PNAS LETTER

Language models align with human judgments on key grammatical constructions

Jennifer Hu[a,1], Kyle Mahowald[b], Gary Lupyan[c], Anna Ivanova[d], and Roger Levy[e]

Affiliations are included on p. 3.

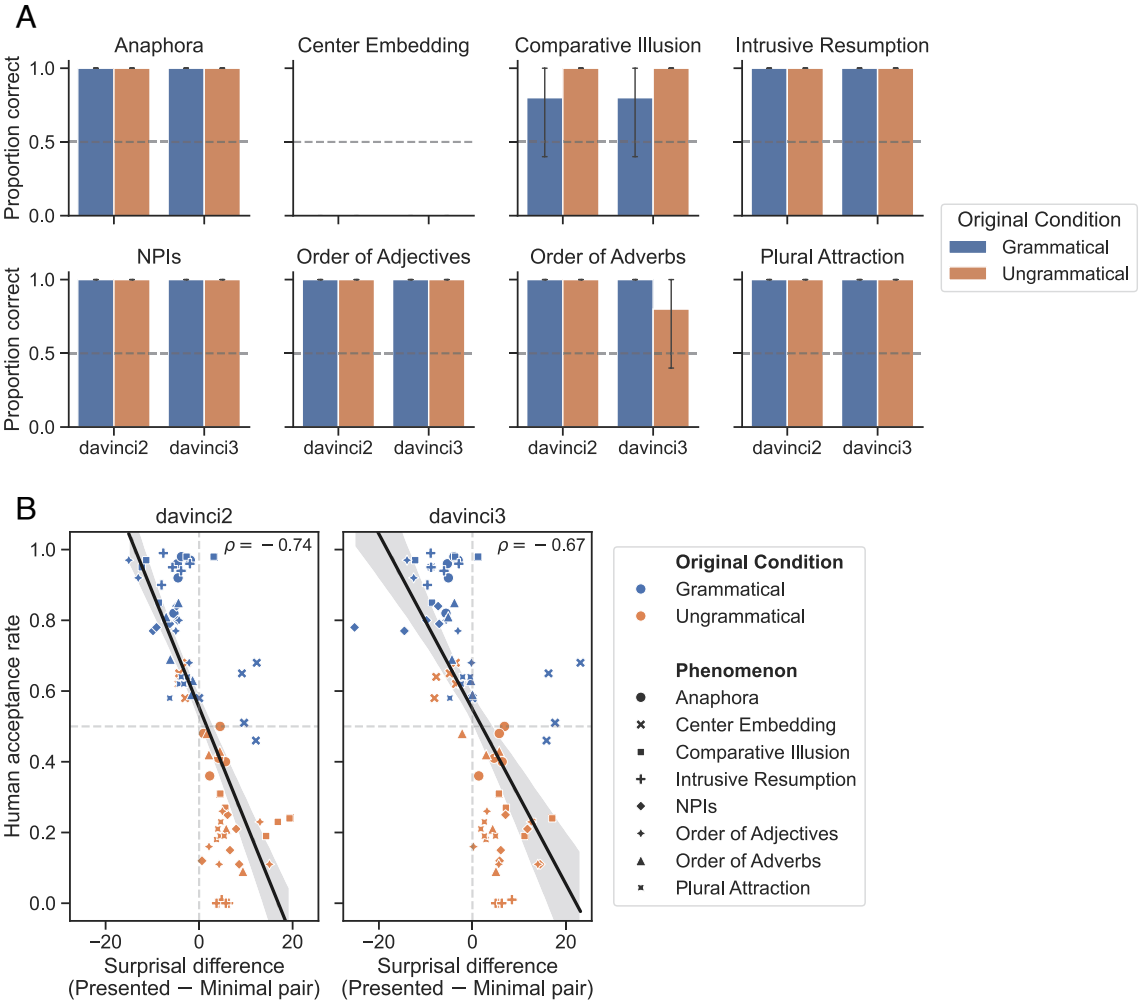

Fig. 1. (*A*) Accuracy scores achieved by models on a version of DGL's original materials with minimal pairs. For each phenomenon, accuracy is computed as the proportion of items in that phenomenon where the model assigns higher probability to the grammatical version of that item (minimal pair) than the ungrammatical version. (*B*) x-axis: Difference in sum surprisal (negative log probability) between the sentence presented to humans in DGL's experiments versus its counterpart in the minimal pair. y-axis: Human acceptance rate (proportion judged as grammatical) for the presented sentence in each minimal pair. Each point represents a minimal pair test item.

Do large language models (LLMs) make human-like linguistic generalizations? Dentella, Günther, and Leivada (1) (DGL) prompt several LLMs ("Is the following sentence grammatically correct in English?") to elicit grammaticality judgments of 80 English sentences, concluding that LLMs demonstrate a "yes-response bias" and a "failure to distinguish grammatical from ungrammatical sentences." We reevaluate LLM performance using well-established practices and find that DGL's data in fact provide evidence for how well LLMs capture human linguistic judgments.

The ability to produce well-formed sentences does not necessarily require being able to articulate the underlying grammatical rules. This distinction has been long noted in linguistics (e.g., refs. 2–4), but is blurred by DGL: Their task tests not only LLMs' grammatical competence but also whether models know what "grammatically correct" means.

Author contributions: J.H., K.M., G.L., A.I., and R.L. designed research; J.H. and G.L. performed research; J.H., K.M., G.L., A.I., and R.L. analyzed data; and J.H., K.M., G.L., and R.L. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: jenniferhu@fas.harvard.edu.

Published August 26, 2024.

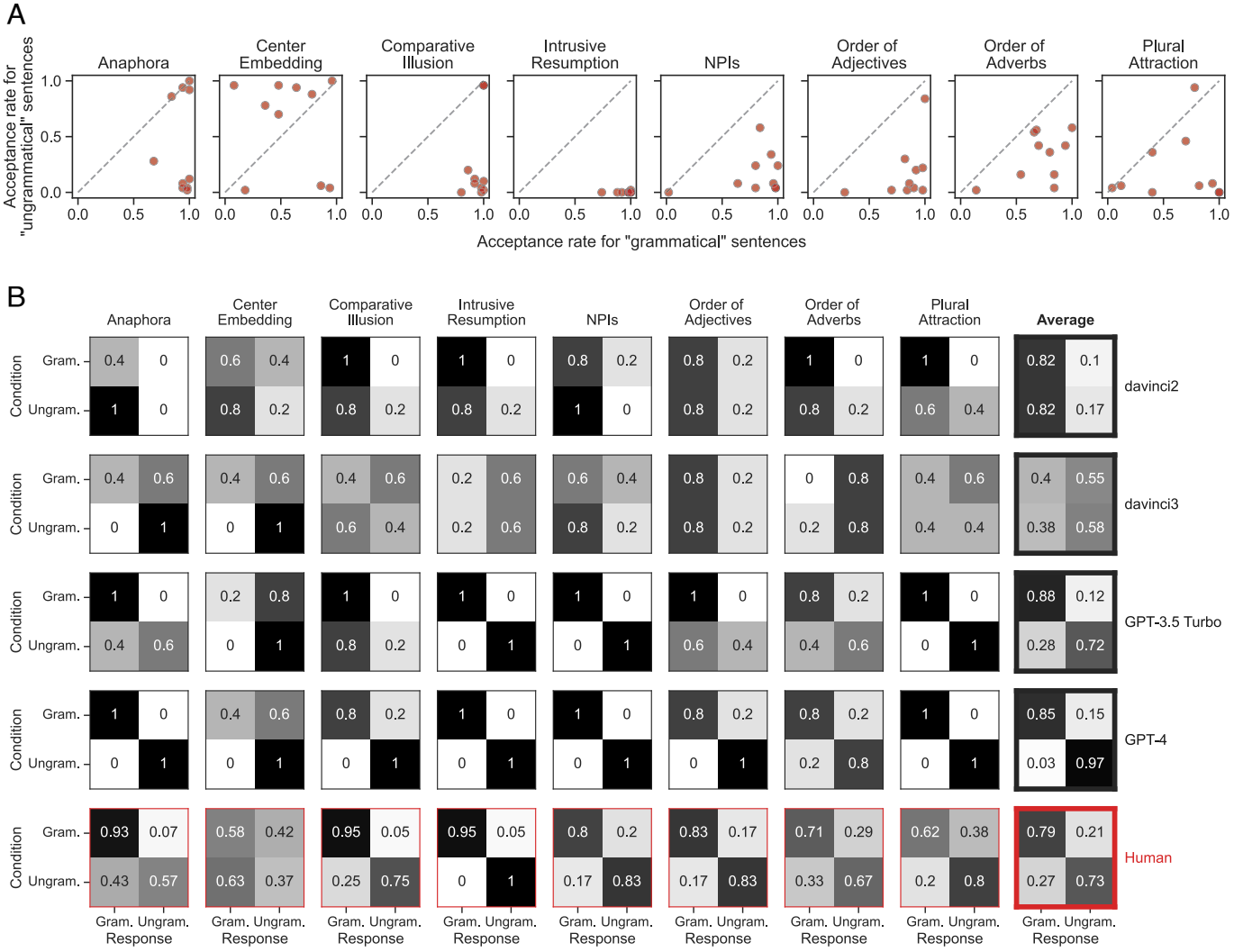

Fig. 2. (*A*) Participant-specific acceptance rates (i.e., rate of judging as grammatical) for sentences that DGL label as "grammatical" (x-axis) versus "ungrammatical" (y-axis). If participants' responses perfectly reflected DGL's normative coding, then all participants would be in the *Bottom Right* corner (as exemplified by intrusive resumption). (*B*) Confusion matrices achieved by models and humans on each phenomenon, when evaluating models using the same prompt that was seen by humans ("Is the following sentence grammatically correct in English? [SENTENCE] Respond with C if it is correct, and N if it is not correct."). "Gram." = grammatical, and "Ungram." = ungrammatical. A small fraction of davinci2 and davinci3's responses (4%) were not codable as corresponding to "C" or "N," resulting in missing data.

To remedy this, we follow standard methods (not discussed by DGL) of evaluating LLMs' linguistic knowledge (e.g., refs. 5–7). Rather than relying on models' metalinguistic skills, a method that systematically underestimates LLM generalization capabilities (8), we directly measure the probabilities models assign to strings (9). For each sentence in DGL's materials, we constructed a lexically matched counterpart differing only in the targeted grammatical feature. This controlled manipulation isolates grammatical differences, so a model that has learned the correct generalizations should assign higher probability to the grammatical sentence in each minimal pair. Minimal-pair analysis reveals at- or near-ceiling performance except on center embedding (Fig. 1), for which humans are also below chance (47.1% accuracy). Furthermore, minimal-pair surprisal (negative log-probability) differences predict item-level variation in human responses: The less surprising a sentence relative to its minimal pair, the more likely humans are to judge it as grammatical (Fig. 1; davinci2: Pearson $\rho = -0.74$; davinci3: Pearson $\rho = -0.67$).

Moreover, although DGL argue that human-judgment inaccuracies reflect "performance factors," their data reveal systematic variation in human acceptability judgments (Fig. 2*A*). For instance, the Anaphora phenomenon shows two groups of participants: one whose judgments conform to DGL's labels (*Bottom Right* cluster) and one judging all sentences as grammatical (*Upper Right* cluster). DGL's logic would imply that only these latter participants suffer performance constraints. Genuine variability in acceptability is a better explanation and is consistent with a wide literature in linguistics (10). For example, the anaphora sentences that DGL label as ungrammatical use the word "themselves" as a singular pronoun, which may be perfectly acceptable to some speakers. Similarly, many participants (43%) judge

order-of-adverbs sentences like "Gary still perhaps drives to work" as grammatical, even though DGL code it as ungrammatical.

Finally, DGL's task differed subtly for models and humans: Models were prompted for open-ended responses (which were subsequently coded as correct/incorrect by DGL), whereas humans had to provide a binary judgment by pressing one of two keys. We reevaluated davinci2, davinci3, GPT-3.5 Turbo, and GPT-4 using the exact prompt seen by humans (Fig. 2*B*). The "yes"-bias reported by DGL disappears for all models except davinci2. While davinci2 and davinci3 still perform near chance, GPT-3.5 Turbo and GPT-4 outperform humans according to DGL's normative grammaticality coding. Overall, we conclude that LLMs show strong and human-like grammatical generalization capabilities.

ACKNOWLEDGMENTS. We thank Evelina Fedorenko, Ted Gibson, and Steve Piantadosi for helpful comments and discussion.

Author affiliations: [a]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA 02138; [b]Department of Linguistics, The University of Texas at Austin, Austin, TX 78712; [c]Department of Psychology, University of Wisconsin-Madison, Madison, WI 53706; [d]School of Psychology, Georgia Tech, Atlanta, GA 30332; and [e]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

1. V. Dentella, F. Günther, E. Leivada, Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2309583120 (2023).
2. D. Birdsong, *Metalinguistic Performance and Interlinguistic Competence, Springer Series in Language and Communication* (Springer, Berlin Heidelberg, 1989).
3. Y. Han, "Grammaticality judgment tests: How reliable and valid are they?" in *Applied Language Learning*, L. Woytak, Ed. (Defense Language Institute Foreign Language Center, 2000), vol. 11, pp. 177–204.
4. N. Chomsky, *Knowledge of Language: Its Nature, Origin, and Use* (Praeger Scientific, 1986).
5. R. Marvin, T. Linzen, "Targeted syntactic evaluation of language models" in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii, Eds. (Association for Computational Linguistics, Brussels, Belgium, 2018), pp. 1192–1202.
6. A. Warstadt *et al*., BLiMP: The benchmark of linguistic minimal pairs for english. *Trans. Assoc. Comput. Linguist.* **8**, 377–392 (2020).
7. J. Hu, J. Gauthier, P. Qian, E. Wilcox, R. Levy, "A systematic assessment of syntactic generalization in neural language models" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Eds. (Association for Computational Linguistics, Online, 2020), pp. 1725–1744.
8. J. Hu, R. Levy, "Prompting is not a substitute for probability measurements in large language models" in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, K. Bali, Eds. (Association for Computational Linguistics, Singapore, 2023), pp. 5040–5060.
9. A. Clark, S. Lappin, *Linguistic Nativism and the Poverty of the Stimulus* (John Wiley & Sons, 2010).
10. J. Bresnan, A. Cueni, T. Nikitina, R. H. Baayen, "Predicting the dative alternation" in *Cognitive Foundations of Interpretation*, G. Bouma, I. Kraemer, J. Zwarts, Eds. (KNAW, 2007), pp. 69–94.