ChatGPT as an informant

Iris Mulders, Utrecht University, https://orcid.org/0000-0002-3695-9059
E.G. Ruys, Utrecht University, https://orcid.org/0000-0001-8997-8547

While previous machine learning protocols have failed to achieve even observational adequacy in acquiring natural language, generative large language models (LLMs) now produce large amounts of free text with few grammatical errors. This is surprising in view of what is known as “the logical problem of language acquisition”. Given the likely absence of negative evidence in the training process, how would the LLM acquire the information that certain strings are to be avoided as ill-formed? We attempt to employ Dutch-speaking ChatGPT as a linguistic informant by capitalizing on the documented “few shot learning” ability of LLM’s. We then investigate whether ChatGPT has acquired familiar island constraints, in particular the CNPC, and compare its performance to that of native speakers. Although descriptive and explanatory adequacy may remain out of reach, initial results indicate that ChatGPT performs well over chance in detecting island violations.

keywords: LLM, GPT, island conditions, negative evidence, Poverty of the Stimulus, CNPC, Adjunct Islands, island constraints, wh-movement

1. Introduction

One familiar argument in favor of an innate Universal Grammar is based on the supposed absence of negative evidence in the child’s intake, an instance of the Poverty of the Stimulus (POS) argument for UG. Consider the following passage from Braine (1971) (see also Baker 1979, Hornstein & Lightfoot 1981, Pinker 1986, Laurence & Margolis 2001; but also Cowie 1997):

> Information about what is not a sentence would appear to be necessary in order for the learner to reject hypothetical grammars and grammatical rules which are "overinclusive" (i.e. which generate all the acceptable strings, and which err only because they also generate unacceptable strings). Since such grammars generate all the good sentences to which the learner is exposed, how can he discover that they are wrong unless his input data contain information about nonsentences? [...] Moreover, in such cases the overinclusive grammar would clearly be simpler

> than the true grammar so that any ordinary simplicity measure would favor the wrong grammar [Braine (1971:157)]

Baker (1979) illustrates the relevant reasoning by means of the following data:

(1) a. The child seems to be happy
    b. The child seems happy
(2) a. John appears to be reluctant to leave
    b. John appears reluctant to leave

We may suppose that a language learner faced with the data in (1) and (2) might postulate a rule deleting *to be* so as to derive the b.-variants:

(3) X – *to be* – Y
    1 2 3
    ⇒ 1, ∅, 3 (Optional)

The resulting grammar overgenerates in that it allows (4b) next to (4a):

(4) a. The baby seems to be sleeping
    b. * The baby seems sleeping

It is difficult to see how a child might acquire the knowledge that (4b) is ill-formed, in the absence of specific evidence to this effect; hence this fact must somehow follow from aspects of UG.

In view of this argument, and what is known more generally as “the logical problem of language acquisition”, it is surprising that generative language models now produce large amounts of free text with few grammatical errors. Given the absence of negative evidence in the training process, how would the LLM have acquired the information that certain strings are to be avoided as ill-formed? In this paper we investigate whether English-speaking and Dutch-speaking ChatGPT has acquired familiar constraints on long movement, in particular the CNPC (Ross 1967), illustrated by the contrast in (5):

(5) a. To whom$_i$ did you tell t$_i$ [$_{NP}$ the story that John was speaking]?

b. * To whom$_i$ did you tell me [$_{NP}$ the story that John was speaking t$_i$]?

The reason we focus on the CNPC and similar constraints on long movement is as follows. Simple constructions such as (1a) and (1b) can be expected to be relatively frequent in everyday use. If so, the relatively low ((near-)zero) frequency of (4b) in comparison with (1b) and (2b) may prompt the successful hypothesis that (4b) is excluded. Such a statistical discovery procedure is less likely to be successful when it comes to complex pairs such as (5), where the well-formed case will also be infrequent. Of course, if the training corpus is large and diverse enough, much larger than the intake of human language learners, as is the case with the training corpus for LLMs such as GPT-3 (Brown et al. 2020), the contrast in (5) may also fall out from statistical considerations. While previous attempts at machine language learning (including statistical n-gram methods, neural networks, and string substitution procedures) have failed to achieve observational adequacy beyond a narrow range (see e.g. Huijbregts 2008, Berwick et al. 2011), ChatGPT's success in producing well-formed free text makes it plausible that it may also have acquired statistically more obscured constraints, such as the CPNC.

It is important to note that, should we find ChatGPT capable of distinguishing (5a) from (5b), the conclusions to be drawn from this observation will remain relatively limited, for several reasons. First of all, observational adequacy (a grammar's ability to distinguish well-formed strings from ill-formed ones, irrespective of structure and meaning) represents at most a very minor part of the knowledge acquired by a native speaker of a natural language, a point reiterated by Chomsky numerous times over the years. Compare the following passage from Berwick et. al. (2011):

> Put another way, language acquisition is not merely a matter of acquiring a capacity to associate word strings with interpretations. Much less is it a mere process of acquiring a (weak generative) capacity to produce just the valid word strings of a language. Idealizing, one can say that each child acquires a procedure that generates boundlessly many meaningful expressions, and that a single string of words can correspond to more than one expression. [Berwick et. al. 2011:1212]

Secondly, even if an LLM were found to have achieved *observational* adequacy –a finding that might potentially result from research along the lines reported here– , and were then found to have moved on to *descriptive* adequacy (the ability to associate strings with their correct structures and meanings) –which our present approach would not allow us to detect– it is obvious that current LLM's will not achieve *explanatory* adequacy as a theory of the human Faculty of Language (Chomsky 2013, 2016, 2023). This is so not only because the diversity and amount of material LLMs are trained on exceeds a human language learner's intake by several orders of magnitude, but also because there is no evidence that the methodology underlying these models would be incapable of acquiring languages that violate known properties of human languages.

A final caveat is in order. While it is plausible that negative evidence represents at most a very small part of ChatGPT's training, strictly speaking it is not entirely absent. First, Brown et al. (2020:8) reports that the training set for GPT-3 contains two unspecified collections of books, as well as English-language Wikipedia. These sources contain at least some discussion of the CNPC, including examples of violations explictly marked as ill-formed, which in principle constitutes negative evidence.[1] In addition, Ouyang (2022) reports that the "Instruct"-variants of GPT-3 are finetuned by reinforcement learning through human feedback, focusing on "writing in clear language [p. 37]", *inter alia*. This also holds for GPT-4 (OpenAI 2023:2). Such fine tuning may also, in principle, inject negative evidence on CNPC violations in the training process. We assume that these factors have not played a meaningful role in overcoming the overall absence of negative evidence, and we expect that a similarly trained LLM without these sources would not perform significantly worse on cases like (5), but this must remain speculation at this time.

In spite of these caveats, we feel there is some interest in testing whether current LLM's refute this naive version of the POS argument: that in the absence of negative evidence, the unacceptability of unacceptable utterances cannot be acquired (without UG). An additional purpose of the experiments reported here was to explore the viability of the methodology we adopted, in service of future work on the linguistic competence of LLM's.

2. Test methodology: ChatGPT as an informant

Several methodologies might be considered for ascertaining an LLM's ability to distinguish well-formed expressions from ill-formed ones.[2] One obvious method would be to search for

[1] See e.g the Wikipedia page *Wh-movement* at https://en.wikipedia.org/w/index.php?title=Wh-movement&oldid=1208346435

[2] Non-categorical results can be obtained by comparing the relative probabilities assigned to strings making up minimal pairs, as in Hu et al. (2024), q.v. for additional references on relevant methodologies. See also Ozaki et

violations of a particular constraint in the LLM’s output. However, the very reasons which make an island constraint like the CNPC an interesting test case for an effect of the absence of negative evidence –the expected rarity of both well-formed and ill-formed structures of the relevant kind– also entails that we could not conclude from the absence of CNPC violations in a corpus of LLM generated text that the LLM obeys the CNPC.

We opted instead for a methodology that treats GPT as a regular informant, by capitalizing on its documented few-shot learning ability. Brown et al. (2020:7) provide the following illustration showing three methods of prompting GPT-3 to translate English to French: *zero shot* (instruction only, no examples), *one-shot* (instruction and one example), and *few-shot* (instruction and a few examples):

zero-shot	**one-shot**	**few-shot**
Translate English to French:	Translate English to French:	Translate English to French:
Cheese =>	sea otter => loutre de mer	sea otter => loutre de mer
	Cheese =>	peppermint => menthe poivrée
		plush girafe => girafe peluche
		Cheese =>

Brown et al. (2020) observe that with larger models, one-shot and few-shot prompts significantly increase performance on a range of tasks (see also Kojima (2022) and references cited there for the importance of prompt construction).

We tested ChatGPT’s grammatical competence by providing prompts consisting either of an instruction to act as a linguistic informant (zero shot), or of an instruction accompanied by one (one shot) or two (few shot) minimal pairs labeled as grammatical or ungrammatical. The English prompts we employed are given in (6)-(8) (see the supplementary online materials for the Dutch prompts):

(6) You are a native speaker of English. Question: is the following sentence grammatical?
(target sentence)
Answer: the sentence is

(7) You are a native speaker of English. Look at the following sentence:

al. (2022) and references cited there for discussion on how different LM “surprisal” measures can be interpreted as reflecting grammaticality.

How many books did you read?

That sentence is grammatical.

Look at the following sentence:

How many did you read books?

That sentence is ungrammatical.

Look at the following sentence:

(target sentence)

That sentence is

(8) You are a native speaker of English. Indicate whether the following sentences are grammatical or ungrammatical.

Sentence: Whose books did you read?

Grammatical.

Sentence: Whose did you read books?

Ungrammatical.

Sentence: I know how many books he has read.

Grammatical.

Sentence: I know how many he has read books.

Ungrammatical.

Sentence: (target sentence)

We expected that the completions to these prompts would be interpretable as grammaticality judgments.[3] We explored the viability of this method of elicitation by testing a small set of example pairs, which resulted in completions that could be interpreted as grammaticality judgments without difficulty or ambiguity.

Another methodological issue we needed to address involves example sentence construction. We needed to prevent the LLM from rejecting island violations based on a proxy measure, falsely creating the impression it had acquired the island constraint. In many cases, proscribed extraction from an island creates a very low frequency n-gram inside the island, rendering it susceptible to learning strategies involving weak string substitutability (Clark &

[3] Note that it is technically irrelevant whether ChatGPT “understands” the question, or “intends” by its completion to indicate that the target sentence is grammatical or not. If we were to find that it performs above chance, it has obviously acquired some ability to distinguish expressions that violate the CNPC from ones that do not, demonstrating that the distinction can be detected in the absence of negative evidence.

Eyraud 2007) or low-plausibility bi-grams or tri-grams (Reali & Christiansen 2005). Consider (9):

(9) * Dit is het portret waar$_i$ ik [$_{NP}$ het kind dat op t$_i$ leek] herkende
This is the portrait which I [the child that looked like] recognized

The labeled complex NP will not appear in the corpus as substitutable for an NP; the underlined sequence is low probability. However, since not all CNPC violations result in such substrings, and island violations that do not are nonetheless ill-formed, rejection of examples like (9) would not properly demonstrate acquisition of the CNPC. For this reason our test items are constructed in such a way that the island violations do not, in our judgment as (near-)native speakers, result in extremely low probability local substrings, so that detection of the island violation requires parsing most or all of the expression. To achieve this, we typically chose predicates inside the islands that (unlike *look like* in (9)) provide an argument position for the extracted element but are also grammatical without that argument, e.g. *aanvallen* 'attack' in (10):

(10) * Dit is de man die$_i$ [$_{NP}$ de vrouw die t$_i$ wilde aanvallen] mij weggejaagd had.
This is the man who [the woman who wanted to attack] chased me away

In experiment 1, we observed LLM performance on our test items that we suspected might in some conditions be at least at human levels. For confirmation, we complemented the experiment with human trials on our Dutch items (experiment 2). This also served to assess whether the LLM's responses as a linguistic informant differed from human responses to a few shot forced choice grammaticality judgment task.

3. Experiment 1: ChatGPT

3.1 Language models

We tested two generative pre-trained transformer models: GPT 3.5 Turbo and GPT 4. Brown et al. (2020) reports that GPT 3 is a 1.7 billion parameters model, trained on ~195×10^9 words of text, ~669×10^6 or 0.34% of which are reported as Dutch. For the later OpenAI models no technical details have been made public (OpenAI 2023:2).

3.2 Materials

Set 1: base line examples. We tested 10 simple well-formed/ill-formed minimal pairs with each prompt type, to ascertain whether the model "understood the task", i.e., whether the LLM's replies matched those one would expect from a competent speaker's response to a request for a grammaticality judgment:

(11) a. Wiens boeken heb je gelezen?
whose books have you read
b. * Wiens heb je boeken gelezen?
whose have you books read?

Set 2. English island violations. We tested 10 English well-formed/ill-formed minimal pairs involving long movement constraints, for comparison with the Dutch target cases:

(12) a. This is a book I think you will like.
b. * This is a book I like the man who was reading.

Set 3. Dutch island violations. We tested 32 Dutch well-formed/ill-formed minimal pairs involving long movement constraints. 26 were CNPC violations (Ross 1967), involving 6 movement types (Relativization / *wh*-movement / Extraposition / V1 / *Tough* movement / comparative operator movement), 6 were violations of the Adjunct Condition case of the CED (Huang 1982).[4]

(13) a. Dit is de man die$_i$ t$_i$ [$_{NP}$ de vrouw die wilde aanvallen] weggejaagd had.
This is the man who$_i$ t$_i$ [the woman who wanted to attack] had chased away
b. * Dit is de man die$_i$ [$_{NP}$ de vrouw die t$_i$ wilde aanvallen] mij weggejaagd had.

[4] Our experiment differs from some previous work on machine learning of filler-gap constructions (see e.g. Ozaki et al.) in that in our items (un)grammaticality depends only on whether a gap is inside an island relative to its filler, not on a mismatch between present filler and absent gap, or vice versa.

This is the man who$_i$ [the woman who wanted to attack t$_i$] had chased me away

All test items can be found in the supplementary materials online.

3.3 Procedure

Each of the (10 + 10 + 32) × 2 = 104 items was presented once to each of the two LLM's using each of the three prompts described in section 2 by means of the "Playground" chat interface.[5] This resulted in 624 completions, which we manually scored as either matching or not matching our own grammaticality judgments, based on our intuitions as (near-)native speakers and the relevant syntactic literature.

3.4 Predictions

Based on our previous experience with ChatGPT, we expected it to provide completions that could be interpreted as grammaticality judgments. We further expected the item set 'baseline Dutch' to be easier than the item sets with the long movement constraints, since it contains much simpler and probably more frequent sentences in the grammatical condition, and the ill-formedness of the structures in the ungrammatical condition could plausibly be inferred from their relative infrequency in comparison with the grammatical cases. We also expected the English data set to be easier than the comparable Dutch data set, considering the vastly larger exposure GPT has had to English than to Dutch, although an effect of transference could not be excluded.[6] For similar reasons, we expected GPT 4 to outperform GPT 3.5 Turbo, although technical data on the difference between these models have not been made public. Furthermore, we expected one-shot and few-shot prompts to improve performance, perhaps more so in GPT 4 than in GPT 3.5 Turbo (Brown et al. 2020).

3.5 Analysis, results and discussion

[5] Parameters were set as follows: temperature 0; max_tokens 256; top_p 1, frequency_penalty 0; presence_penalty 0.

[6] See Wendler et al. (2024) for some relevant findings.

To get a first impression of the results, a plot showing the proportion of matching completions in the three item sets for each model, prompt, and grammaticality level can be found in Figure 1.

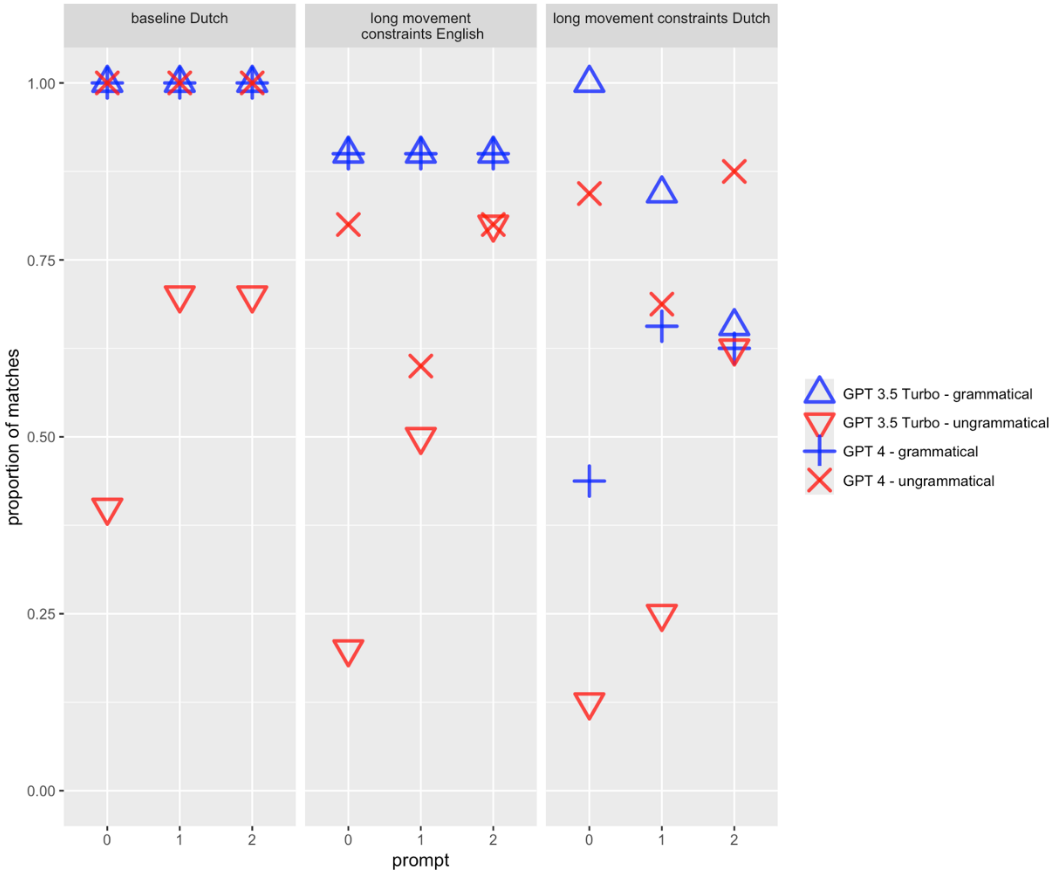

Figure 1. Proportion of matching completions in the GPT experiment for each item set, model, prompt, and grammaticality level (grammatical vs. ungrammatical). Each point is calculated over 10 completions in the baseline Dutch and in the English item sets, and over 32 completions in the long movement Dutch item set.

We performed a generalized linear mixed effects analysis of the relationship between the matching completions, and item set, prompt, model, and grammaticality, using R 4.3.2 (R Core Team 2023), Rstudio 2024.4.1.748 (Posit team 2024), and lme4 1.1.35.2 (Bates et al. 2015).

As a first approximation of an answer to the question whether ChatGPT's completions in this task can indeed be interpreted as grammaticality judgments, we looked at the overall proportion of matches. In the raw data, the overall proportion of matches was .71; the question is whether this is better than chance. To answer this question, we built an intercept-only generalized mixed model with 'match' as the dependent variable, and a random intercept for item.[7] The estimate for the proportion of matches in this model is .73, and its 95% confidence interval is .67 - .78. Since this confidence interval does not include .50, we conclude that overall, ChatGPT performs better than chance on this task, making it likely that it does “know” the concept of grammaticality in some sense, and employs it in completing the prompts.

In view of the fact that our research question focuses specifically on the long movement condition we felt justified in asking whether the LLM’s performance on item sets 2 and 3 (English and Dutch long movement) combined (leaving out the base line data in set 1) also exceeded chance. This turned out to be the case (estimate of the proportion of matches .67; 95% CI .62 - .72).

To the intercept-only model we then added item set, prompt, model, and grammaticality as fixed factors, in a step-wise fashion (in this order), as well as a by-item random slope for grammaticality, using AIC comparisons to ensure improved model fit (Hamaker et al. 2011). Next, we added two-way interactions in a step-wise fashion, in order of theoretical interest. We started with the interaction between prompt and model (cf. Brown et al. 2020), which did not improve model fit, so it was removed from the model. Next up were the interaction between model and grammaticality and the interaction between prompt and grammaticality (both improved model fit). The interaction between item set and prompt did not improve model fit so it was removed. Adding the interaction between item set and grammaticality, or item set and model, resulted in non-converging models. Finally, we added by-item random slopes for model. Attempts to fit additional random slopes resulted in models not converging, or giving singularity warnings.

To test the fixed effects for significance, we performed type III Wald chi-square tests using the car package 3.1.2 (Fox 2019) – the results, along with the formula for the final model, can be found in Appendix A. We also performed planned comparisons for item set, using emmeans 1.9.0 (Lenth 2023), see Appendix B.

For the differences between the three item sets, results were in line with our expectations. Performance on the baseline-Dutch item set was better than on the long-movement-constraints-Dutch set, suggesting that long-movement constraints may be more difficult to “learn” for the models than basic sentences, at least in Dutch. The models did better on the long-movement-constraints-English set than on the long-movement-constraints-Dutch set, which is in line with our expectation that the models will do better on English than on Dutch because of their larger exposure to English.

[7] Formula: match ~ (1|item)

We expected GPT 4 to do better than GPT 3.5 Turbo. We found a significant interaction between model and grammaticality, indicating that GPT 3.5 Turbo has a “yes-bias”: it likes to say ‘grammatical’ more than it likes to say ‘ungrammatical’ (overall it says ‘grammatical’ 73% of the time, vs. GPT 4 only 45% of the time), resulting in seemingly great performance on the grammatical sentences, but poor performance on the ungrammatical sentences (cf. Dentella et al. 2023). As Figure 2 illustrates, GPT 4 is far more balanced in its completions, and consequently the better model of the two.

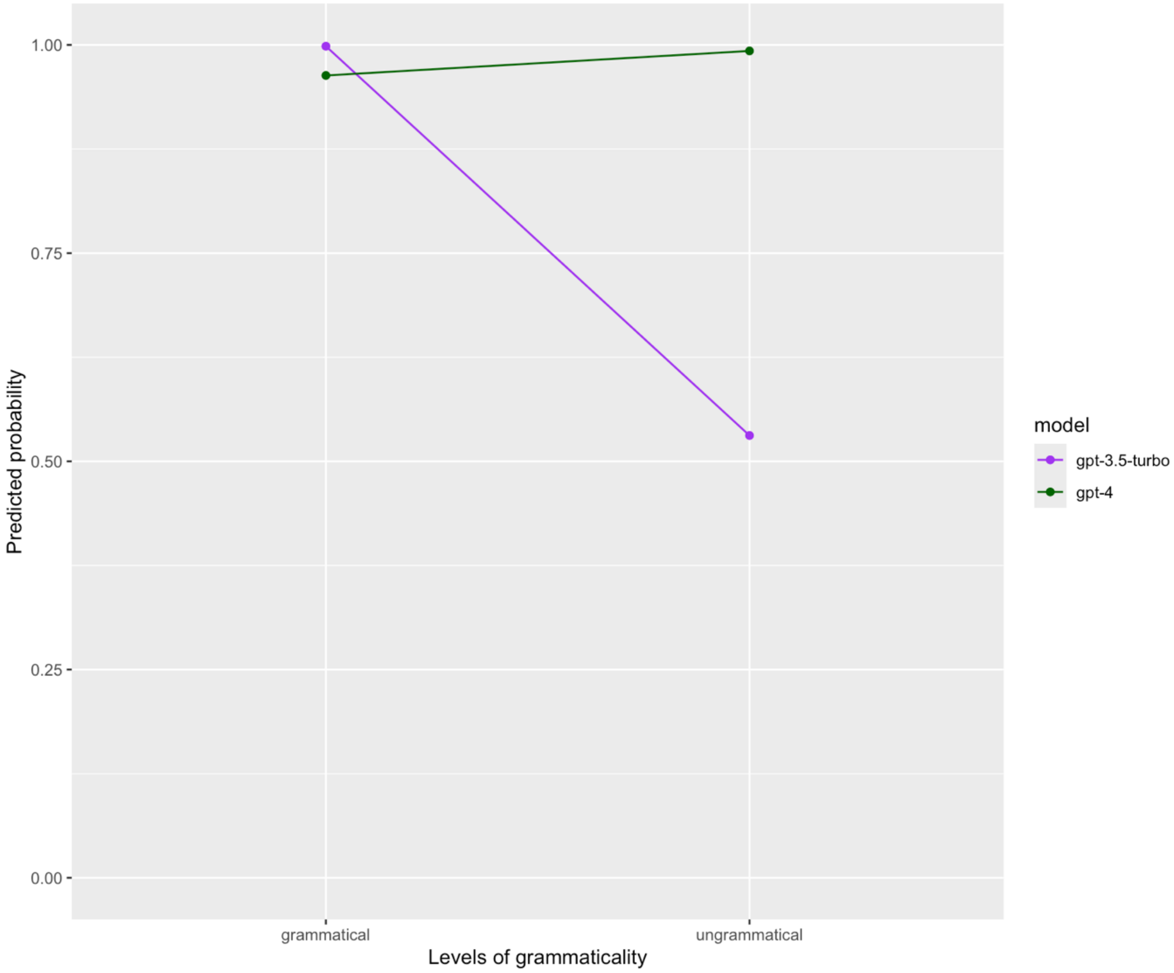

Figure 2. Predicted match probability for grammatical/ungrammatical sentences for GPT 3.5 T urbo and GPT 4.

Finally, we found a significant interaction between prompt and grammaticality, indicating that in ungrammatical items, having prompts with more shots significantly improves performance (see Figure 3). Using few-shot prompts helps to close the performance gap between grammatical and ungrammatical items (i.e. it helps to “fix” the grammaticality bias).

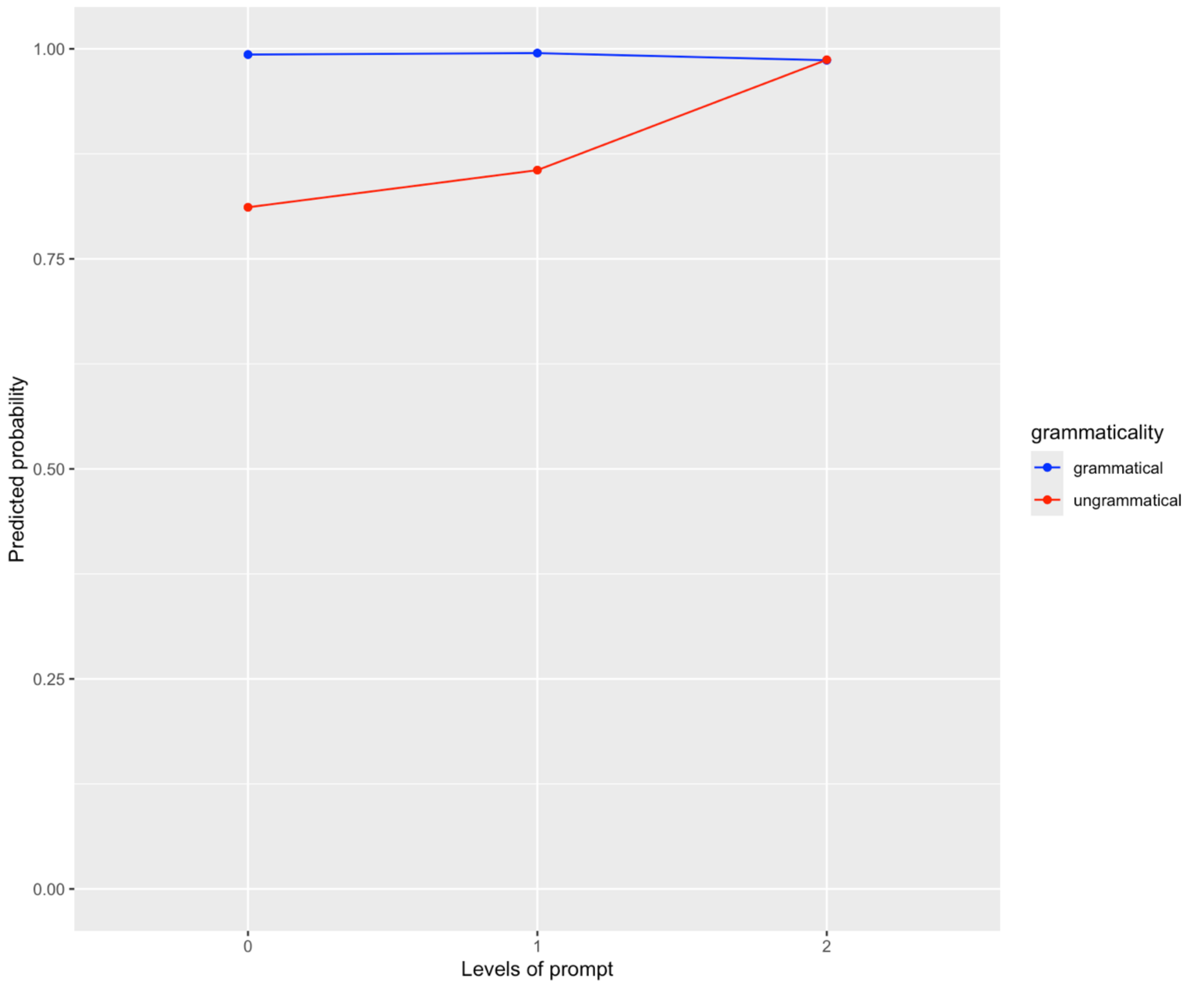

Figure 3. Predicted match probability for grammatical/ungrammatical sentences for zero vs. one vs. few-shot prompts.

Overall, GPT 4 with a few-shot prompt has the best performance.

4. Experiment 2: humans versus few-shot ChatGPT 4

As we have seen, the highest performing LLM (GPT 4 with a few-shot prompt) performed surprisingly well in experiment 1. In order to assess whether it reaches (or perhaps even surpasses) a human level of performance, we ran the same task[8] on native speakers of Dutch in experiment 2.

4.1 Participants

[8] For an evaluation of yes/no acceptability judgment tasks very similar to the one we used, and for more sensitive grammatical island detection methods, see Sprouse & Almeida (2017), Sprouse, Wagers & Phillips (2013), and references cited there.

We recruited 20 participants on Prolific, all (self-reported) Dutch nationals residing in the Netherlands with Dutch as their first language, without dyslexia, 12 females and 8 males, ranging in age from 20 to 54, mean age 35 years. Participants were paid £3 for their participation.

4.2 Materials

The materials were identical to the Dutch materials described in section 3.2. They were divided into two lists following a Latin square design such that each participant saw each item in either its grammatical or ungrammatical version, and as many grammatical as ungrammatical sentences overall.

4.3 Procedure

We conducted a grammaticality judgment experiment on Qualtrics. First, participants confirmed their Prolific ID. Next, they saw an information letter containing the approval from the ethical committee. They were then shown instructions describing the few-shot prompt (which was repeated for every sentence) and instructing them to read each (bold faced) test sentence carefully and indicate if it was grammatical or ungrammatical by pressing the corresponding button. Full instructions can be found in the supplementary materials online.

We first presented 10 items from stimulus set 1 (base line examples) in a randomized block. This was followed by a short instruction explaining that the next 32 sentences would be more complicated, and reminding the participants to assess them only for grammaticality, not on whether a paraphrase might be preferred. We then presented 32 items from stimulus set 3 (Dutch island violations) in a second randomized block. After a thank-you page, participants were redirected to Prolific and paid.

On average, participants completed the task in 10 minutes (range 6-14 minutes).

4.4 Analysis and results

The anonymous data can be found in the online supplementary materials.

4.4.1. Humans

We performed a GLMER analysis of the relationship between the matching completions and item set and grammaticality. Unsurprisingly, humans performed above chance (estimate in the intercept-only model for both item sets together is .87 matches, 95% CI .80-.92; for the island item set .79 matches, 95% CI .73-.83). The formula and results for the full model can be found in Appendix C. Humans performed better on the base line item set than on the island item set, and they showed a marginally significant ungrammaticality bias.

4.4.2 Humans versus GPT 4 with a few-shot prompt

Table 1 provides the raw mean proportions of matches for the human participants for both item sets, compared to GPT 4 with a few-shot prompt. Table 2 shows the same for grammatical versus ungrammatical items.

	Item set 1 (base line)	Item set 3 (islands)	Grand total
GPT 4	1.00	0.75	0.81
Humans	0.97	0.77	0.82

Table 1. Proportions of matching responses on both item sets, GPT 4 vs. humans.

	Grammatical	Ungrammatical	Grand total
GPT 4	0.71	0.90	0.81
Humans	0.77	0.86	0.82

Table 2. Proportions of matching responses on (un)grammatical items, GPT 4 vs. humans.

To be able to compare the data from the humans with the data from few-shot prompt GPT 4 in a balanced way, we calculated the proportion of matches for the humans per item. We then arcsine transformed these proportions, as well as GPT's match variable. We built a linear mixed effects model using nlme 3.1.164 (Pinheiro & Bates 2000, Pinheiro et al. 2023), containing item set, grammaticality, entity type (human or GPT) and its interactions with item set and grammaticality, and random intercepts for items. The results and model formula can be found in Appendix D.

Overall, performance was better on the baseline item set than on the island item set, and significantly more matches were observed on ungrammatical items than on grammatical items. None of the interactions were significant, indicating that few-shot GPT 4 did not do better or worse than humans, as could already be expected from the raw means in tables 1 and 2.

5. Conclusions

We found that the completions to our prompts provided by the LLMs (especially GPT 4) could be interpreted as grammaticality judgments. We observed that, in the vast majority of cases, completions to our prompts consisted merely of the strings "grammatical." and "ungrammatical." or their Dutch translations. Together with our finding that overall, the LLMs performed above chance, this indicates that our methodology can successfully be used to assess an LLM's grammatical competence. As expected, selecting a few-shot prompt can help elicit correct responses. Overall, we conclude that ChatGPT can work well as an informant for purposes of assessing its linguistic competence.

We conclude further, unsurprisingly at this point, given the overall performance of ChatGPT, that large transformer models provide a means of overcoming the lack of negative evidence in acquiring the effects of constraints on grammaticality that operate in local contexts (item set 1), presumably due, at least in part, to the large size of their training sets. More surprisingly, we conclude that these means are probably also sufficient to overcome the lack of negative evidence in acquiring the effects of constraints on long movement in complex multi-clausal examples (item sets 2 and 3), as the LLMs also performed better than chance in these two item sets. In fact, we cannot distinguish the performance of GPT 4 with a few-shot prompt from the performance of humans on this task. Observe, furthermore, that this performance was achieved on examples that were, for the most part, deliberately constructed so as to exclude local detection of the island constraint violation inside the island, so that the violation could be expected to be relatively difficult to detect, and for a large part on examples in a language that formed only a small proportion of the training data. Thus, in response to the question raised in section 1, we conclude that large language models indeed have the ability to acquire a weak generative capacity that includes the effects of statistically relatively obscured constraints such as the CNPC and the Adjunct Condition, thus providing one partial solution to "the logical problem of language acquisition".

Supplementary materials and data archive

All test items, participant instructions, raw data and analysis scripts can be found at https://doi.org/10.24416/UU01-WQBJ1N.

Acknowledgments

Many thanks to the audience at De Grote Taaldag 2024 and to two anonymous *Nota Bene* reviewers for helpful comments, and to Kirsten Schutter, Piet van Tuijl and Hugo Quené for their invaluable help with the statistical analysis. All remaining errors are our own.

References

Baker, C.LeRoy. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 10(4). 533-581.

Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1-48. https://doi.org/10.18637/jss.v067.i01.

Berwick, Robert C., Paul Pietroski, Beracah Yankama & Noam Chomsky (2011), Poverty of the stimulus revisited. *Cognitive Science* 35. 1207–1242.

Braine, Martin D. S. 1971. On two types of models of the internalization of grammars. In D. I. Slobin (ed.), *The Ontogenesis of Grammar: A Theoretical Symposium*. New York: Academic Press.

Brown, Tom B., Benjamin Mann, Nick Ryder, Malanie Subbiah, Jared D. Kaplanet al. 2020, Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan & H. Lin (eds.), *Advances in Neural Information Processing Systems* 33.

Chomsky, Noam. 2013. Problems of projection. *Lingua* 130. 33-49.

Chomsky, Noam. 2016. *What kind of creatures are we?* Columbia University Press.

Chomsky, Noam. 2023. The false promise of ChatGPT. *The New York Times*, 2023-03-08.

Clark, Alexander & Eyraud, Rémi. 2007. Polynomial identification in the limit of substitutable context-free languages. *Journal of Machine Learning Research* 8. 1725-1745.

Cowie, Fiona. 1997. The logical problem of language acquisition. *Synthese* 111. 17–51.

Dentella, Vittoria, Fritz Günther & Evelina Leivada. 2023. Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, *120*(51). doi:10.1073/pnas.2309583120.

Fox, John & Sanford Weisberg. 2019. *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Hamaker, Ellen L., Pascal van Hattum, Rebecca M. Kuiper & Herbert Hoijtink. 2011. Model selection based on information criteria in multilevel modeling. *Handbook of advanced multilevel analysis*. 231-255.

Hornstein, Norbert & David Lightfoot (eds.). 1981. *Explanation in Linguistics, The Logical Problem of Language Acquisition*. Longman.

Hu, Jennifer, Kyle Mahowald, Gary Lupyan, Anna Ivanova & Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. doi:10.48550/arXiv.2402.01676

Huang, C.-T. James. 1982. *Logical Relations in Chinese and the Theory of Grammar*. Diss. MIT.

Huijbregts, Riny. 2008. *Linguistic Argumentation and Poverty of the Stimulus Arguments*. ms., Utrecht University.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo & Yusuke Iwasawa. 2022. Large Language Models are zero-shot reasoners. doi:10.48550/arXiv.2205.11916

Laurence, Stephen & Eric Margolis . 2001. The poverty of the stimulus argument. *The British Journal for the Philosophy of Science* 52(2). 217-276.

Lenth Russell V. 2023. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.9.0, https://CRAN.R-project.org/package=emmeans.

OpenAI . 2023. *GPT-4 Technical Report*. doi:10.48550/arXiv.2303.08774

Ouyang, Long et al. 2022. Training language models to follow instructions with human feedback. doi:10.48550/arXiv.2203.02155

Ozaki, Satoru, Dan Yurovsky & Lori Levin. 2022. How well do LSTM language models learn filler-gap dependencies? *Proceedings of the Society for Computation in Linguistics (SCiL)* 5(1). 76-88. doi:10.7275/414y-1893

Pinheiro, José, Douglas Bates & R Core Team. 2023. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-164. https://CRAN.R-project.org/package=nlme.

Pinheiro, José & Douglas Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer. doi:10.1007/b98882

Pinker, Steven. 1986. Productivity and conservatism in language acquisition. In W. Demopoulos & A. Marras (eds.), *Language Learning and Concept Acquisition: Foundational Issues*. 54-79. Norwood, NJ: Ablex.

Posit team. 2024. *RStudio: Integrated Development Environment for R*. Boston, MA: PBC. http://www.posit.co/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reali, Florencia & Morten H. Christiansen. 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science* 29. 1007-1028.

Ross, John R. 1967. Constraints on variables in syntax. Doctoral dissertation, MIT, Cambridge, MA. Reprinted as Infinite Syntax! Norwood, NJ: Ablex, 1986.

Sprouse, Jon, Matt Wagers & Colin Phillips. 2012. A test of the relation between working-memory capacity and syntactic island effects. *Language*. 82-123.

Sprouse, Jon & Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: a Journal of General Linguistics* 2(1).

Wendler, Chris, Veniamin Veselovsky, Giovanni Monea & Robert West. 2024. Do Llamas work in English? On the latent language of multilingual transformers. doi:10.48550/arXiv.2402.10588

Appendix A

Formula: match ~ prompt + model + item_set + grammaticality + model:grammaticality + prompt:grammaticality + (1 + grammaticality + model | item)

	Chisq	df	Pr(>Chisq)	
prompt	7.5407	2	0.023044	*
model	2.4790	1	0.115374	
item_set	19.8478	2	<0.0001	***
grammaticality	9.1166	1	0.002533	**
model:grammaticality	26.1993	1	<0.000001	***
prompt:grammaticality	24.1099	2	<0.000001	***

Appendix B

Contrast	Raw match proportions	Estimate	SE	df	z.ratio	p.value	
baseline Dutch – long mov’t constraints English	.90 - .76	1.54	1.015	Inf	1.513	.2847	
baseline Dutch – long mov’t constraints Dutch	.90 - .64	3.61	0.894	Inf	4.040	.0002	***
long mov’t English – long mov’t constraints Dutch	.76 - .64	2.08	0.743	Inf	2.797	0.0143	*

Appendix C

Formula: match ~ item_set + grammaticality + (1+grammaticality|item) + (1+grammaticality|subject)

	Raw match proportions	Estimate	SE	z value	Pr(>|z|)	
item set	.97-.77	-2.7153	0.6133	-4.427	<0.000001	***
grammaticality	.77-.86	1.1981	0.6679	1.794	0.0729	.

Appendix D

Formula:

arcsine_match ~ human_or_GPT*item_set + human_or_GPT*grammaticality + human_or_GPT, random= ~ 1 | item

	Estimate	SE	df	t-value	p-value	
human_or_GPT	-0.0707058	0.08153839	122	-0.86715	0.3876	
item_set	-0.3809063	0.08153839	40	-4.67150	0.0000	***
grammaticality	0.1918326	0.06945729	122	2.76188	0.0066	*
human_or_GPT:item_set	0.0235855	0.16307677	122	0.14463	0.8852	
human_or_GPT:grammaticality	-0.2147333	0.13891457	122	-1.54579	0.1247	

Address for correspondence

E.G. Ruys
Institute for Language Sciences
Utrecht University
Trans 10
3512 JK Utrecht
The Netherlands
e.g.ruys@uu.nl

Co-author information

Iris Mulders
Institute for Language Sciences
Utrecht University
Drift 15
3512 BR Utrecht
The Netherlands

i.c.m.c.mulders@uu.nl