biolinguistics

PsychOpen GOLD

Forum

# Evaluating the Language Abilities of Large Language Models vs. Humans: Three Caveats

Evelina Leivada [1,2], Vittoria Dentella [3], Fritz Günther [4]

**[1]** *Department of Catalan Philology, Universitat Autònoma de Barcelona, Barcelona, Spain.* **[2]** *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.* **[3]** *Department of English and German Studies, Universitat Rovira i Virgili, Tarragona, Spain.* **[4]** *Institut für Psychologie, Humboldt-Universitat zu Berlin, Berlin, Germany.*

Biolinguistics, 2024, Vol. 18, Article e14391, https://doi.org/10.5964/bioling.14391

**Published (VoR):** 2024-04-19

**Handling Editor:** Patrick C. Trettenbrein, Max Planck Institute for Human Cognitive and Brain Sciences & University of Göttingen, Germany

**Corresponding Author:** Evelina Leivada, Universitat Autònoma de Barcelona, Departament de Filologia Catalana, 08193 Barcelona, Spain. E-mail: evelina.leivada@uab.cat

## Abstract

We identify and analyze three caveats that may arise when analyzing the linguistic abilities of Large Language Models. The problem of unlicensed generalizations refers to the danger of interpreting performance in one task as predictive of the models' overall capabilities, based on the assumption that because a specific task performance is indicative of certain underlying capabilities in humans, the same association holds for models. The human-like paradox refers to the problem of lacking human comparisons, while at the same time attributing human-like abilities to the models. Last, the problem of double standards refers to the use of tasks and methodologies that either cannot be applied to humans or they are evaluated differently in models vs. humans. While we recognize the impressive linguistic abilities of LLMs, we conclude that specific claims about the models' human-likeness in the grammatical domain are premature.

## Keywords

Artificial Intelligence, grammaticality, Large Language Models, probabilities

## 1. Introduction

In most subfields of Artificial Intelligence (AI), the success or failure of cutting-edge applications depends on the quantity and quality of the data. The quality of the data is also relevant in the context of our evaluations of the capabilities of such applications. Recent Large Language Models (LLMs) have been heralded as the most important innovation of the last decades, having the potential to transform fields like health care and education (Gates, 2023). In this context, rigorously determining the linguistic abilities of such models is critical. While data is important, language learning in humans, unlike LLMs, is not anchored exclusively on data-driven prediction (Felin & Holweg, 2024). In the process of language learning, humans also form hypotheses and theories about the input, while LLMs are trained to predict the next token in a sequence of tokens. In light of such differences, do LLMs behave linguistically in a way that can be called human or human-like?

To answer this question, Dentella et al. (2023) tapped into the grammatical understanding of LLMs. Specifically, they examined three models (GPT-3/text-davinci-002, GPT-3/text-davinci-003, and ChatGPT 3.5) and 80 humans in a judgment task that featured different linguistic phenomena. The results suggested that humans performed better than the tested LLMs, especially in recognizing the grammatically ill-formed prompts as such. Unlike humans, the LLMs showed a strong bias towards providing yes-responses irrespective of the grammaticality of the prompts. Moreover, upon repeated prompting of a sentence, humans remained largely stable in their opinions about its well-formedness, whereas the models oscillated a lot, revealing a stark absence of response stability (Dentella et al., 2023).

Hu et al. (in press) provide an illuminatingly different perspective on these results. Specifically, they reach different conclusions, arguing that the tested LLMs perform well to the point of aligning with human judgments on key grammatical constructions and of showing human-like grammatical generalization capabilities. Leivada et al. (in press) provide a very brief reply to Hu et al., but given the space constraints, several important issues that merit clarification and correction have not been addressed. We address them in this work. Before delving into them, it is important to highlight that Hu et al. make a valuable contribution towards understanding the language abilities of LLMs and, more importantly, towards figuring out why some methods of LLM evaluation show different results than others. Juxtaposing Dentella et al. (2023) and Hu et al. (in press), we identify three specific caveats that explain why different teams of scholars reach diametrically opposite conclusions about the language abilities of LLMs.

## 2. The Problem of Unlicensed Generalizations

The notion of unlicensed generalizations refers to the danger of interpreting results from one task as indicative of the model's overall capabilities, assuming that because a specific task performance entails certain abilities in humans, the exact same relation holds ipso facto for LLMs. Let us illustrate the problem with an example: Hu et al. (in press) argue that they re-evaluate LLM performance "using well-established practices and find that DGL's [Dentella et al.'s] data in fact provide evidence for how well LLMs capture human behaviors". Claiming that LLMs accurately capture human behavior on the basis of observing that their performance is equal or better than that of humans in a certain task (which it is not, but let us assume for now that it is; we will explain why this assumption is not correct in the section 'The problem of double standards') is analogous to claiming that a clock that shows the right time works correctly. As Guest and Martin (2023) argue, in the context of evaluating LLMs, such analogies are unfounded, because they *reverse* the order of the argument: If a clock works correctly, then it shows the correct time. Similarly, if the models are "human-like", we can infer that they will capture well human performance across tasks. However, we cannot legitimately infer human-like competence based on a task; similar to how we would be wrong to infer that a clock that shows the target time once a day works correctly. This reverses the nature of the argument. First, we need to determine both computational and behavioral alikeness through systematic testing that covers not only superficial similarity in terms of accuracy in a task, but also the type of reasoning used to perform the task, and then we can assert human-likeness.

It is interesting to observe how such differences are framed, depending on whether one finds that the models outperform humans or not. On the one hand, Dentella et al. (2023) suggest that their results challenge the claim that LLMs, at their current state of development, possess human-like language abilities, because they found that the tested models differ from humans in a specific task that taps into grammatical well-formedness. On the other hand, Hu et al. (in press) framed Dentella et al. (2023) as if making a more general claim about an alleged inability of LLMs to form *linguistic generalizations*. In fact, when presenting the scope of Dentella et al. in social media, they framed it in an even broader way: "Are Large Language Models good at language? A recent paper by Dentella, Günther, & Leivada (DGL) argues no."[1] Yet, Dentella et al. (2023) argue no such thing. Instead, they make a narrower claim about the inability of the tested models to discern the boundaries of grammar on a par with humans, as evidenced through a judgment task on specific grammatical phenomena. Similar to how target performance in a task does not license the generalization that LLMs have human-like capabilities, non-target performance in a task does not predict a ubiquitous failure across all possible tasks. Effectively, Hu et al. (in press) launch a strawman when they argue against a claim that Dentella et al. (2023) never made.

## 3. The Human-Like Paradox

Unlicensed generalizations and inappropriate framing of the linguistic abilities of LLMs have given rise to a paradox which we term 'the human-like paradox'. This refers to the problem of interpreting the results of an experiment through simultaneously affirming that LLMs behave in a human or human-like way, while lacking human comparisons.

[1] https://twitter.com/_jennhu/status/1754891894704746789

To illustrate this paradox with an example, Hu et al. (in press) argue that "LLMs show strong and human-like grammatical generalization capabilities". Yet, as also noted in Leivada et al. (in press), this claim is not backed up with human data. While Dentella et al. compare humans and LLMs using the same judgment task,[2] Hu et al. suggest that a better way of tapping into the language abilities of LLMs goes through obtaining direct probability measurements. Hu et al. believe that the results obtained from the two methodologies, grammaticality judgments and minimal-pair direct probabilities, *align* (i.e. the grammatical sentence in a 'grammatical-ungrammatical' pair is the one that both humans and models "prefer": Humans give it a higher rate of acceptability, while models find it less surprising than its counterpart). Yet we argue that this is not evidence that the tested models are sensitive to grammaticality or that a generalization has been learned. Another explanation for the observed alignment is possible: for instance, it could be the case that LLMs are less surprised by the sentences that humans find well-formed because these are more abundant in their web-scrapped training data. However, there is another elephant in the room: these claims about alignment and human-like abilities in a specific task occur in the *absence* of human comparisons. Hu et al. compare probability measurements in models to judgments in humans. This happens because they do not and cannot look *directly* into the minds of humans and observe probabilities over strings of words—assumptions about such internal representations and processes always have to be inferred from some other associated outcome variable (behavioral, electrophysiological, or neuroimaging, among others). In fact, the problem is not only practical; it is known that even when asked directly, humans are not good at probability assignment (Kahneman & Tversky, 1982). If "[t]he metalinguistic judgments elicited from LLMs through prompting are not the same as quantities directly derived from model representations" (Hu & Levy, 2023), the two methods are not the same, yet Hu et al. compare them as if they were.

Even if we grant that Hu et al. are right that LLMs perform well in probability measurements, this does not cast doubt on main claim put forth by Dentella et al.: The tested LLMs, at their current stage of development, do not perform in a human-like way in terms of providing judgments about grammatical well-formedness. On the other hand, there are tasks in which LLMs will perform better than humans (e.g., name 100 animals that start from 'm' in one minute); but outperforming humans would again run contrary

[2] While Hu et al. argue that the task reported in Dentella et al. was not exactly the same in humans vs. models—because humans were asked to press a button to answer as to whether a sentence was correct or not, while models answered the same questions without pressing a button—, we believe that these are differences that come with the territory. In fact, Hu et al. also elicited judgments, asking the models to respond with C or N (corresponding to the buttons humans pressed), but this instruction was ignored. As discussed in the next section, the models often mentioned other (task-irrelevant) things in their replies, in parallel with or even in the absence of C or N. If Hu et al. indeed maintain that pressing a button changes the results, enhancing human accuracy, the straightforward path forward would be to rerun the test asking humans to respond writing their answer in a text box. Our prediction is that the results will be the same, because pressing a button does not alter judgments in humans.

to the claim that the models behave in a human-like manner. The question is whether one chooses to direct their attention exclusively to tasks that mask the differences between humans and models to sustain a claim that LLMs have human-like language abilities, and whether such claims rely on the right level of comparison between the two. All in all, the buzzword 'human-like' should be used with caution for it is meaningful only when one has established what counts as 'human' in a specific experimental setting, using the *right* type of comparisons.

## 4. The Problem of Double Standards

Comparisons fulfil their function when the standards of evaluation are kept uniform. To reach their conclusions, Hu et al. (in press) replace grammaticality judgments with minimal-pair probability measurements. In other words, they replace absolute judgments for individual sentences with respect to a specific criterion (grammatical well-formedness) with relative judgments for pairs of sentences without any target criterion. Yet, grammaticality is not a matter of comparison. Humans are able to judge the well-formedness of individual sentences and they do not need minimal pairs to anchor or adjust their judgments, as demonstrated by the human participants' performance in Dentella et al. (2023). Also, grammaticality is not a matter of degree either (Leivada & Westergaard, 2020). A sentence is either grammatical or ungrammatical, in the sense that either it contains a violation of at least one rule of grammar, or it does not. From this perspective, asking the models 'Which of the following two sentences is more grammatically correct in English?', as Hu and Frank (2024) do, is wrong. No rule of grammar can be violated just a bit in sentence A but significantly more in sentence B such that A is more grammatical than B (Leivada & Westergaard, 2020). Hence, the question cannot be one of degree.[3] The question can be one of degree for humans, because unlike LLMs, humans also have judgments of *acceptability* (see Dentella et al., 2023 for the distinction): A person may like a sentence better than another one and assign it a higher acceptability rating. Models lack such preferences, likes, and dislikes; certain words do not make them feel in different ways, and they do not voluntarily (i.e. unprompted) project a specific identity through the use of specific words. This is probably the reason why the LLMs tested in Dentella et al. (2023) oscillate a lot in their judgments, unlike humans. Humans can consistently decide whether a sentence looks good or bad; LLMs cannot do so, thus they oscillate to maximize the probability of providing some target answers.

[3] This has consequences for the obtained results too. Hu and Frank (2024) classify the sentences in absolute terms, assigning them the label 'grammatical/good' or 'ungrammatical/bad', when presenting the results in the repository. Yet in the actual experiment, the question was one of *degree* (i.e. which sentence is more grammatically correct than the other), not of *kind*. An answer that suggests that a sentence A is more grammatical than a sentence B does not legitimize the inference that B is ungrammatical or bad, if the question is one of degree.

If we evaluate humans on a judgment task and models through direct probability measurements, which is the comparison Hu et al. (in press) make to find the alignment they claim, essentially, we hold different standards of evaluation for the two agents. This is the problem of double standards: using tasks and methodologies that either cannot be administered to both humans and models, or they are evaluated differently across the two. Hu et al. (in press) not only compare results obtained from different methods (judgments for humans vs. probabilities for LLMs), but they also adopt a different locus of comparison for each: *individual sentences* for humans vs. *pairs* of sentences for LLMs. As Leivada et al. (in press) note, one needs to compare apples to apples. This is a matter with important consequences because double standards can help inflate the performance of the models, leading to erroneous claims about "human-like" LLM behavior.

Take for instance Hu et al.'s (in press) claim that a minimal-pair analysis of probabilities shows "at- or near-ceiling performance" for LLMs. If we hold the models accountable at the same level as humans (i.e. individual sentences), a very different picture is observed. Assuming that probabilities really are informative of grammaticality, a surprisal threshold should exist in order to discriminate what is grammatical from what is not. Yet, when re-analyzing the probabilities given in the Hu et al. (in press) dataset, Leivada et al. (in press) find that the mean difference between ungrammatical and grammatical sentences exists in the context of a *massive overlap* between the two distributions. Even with an optimal surprisal threshold that results in the highest possible classification accuracy, this accuracy is only at .60 for davinci2 and .58 for davinci3[4]. While this is significantly better than random guessing ($p$ = .005 for davinci2, $p$ = .019 for davinci3), it is very similar to the judgment-based overall accuracies of .59 (davinci2) and .56 (davinci3) already reported as significant by Dentella et al., and far from the "at- or near-ceiling performance" that Hu et al. claim based on their minimal-pair comparisons (Leivada et al., in press). To return to the critical assumption about human-likeness mentioned in Section 2, Hu et al.'s results do not provide reliable "evidence for how well LLMs capture human behaviors", because they have been compromised by the problem of double standards. Consequently, when the LLM probabilities are subjected to a test regime more similar to that applied to humans, the claims made by Hu et al. (in press) need to be modified to a level where they are far more in line with the original findings by Dentella et al. (2023). Of course, one may argue that the minimal-pair comparison is necessary to isolate grammaticality from other factors that may influence surprisal, such as word frequencies or semantic effects, and is thus necessary to arrive at controlled comparisons. Rather than salvaging this approach, this argument further reveals its shortcomings: Surprisal in the manner derived by Hu et al. is just a function of the

[4] The accuracy rates are slightly lower when using alternative measures to the sum surprisal measure over all words in a sentence used by Hu et al. (in press) and for this analysis, namely the average surprisal (sum surprisal divided by the number of words in a sentence) or the maximum surprisal for a word in a sentence.

raw probability of observing a sentence in a corpus, without being anchored to any specific target criterion (such as grammaticality). If LLMs need specific comparisons in order to tell apart grammatical from ungrammatical sentences, this already counts as an inherent discrepancy from humans, who are able to make such judgments without such a comparison.

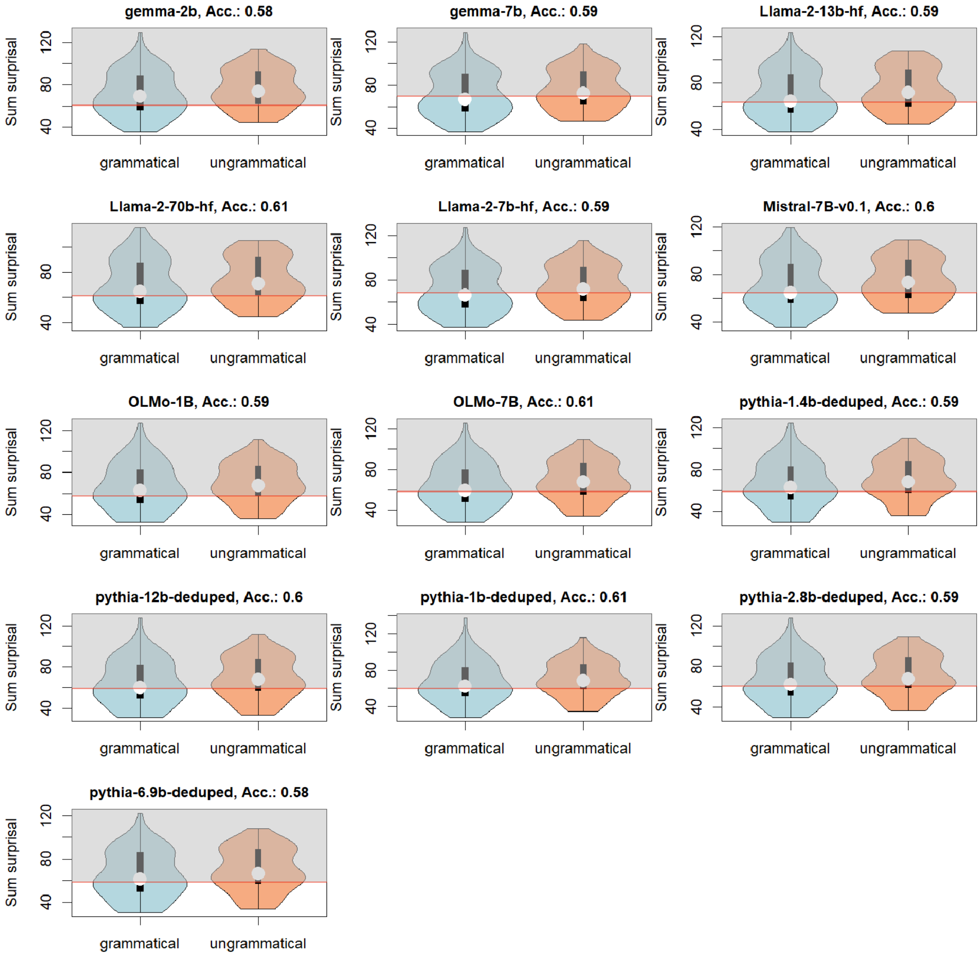
Selecting the right tasks for evaluating LLMs lies at the heart of the problem of double standards. Returning to the alleged superiority of probability measurements to judgment tasks, Hu and Frank (2024) re-affirm this position. They obtain probabilities from 13 open-source autoregressive language models that range in size from 1B to 70B parameters, using two datasets: DGL (Dentella et al., 2023) and BLiMP (Warstadt et al., 2020). Their results lead them to put forth the claim that elevated task demands (which they take to be high in judgment tasks and low in probability measurements) may mask the true linguistic abilities of smaller models. This is another instance of the problem of double standards hampering the results and weakening the claims made on their basis: The association 'judgments-high task demands' and 'probabilities-low task demands' does not hold for humans; hence it cannot serve as an adequate basis of experimentally comparing humans and LLMs. We cannot obtain direct probabilities from humans through peeking into their neurons, and definitely not by peeking directly into their mental representations either. Moreover, providing judgments of well-formedness is extremely easy for humans, making hard to justify the link of this task to high cognitive demands.

Since Hu and Frank (2024) obtain new probability values for the two tested datasets, DGL and BLiMP, we re-analyzed both sets of probabilities individually for each of the 13 models, along the lines described in Leivada et al. (in press) for the re-analysis of Hu et al. (in press). Specifically, we considered the absolute sum surprisal of individual sentences, both grammatical and ungrammatical, and checked whether a surprisal threshold that acts as a cut-off for grammaticality exists. The existence of this threshold is of paramount importance, because without it, the claim of Hu and Levy (2023) and Hu et al. (in press) about the alleged superior ability of probability measurements to capture the generalization capabilities of LLMs lacks foundation. Succinctly put, if the most optimal of all possible surprisal thresholds does not boil down to a cut-off point that permits mapping probabilities above it to grammatical sentences and probabilities below it to ungrammatical sentences, it is not clear that probabilities truly are an index of internalized grammatical knowledge (i.e. of having internally generalized anything that helps the model to tell apart well- from ill-formed sentences). If they are not, any study that compares probabilities to judgments and finds the former faring better (e.g., Hu & Frank 2024; Hu & Levy, 2023; Hu et al., in press) is possibly built on a foundation that merits reconsideration.

The results of this new analysis confirm the findings of Leivada et al. (in press). As Figure 1 shows for DGL and Figure 2 for BLiMP, accuracies in both datasets remain just slightly above chance, ranging from .55 to .61. We find that roughly comparable parts of the two distributions end up in areas above and below the optimal threshold to discriminate grammatical from ungrammatical sentences. If probabilities were a good proxy for grammaticality, one would expect a clearer difference between the distributions, with ungrammatical sentences clustering together and showing a clear concentration above the threshold, and vice versa for grammatical sentences.

**Figure 1**

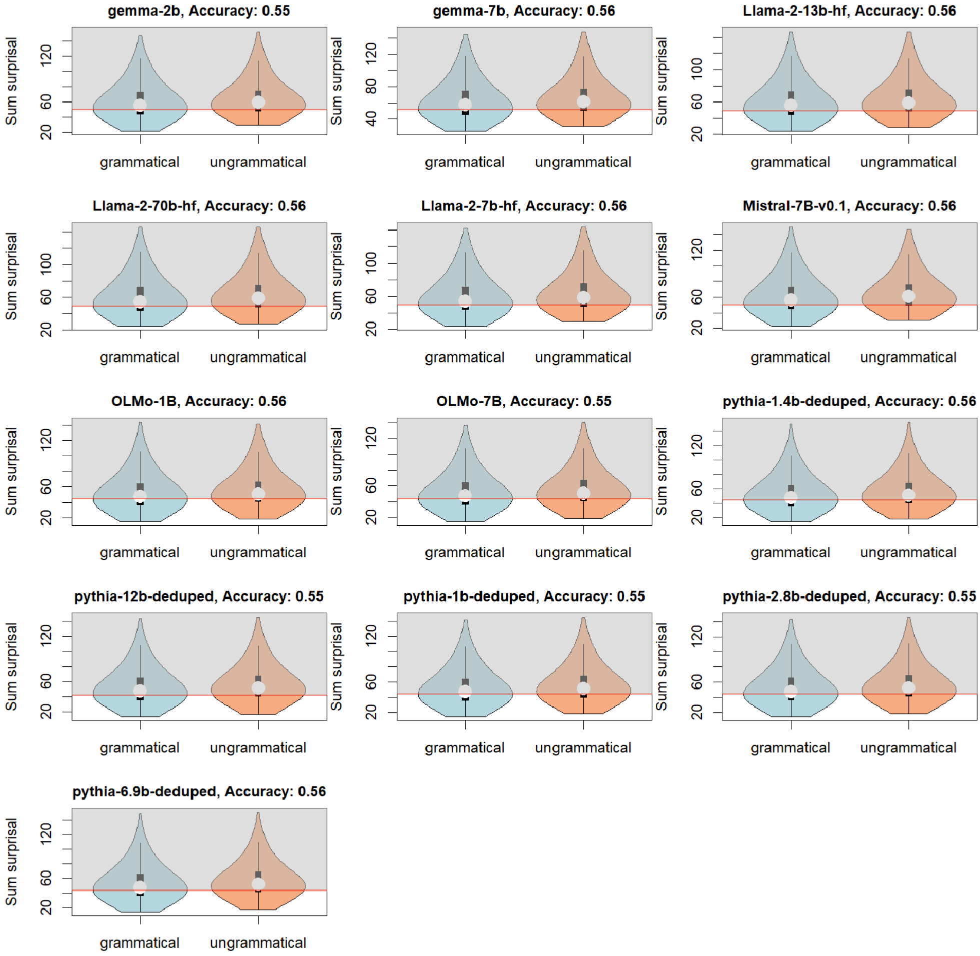
*Distributions of Sum Surprisal Using the DGL Dataset in 13 Models*



*Note.* The surprisal threshold that results in the classification with the highest accuracy is indicated by the horizontal red line. Any sentence with a sum surprisal higher than or equal to that threshold is classified as “ungrammatical”, while any sentence with a sum surprisal lower than that threshold is classified as “grammatical”. The probabilities are taken from Hu and Frank (2024).

**Figure 2**

*Distributions of Sum Surprisal Using the BLiMP Dataset in 13 Models*



*Note.* The probabilities are taken from Hu and Frank (2024).

The problem of double standards also includes tasks that in principle can be applied to both humans and LLMs but their results are assessed differently across the two. To provide an example, Hu et al.'s (in press) strongest argument comes from finding that GPT-3.5 Turbo and GPT-4 (not tested in Dentella et al., 2023) outperform humans in providing judgments of grammatical well-formedness (note again that LLMs outperforming humans would pull into question the alleged "human-like behavior" of LLMs). Before explaining why this finding is spoiled by the problem of double standards, let us briefly add that this difference from the models tested in Dentella et al. does not entail that the more recent models have developed an understanding of grammatical correctness that even surpasses that of humans sometime after the first testing took place. Hu et al. do not acknowledge so, but potential alternative explanations for this better performance exist: for instance, that the DGL dataset was available online and discussed in social media by the time Hu et al. performed their testing. LLMs are trained on thousands of scientific papers (Frank, 2023), and algorithmic fudging based on continuous social media monitoring has been noted, affecting performance in linguistic tasks too (Leivada et al., 2023).

Admittedly, it is easy to understand how this very good performance of the recent models could legitimize Hu et al.'s claim that LLMs show human-like abilities. This looks compelling, until one checks the raw material Hu et al. make available. The LLMs, according to the task instructions, should return one of the two outcomes: C if a prompt is correct, and N if it is not. Yet the raw responses show some peculiar answers that deviate from the instructions, such as "The teacher is going to teach a lesson on the Civil War" or "I will be meeting with John and Karen C". Instead of coding these answers as incorrect (since the target response C was not unambiguously provided), Hu et al. removed the first one (which lacks either C or N) and coded the other as target/accurate. Despite the alleged alignment with human responses, it is very hard to imagine that a human would respond this way, and anyone would count it as target behavior in this specific task.

If we recode all the "hallucinating replies" as non-target/inaccurate, Hu et al.'s dataset from judgment prompting effectively replicates Dentella et al.'s main finding: humans *are* more accurate than davinci2 and davinci3. More recent models such as GPT-3.5 Turbo and GPT-4 indeed do better, but as argued above, more than one reason could be responsible for this better performance. Also, it is interesting to note that even the best performing model (GPT-4) includes a paradoxical reply that asserts that a given prompt is both correct and incorrect. We interpret these results as showing that scaling mitigates but does not completely cover the *qualitative* differences in the performance of humans vs. LLMs. The same is true for the problem of double standards: Holding the models accountable to a different level of performance may cover quantitative differences, but their distinctly non-human errors are still present and harder to explain away.

## 5. Conclusion

We identified three caveats that may give rise to an inaccurate picture of the linguistic abilities of LLMs: the problem of unlicensed generalizations, the human-like paradox, and the problem of double standards. Moreover, we provided a detailed comparison of two different methodologies for tapping into the linguistic abilities of the models: judgment tasks vs. direct probabilities. To establish the comparison, we re-analyzed Hu et al. (in press) and Hu and Frank (2024), and we found that probabilities do not unambiguously tell apart grammatical from ungrammatical sentences. This raises concerns about whether this truly is a superior method for approaching the internalized grammatical abilities of LLMs. In relation to the validity of this method, it is also worth observing that Hu et al. (in press) did not obtain direct probability measurements—which is methodologically superior to judgments according to them (see also Hu & Levy, 2023, for the same claim)—from the more recent models GPT-3.5 Turbo and GPT-4; they only elicited judgments from them. They do not explain why they chose to use the "inferior" method with the recent models, but it is because OpenAI no longer returns log probabilities in most of their models. This option has been disabled probably because it reveals too much information about the black-box training data of proprietary, closed-source models.

Ronald Coase once said that if data is tortured long enough, it will confess to anything. For models that have been linked to a tendency to amplify misinformation (Kidd & Birhane, 2023), training data is the weakest link: If tortured long enough, they may give rise to uncomfortable confessions that are hard to reconcile with the current AI hype that (at times, uncritically) endows LLMs with human-like abilities.

**Funding:** The authors have no funding to report.

**Acknowledgments:** We thank Jennifer Hu for her help in interpreting her datasets.

**Competing Interests:** The authors have declared that no competing interests exist.

## References

Dentella, V., Günther, F., & Leivada, E. (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 of the United States of America, 120*(51), Article e2309583120. https://doi.org/10.1073/pnas.2309583120

Felin, T., & Holweg, M. (2024). Theory is all you need: AI, human cognition, and decision making. https://doi.org/10.2139/ssrn.4737265

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology, 2*, 451–452. https://doi.org/10.1038/s44159-023-00211-x

Gates, B. (2023, March 21). *The age of AI has begun.* GatesNotes. https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and Artificial Neural Networks. *Computational Brain & Behavior, 6*, 213–227. https://doi.org/10.1007/s42113-022-00166-x

Hu, J., & Frank, M. C. (2024). *Auxiliary task demands mask the capabilities of smaller language models.* arXiv. https://doi.org/10.48550/arXiv.2404.02418

Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in Large Language Models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5040–5060). Association for Computational Linguistics.

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (in press). Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences of the United States of America.*

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11*(2), 123–141. https://doi.org/10.1016/0010-0277(82)90022-1

Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science, 380*(6651), 1222–1223. https://doi.org/10.1126/science.adi0248

Leivada, E., Dentella, V., & Murphy, E. (2023). The quo vadis of the relationship between language and Large Language Models. To appear in J.-L. Mendívil-Giró (Ed.), *Artificial knowledge of language: A linguist's perspective on its nature, origins and use.* https://doi.org/10.48550/arXiv.2310.11146

Leivada, E., Günther, F., & Dentella, V. (in press). Reply to Hu et al: Applying different evaluation standards to humans vs. Large Language Models overestimates AI performance. *Proceedings of the National Academy of Sciences of the United States of America.*

Leivada, E., & Westergaard, M. (2020). Acceptable ungrammatical sentences, unacceptable grammatical sentences, and the role of the cognitive parser. *Frontiers in Psychology, 11*, Article 364. https://doi.org/10.3389/fpsyg.2020.00364

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics, 8*, 377–392. https://doi.org/10.1162/tacl_a_00321

PsychOpen GOLD is a publishing service by
Leibniz Institute for Psychology (ZPID), Germany.
www.leibniz-psychology.org

PsychOpen GOLD