.001).

Model	*Acceptability*	*Complexity*	*Coherence*
baseline	r(342) = .30***	r(514) = .50***	r(1010) = .34***
Winning fine-tuned model	r(342) = .88***	r(514) = .83***	r(1010) = .81***
GPT-3 (v. 03/2023)	r(342) = .37***	r(129) = .25 $^{p=0.003}$	r(198) = .11 $^{p=0.124}$
GPT-3.5 (davinci 002)	r(342) = .44***	r(509) = .25***	r(1007) = .23 $^{p=0.001}$
GPT-3.5 (davinci 003)	r(342) = .52***	r(509) = .28***	r(1007) = .37***
GPT-4 (gpt-4-1106-preview)	r(342) = .80***	r(509) = .61***	r(1007) = .60***

These results highlight four main findings:

1. with the exception of GPT-4, all other GPT models perform inconsistently across tasks; sometimes below the baseline (e.g., in the Complexity task), sometimes above it (e.g., in the Acceptability task), and sometimes roughly in line with it (e.g., in the Coherence task).
2. in a zero-shot learning context, all GPT models—including GPT-4—are outperformed by significantly smaller, fine-tuned transformer models that set the gold standard for each task.
3. although the `davinci-003` model shows a numerical improvement over `davinci-002` (as indicated by a higher Pearson correlation coefficient), the strong similarity of responses across the two models for the vast majority of items suggests that this difference is not substantively meaningful.
4. GPT-4 substantially outperforms all earlier models, achieving a correlation of up to 80% with human judgments on the Acceptability task, marking a notable advancement in performance.

In more detail, the state-of-the-art fine-tuned transformers — UmBERTO-MTSA for Acceptability and Complexity (Sarti 2020), and DeBERTa-v3 for Coherence (Xia et al. 2022) — are all based on *bi-directional Transformer architectures* (*BERT*). These models

operate at parameter scales that are orders of magnitude smaller than those of GPT-based transformers, using millions of parameters rather than the hundreds of billions used in GPT-X models. Our results suggest that task-oriented fine-tuning is the critical factor for enhancing transformer performance. It leads to fewer and more consistent errors, with mistakes typically concentrated within a smaller set of morphosyntactic categories.

From a qualitative analysis, while GPT-04 consistently outperforms all GPT-3 model, no significant difference (aside from a numerical advantage for `davinci-003`) is revealed between `davinci-003` and `davinci-002`. Both GPT-3.5 models frequently produce judgments that significantly diverge from human ratings, either positively or negatively.

Among the sentences rated substantially more negatively by the models than by humans, we observe several notable patterns. For instance, inverse copular constructions, such as "La causa della rivolta sono le foto del muro" (literally: "The cause of the riot are the pictures of the wall"; in standard English: "The cause of the riot is the pictures of the wall"), are assigned a mid-range score of 4 by GPT-3 models, whereas human judgments cluster around 6. Similarly, non-canonical word orders, such as those involving post-verbal subjects (e.g., "Hanno svaligiato i ladri la villa", literally: "have robbed the thieves the villa"; in standard English: "The thieves have robbed the villa"), receive an average score of 2 from the models, compared to an average human score of 4.

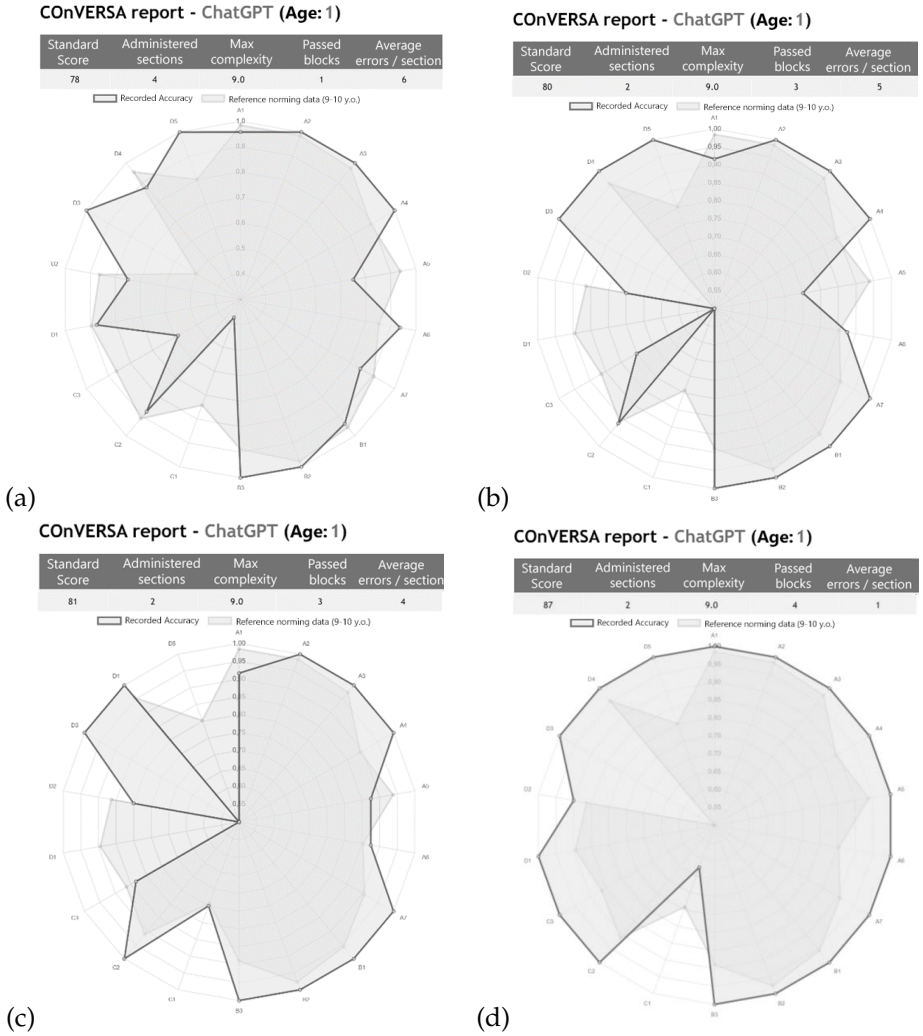
Significant negative discrepancies also appear with less frequent lexical items. For example, "Gli scalatori sono precipitati." ("The climbers fell") is rated approximately

Table 2
Models' CONVERSA Standard Scores and phenomenon-specific accuracy measures (e.g., Agreement, Thematic Structure), with percentile rankings computed relative to 9-year-old children (norming data: Chesi et al., 2024)

	GPT-3 (v.03/2023)	GPT-3.5 (davinci 002)	GPT-3.5 (davinci 003)	GPT-4 (1106-pr.)
Standard Score	78 (50° perc.)	80 (60° perc.)	81 (60° perc.)	87 (90° perc.)
Agree	93.13 (70° perc.)	95 (80° perc.)	96.25 (80° perc.)	100 (90° perc.)
Thematic structure	97.5 (80° perc.)	100 (90° perc.)	100 (90° perc.)	100 (90° perc.)
Pronouns	64.06 (10° perc.)	75 (30° perc.)	87.5 (70° perc.)	90.63 (80° perc.)
Questions	85 (70° perc.)	82.5 (70° perc.)	75 (50° perc.)	95 (90° perc.)
Working memory	85.86 (50° perc.)	90 (70° perc.)	88 (60° perc.)	99 (90° perc.)
Features sensitivity	77.72 (40° perc.)	82.61 (60° perc.)	83.7 (60° perc.)	94.57 (90° perc.)
Lexical richness	84.96 (50° perc.)	91.67 (80° perc.)	95 (80° perc.)	98.33 (90° perc.)

three points lower by GPT models compared to human judgments. Another consistent deviation is observed in sentences involving filled-gap dependencies: GPT-3.x models tend to overaccept these structures. Sentences like "Chi è che lo studente dovrebbe considerarlo?" ("Who is it that the student should have considered him?") receive a score of 6 from the models, in contrast to a human average of 2. A similar leniency is found with wh-island violations. For example, "Cosa ti chiedi chi ha costruito negli anni Sessanta?" ("What do you wonder who built in the 1960s?") is judged with a +4 difference by GPT models compared to human ratings, despite being widely considered ungrammatical in the literature due to a *superiority violation* (Chomsky, 1973). Although these issues are less frequent in GPT-4, the model still fails to align with human judgments in some

Table 3
Experiments results using the COnVERSA test with GPT models (the light-gray area in the background indicates the average performance obtained with 9 y.o. children)



cases, particularly with unlicensed negative polarity items (NPIs), such as "alcunché" ("anything"). Additionally, davinci models occasionally produce gross incorrect judgments in longer sentences. For example, "Per concimare il terreno noi utilizzavo un prodotto biologico." ("To fertilize the soil we used.1SG an organic product"), which contains a subject–verb agreement error, is incorrectly rated as fully grammatical (+5 compared to the human average of 2).

No clear error trends were observed in the Complexity and Coherence tasks for GPT-3 models, as no consistent categorical generalizations emerged, unlike in human responses.

As for the forced-choice COnVERSA test, the performance of GPT-3.X models is globally comparable with the one of a child around 9 years old as illustrated in the summary results reported in Table 2. Again, the GPT performance variance within relevant morphosyntactic items groups is high, that is, no simple generalizations can be made on the abstraction capabilities over natural morphosyntactic classes that are evident in human/children performance.

It is useful to examine performance trends across models—particularly the progression from `davinci-002` to `davinci-003`, and from the `davinci-00X` series to GPT-4 (see Table 3). Since the evaluation dimensions (e.g., A1–A7, D1–D5) are ordered according to increasing morphosyntactic complexity, a clear contrast emerges in model behavior.

The `davinci-002` model, for example, performs near ceiling on object relative clause comprehension (D5) but only at chance level on wh-adjuncts (D1). In the latter case, the model often prefers a plausible sentence continuation (e.g., "Where did the child sleep? *Last night*.) rather than a plausible answer (e.g., "Where did the child sleep? *In the bed*."). A similar pattern is observed in the agreement domain: the model shows higher error rates with indefinite determiner–noun agreement (e.g., "dei libri", "some.PL books" vs. *"del libri", "*some.SG books") than with subject–verb agreement involving psychological predicates (e.g., "Al professore servono i gessi.", "To the professor are needed the chalks" vs. *"Al professore serve i gessi.").

Interestingly, this trend is the inverse of what is observed in children. No child participant, whether hearing, deaf, or L2, exhibited this pattern of performance. In contrast, GPT-4 performs in a generally adult-like manner, except in category C1 (personal pronoun shift), which continues to present difficulties. Curiously, `davinci-003` outperformed GPT-4 on this specific category. This suggests that improvements across morphosyntactic domains are neither homogeneous nor systematic, and that performance gains do not follow a trivially predictable developmental trajectory as these models increase in scale.

3.4 Discussion

The results demonstrate a significant improvement in performance, measured as Pearson's correlation with average human judgments, across successive versions of the GPT models. GPT-3.5 models outperform GPT-3 on all tasks, and GPT-4, in turn, outperforms the GPT-3.5 models. The most substantial improvement is observed in the transition from GPT-3.5 to GPT-4: in all tasks, GPT-4 surpasses the baseline established by an SVM model trained on n-gram features. Nevertheless, GPT-4's performance remains consistently below that of the fine-tuned transformer-based models that won the shared tasks in Acceptability, Complexity, and Coherence.

The necessity of appropriate fine-tuning highlights an important insight: these types of judgments are not natural for vLLMs. In fact, LLMs benefit significantly from

explicit instruction or explanation of what is meant by complexity, acceptability, and coherence.This, in turn, requires a substantial amount of training data and examples, and may help the model exploit extra-morphosyntactic or semantic cues arising from linear patterns used to generate linguistic variations of certain items (Kodner, Payne, and Heinz 2023). All the top-performing systems in the shared tasks implemented various forms of data augmentation to expand the training sets, despite the fact that the original Complexity task dataset already included 2,530 examples.

In contrast, human participants demonstrate consistent and coherent performance across all tasks, characterized by low intra- and inter-subject variance, without the need for explicit instruction or task-specific training.

The best performances are observed in the Acceptability and Linguistic Development tasks. Apart from GPT-4, all other models display inconsistent performance with respect to the categorical generalizations underlying each task's design. For example, in the COnVERSA test, the evaluation dimensions (A1–D5) are ordered along a gradient of increasing morphosyntactic complexity. Despite known pragmatic difficulties in children related to the written modality, such as 1st–2nd person shifts in questions and yes–no question answering, it is entirely unexpected that local agreement phenomena (A1–A3) will yield lower performance than less local structures such as A5, which involves agreement with a "quirky" subject (i.e., an experiencer introduced by a preposition).

Except for GPT-4, GPT models exhibit this implausible performance trends, suggesting that they fail to abstract over morphosyntactic structure and are overly influenced by surface lexical variation. Specifically, the models often perform better on items that pose major difficulties for human participants (e.g., relative clauses), while performing worse and less consistently on simpler items involving infrequent lexical items or unusual phrasing (e.g., inverse copular constructions).

Furthermore, the improvement from `davinci-002` to `davinci-003` is not systematic across evaluation dimensions. For example, `davinci-003` performs better on category C3, while `davinci-002` outperforms it on D5. Also noteworthy is the models' insensitivity to well-defined syntactic constraints, such as island violations. In particular, in cases of subextraction from complex noun phrases, all GPT models tend to overaccept sentences when the gap is filled (either with a pronoun or a full DP) assigning them significantly higher acceptability scores than human participants.

4. Conclusion

In the experiments reported in this paper, we tested vLLMs from the GPT family using a small set of shared datasets within a zero-shot learning framework. Our goal was to evaluate the models' capacity to generalize to novel inputs and tasks that, based on human performance, are known to engage implicit morphosyntactic and semantic knowledge (including inferences based on taxonomic implicatures), largely independent of word frequency and lexical variation.

With the exception of GPT-4 on the COnVERSA test, all models exhibited performance patterns that diverge markedly from human judgments, often in implausible ways. We conclude that the underlying Transformer architecture adopted by the GPT tested models and the current training paradigms are insufficient to elicit the type of categorical generalizations that native speakers reliably and implicitly possess. The Poverty of Stimulus challenge (namely, whether the tested Transformer architectures trained with autoregressive or masked objectives on massive corpora can ultimately generalize across morphological, syntactic, and semantic dimensions in a human-like manner) appears unsupported, at least in the context of the tasks analyzed.

While this finding does not rule out the possibility that architectural modifications might improve model performance in this respect, this result is particularly noteworthy given the scale of these models. GPT-3, for instance, comprises approximately 175 billion parameters, three orders of magnitude fewer than the estimated number of synaptic connections in the adult human brain (approximately 164 trillion (Tang et al. 2001)). This gap has likely narrowed further in more recent models such as GPT-4. Nevertheless, our results suggest that this approach does not yield a model that is closer to a descriptively adequate theory of language competence (cf. Piantadosi, 2023). The performance patterns exhibited by the tested models are not predictive of human linguistic behavior, whether in hearing adult or child native speakers, deaf individuals, or second-language (L2) learners.

On the other hand, GPT-4's strong performance, particularly in the COnVERSA test, does raise intriguing questions for the poverty of the stimulus hypothesis. Despite the unrealistic scale of its training data, GPT-4 appears capable of capturing more subtle grammatical generalizations, including constraints on question formation — a classic argument to support the Poverty of Stimulus hypothesis (Crain and Nakayama 1987) — and the avoidance of resumptive pronouns in gap positions (Wilcox, Futrell, and Levy 2024; Lan, Chemla, and Katzir 2025). Nonetheless, the results contribute meaningfully to ongoing debates about the nature of language acquisition, representation, and the role of inductive biases in model design.

5. Limitations

This work originated in the context of the growing interest in GPT models in early 2023. The first tests included only the ChatGPT model available at that time. Before and after the LVI International SLI (Società di Linguistica Italiana) Congress Workshop (Linguistica Teorica e Trattamento Automatico delle Lingue: verso nuove sinergie), where we presented a preliminary version of this study, we added two additional versions of GPT-3 and, later the GPT-4 model, all accessible through API calls. Since our initial experiments, several other commercial models have been released. However, because the architecture and training data of most of these models remain undisclosed, we have chosen not to extend our evaluation to them. For this reason, the relevance of the findings discussed in the present paper should be understood within their historical context, tracing the developmental trajectory from early GPT-3 models to GPT-4 rather than offering a comprehensive assessment of current state-of-the-art models.

Acknowledgments

We are grateful to the audience of the SLI 2023 Workshop "Linguistica Teorica e Trattamento Automatico delle Lingue: verso nuove sinergie" held in Turin, 14-16 September 2023.
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 (202223PL4N) – 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. This work contains simulations carried out on the High Performance Computing DataCenter at IUSS, co-funded by Regione Lombardia through the funding programme established by Regional Decree No. 3776 of November 3, 2020.

References

Baker, Mark C. 2001. *The atoms of language*. Basic Books, New York, New York, first edition.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Biberauer, Theresa and Ian Roberts. 2017. Parameter Setting. In Adam Ledgeway and Ian Roberts, editors, *The Cambridge Handbook of Historical Syntax*. Cambridge University Press, 1 edition, pages 134–162. DOI: 10.1017/9781107279070.008.

Bosco, Cristina, Felice Dell'Orletta, and Simonetta Montemagni. 2014. The Evalita 2014 Dependency Parsing Task. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014, 9-11 December 2014, Pisa*. Pisa university press. DOI: 10.12871/clicit201421.

Brennan, Jonathan R., Edward P. Stabler, Sarah E. Van Wagenen, Wen-Ming Luh, and John T. Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, 157-158:81–94, 6.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, 7. arXiv: 2005.14165.

Brunato, Dominique, Cristiano Chesi, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi, and Roberto Zamparelli. 2020. Accompl-it@ EVALITA2020: Overview of the acceptability & complexity evaluation task for italian. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020). CEUR. org*, Online, December.

Brunato, Dominique, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this Sentence Difficult? Do you Agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium, October. Association for Computational Linguistics.

Chesi, Cristiano and Paolo Canal. 2019. Person Features and Lexical Restrictions in Italian Clefts. *Frontiers in Psychology*, 10(2019).

Chesi, Cristiano, Giorgia Ghersi, Valentina Musella, and Debora Musola. 2024. *COnVERSA: Test di Comprensione delle Opposizioni morfo-sintattiche VERbali attraverso la ScritturA*. Hogrefe, Firenze.

Chesi, Cristiano and Andrea Moro. 2015. The subtle dependency between Competence and Performance. *MIT Working Papers in Linguistics*, 77(2015):33–46.

Chomsky, Noam. 1956. Three models for the description of language. *IEEE Transactions on Information Theory*, 2(3):113–124, 9.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*, volume 11. MIT press, Cambridge, Massachusetts.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Number 9 in Studies in Generative Grammar. Foris, Dordrecht, Netherlands.

Chomsky, Noam. 1986. *Knowledge of language: its nature, origin, and use*. Convergence. Praeger, New York.

Chomsky, Noam. 1995. *The minimalist program*. MIT press, Cambridge, MA.

Chowdhury, Shammur Absar, Cristiano Chesi, Giulia Venturi, Dominique Brunato, Felice Dell'Orletta, Simonetta Montemagni, and Roberto Zamparelli. 2022. Evaluating presuppositional knowledge in language models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SEMEVAL 2022)*, Seattle, Washington, USA, July.

Cignarella, Alessandra Teresa, Cristina Bosco, Rosso Paolo, and others. 2019. Presenting TWITTIRO-UD: An italian twitter treebank in universal dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France, August. ACL.

Crain, Stephen and Mineharu Nakayama. 1987. Structure dependence in grammar formation. *Language*, 63(3):522–543.

De Mauro, Tullio. 2016. Il Nuovo vocabolario di base della lingua italiana. *Internazionale.[28/11/2020]. https://www. internazionale. it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana*, 11.

Elman, Jeffrey L. 1993. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99, 7.

Fong, Sandiway. 1991. *Computational properties of principle-based grammatical theories*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts.
Franck, Julie, Glenda Lassi, Ulrich H. Frauenfelder, and Luigi Rizzi. 2006. Agreement and movement: A syntactic analysis of attraction. *Cognition*, 101(1):173–216. Elsevier.
Friedmann, Naama, Adriana Belletti, and Luigi Rizzi. 2009. Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua*, 119(1):67–88. Elsevier.
Futrell, Richard, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.
Gianollo, Chiara, Cristina Guardiano, and Giuseppe Longobardi. 2008. Three fundamental issues in parametric linguistics. In Theresa Biberauer, editor, *Linguistik Aktuell/Linguistics Today*, volume 132. John Benjamins Publishing Company, Amsterdam, Netherlands, pages 109–142. DOI: 10.1075/la.132.05gia.
Greco, Matteo, Paolo Lorusso, Cristiano Chesi, and Andrea Moro. 2020. Asymmetries in nominal copular sentences: Psycholinguistic evidence in favor of the raising analysis. *Lingua*, 245:102926, 10.
Grillo, Nino. 2008. *Generalized minimality: Syntactic underspecification in Broca's aphasia*. LOT, Utrecht, Netherlands.
Guardiano, Cristina and Giuseppe Longobardi. 2016. Parameter Theory and Parametric Comparison. In Ian Roberts, editor, *The Oxford Handbook of Universal Grammar*. Oxford University Press, pages 376–398. DOI: 10.1093/oxfordhb/9780199573776.013.16.
Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780. MIT Press.
Jakubowicz, Celia. 2011. Measuring derivational complexity: New evidence from typically developing and SLI learners of L1 French. *Lingua*, 121(3):339–351, 2.
Kager, René. 1999. *Optimality theory*. Cambridge University Press, Cambridge, United Kingdom. OCLC: 56218324.
Kodner, Jordan, Sarah Payne, and Jeffrey Heinz. 2023. Why linguistics will thrive in the 21st century: A reply to piantadosi (2023). arXiv preprint arXiv:2308.03228.
Lan, Nur, Emmanuel Chemla, and Roni Katzir. 2025. Large language models and the argument from the poverty of the stimulus. Linguistic Inquiry. In press HAL Id: hal-05404737.
Lau, Jey Han, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241, 7.
Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177. publisher: Elsevier.
Mancini, Simona, Paolo Canal, and Cristiano Chesi. 2018. The acceptability of person and number agreement/disagreement in Italian: an experimental study. *Lingbuzz: 005514*.
Marr, David. 1982. *Vision: a computational investigation into the human representation and processing of visual information*. Freeman, San Francisco, California, w.h. freeman edition.
OpenAI. 2023. Gpt-4 Technical Report, 3. arXiv:2303.08774 [cs].
Piantadosi, Steven T. 2024. Modern language models refute chomsky's approach to language. In Edward Gibson and Moshe Poliak, editors, *From fieldwork to linguistic theory: A tribute to Dan Everett (Empirically Oriented Theoretical Morphology and Syntax 15)*. Berlin: Language Science Press, pages 353—414.
Rizzi, Luigi. 1990. *Relativized minimality*. Number 16 in Linguistic inquiry monographs. MIT Press, Cambridge, Massachusetts.
Roberts, Ian. 2019. *Parameter Hierarchies and Universal Grammar*. Oxford University Press, Oxford, United Kingdom, 1 edition, jun 20. DOI: 10.1093/oso/9780198804635.001.0001.
Sanguinetti, Manuela, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. Postwita-UD: an Italian Twitter Treebank in universal dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May.
Sarti, Gabriele. 2020. Umberto-MTSA@ AcCompl-It: Improving Complexity and Acceptability Prediction with Multi-task Learning on Self-Supervised Annotations. In Valerio Basile, Cristina Bosco, Rodolfo Delmonte Rodda, Pierpaolo Marcuzzo, Viviana Patti, Giovanni

Rossato, Tommaso Caselli, Rachele Sprugnoli, Andrea Cimino, Felice Dell'Orletta, Elisa Ferracane, and Gabriele Sarti, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Communication Tools for Italian (EVALITA 2020)*, Online, December. CEUR-WS.org.

Shieber, Stuart M. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8(3):333–343, 8.

Siegelman, Noam, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, Marco Marelli, Timothy C. Papadopoulos, Athanassios Protopapas, Satu Savo, Diego E. Shalom, Natalia Slioussar, Roni Stein, Longjiao Sui, Analí Taboh, Veronica Tønnesen, Kerem Alp Usal, and Victor Kuperman. 2022. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863, 12.

Sprouse, Jon and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, 2(1):1–32. Ubiquity Press.

Starke, Michal. 2001. *Move Dissolves into Merge: a Theory of Locality*. Ph.D. thesis, Université de Genève, Genève, Switzerland.

Tang, Yong, Jens R. Nyengaard, Didima M.G. De Groot, and Hans Jrgen G. Gundersen. 2001. Total regional and global number of synapses in the human brain neocortex. *Synapse*, 41(3):258–273, sep 1.

Trotta, Daniela, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and Cross-Lingual Acceptability Judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December. arXiv: 1706.03762.

Villata, Sandra, Paolo Canal, Julie Franck, Andrea Moro, and Chesi Cristiano. 2015. Intervention Effects in Wh-Islands: An Eye-Tracking Study. In *Architectures and Mechanisms for Language Processing (AMLaP 2015)*, pages 195–195, Valletta, Malta, September.

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848, 10.

Xia, Fei, Bin Li, Yixuan Weng, Shizhu He, Bin Sun, Shutao Li, Kang Liu, and Jun Zhao. 2022. Lingjing at SemEval-2022 Task 3: Applying DeBERTa to lexical-level presupposed relation taxonomy with knowledge transfer. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 239–246, Seattle, United States, July. Association for Computational Linguistics.

Zamparelli, Roberto, Absar Chowdhury, Shammur, Dominique Brunato, Cristiano Chesi, Felice Dell'Orletta, and Giulia Venturi. 2022. Semeval-2022 Task3 (PreTENS): Evaluating Neural Networks on Presuppositional Semantic Knowledge. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle, United States, July. Association for Computational Linguistics.