nature human behaviour

Supplementary information

https://doi.org/10.1038/s41562-025-02297-0

Active use of latent tree-structured sentence representation in humans and large language models

In the format provided by the authors and unedited

Supplementary Figures

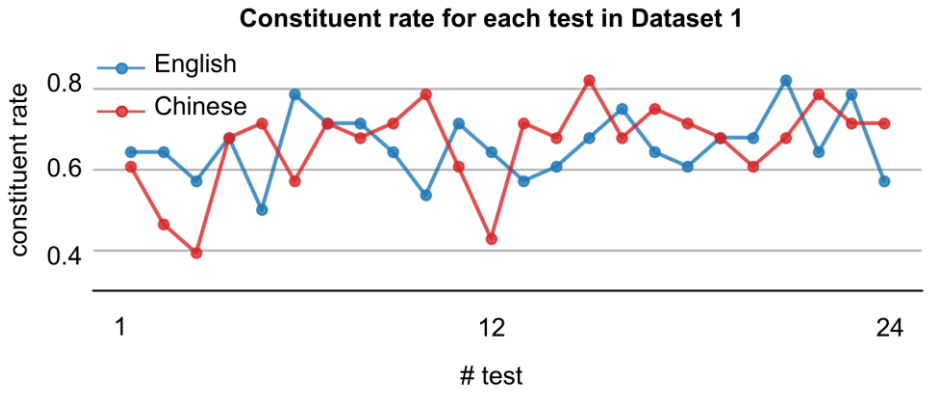

Figure S1. Constituent rate of human participants in each of the 24 tests in Dataset 1. Test 1 is the first test in the experiment and test 24 is the last one. The constituent rate does not significantly change across tests, suggesting no learning effect.

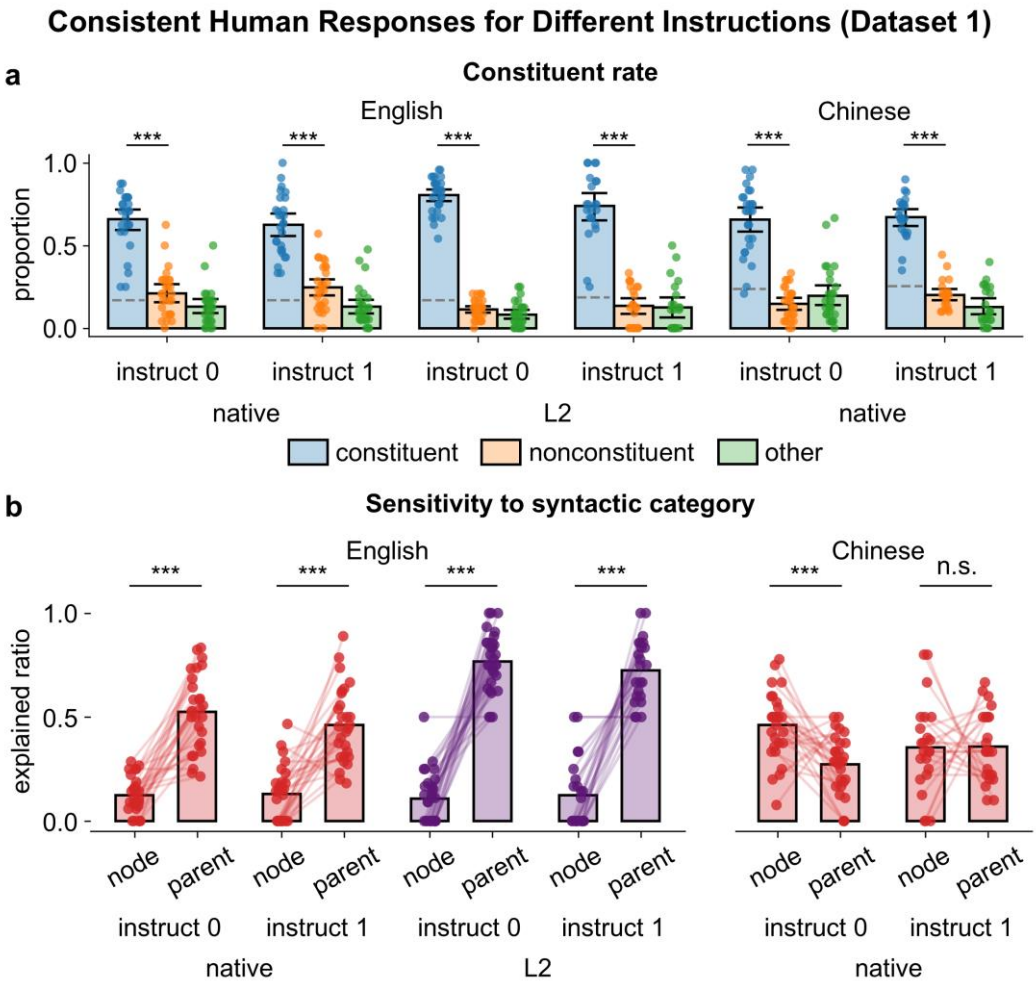

Figure S2. Human results on Dataset 1 under different instructions. The results using instruction 0 (the same as prompt 0 for ChatGPT) are reported in the main text. **a.** Properties of the deleted word string. **b.** The explained ratio for the node- and parent-category rules. Each dot on the bar graph represents data from a human participant (N = 30 for each instruct). Error bars represent 95% confidence intervals (CIs) of the mean across participants, estimated using bootstrap. Human behavior is not strongly affected by the prompt either. ***p < 0.001, paired two-sided bootstrap, FDR corrected. See full statistics in Supplementary Data 1.

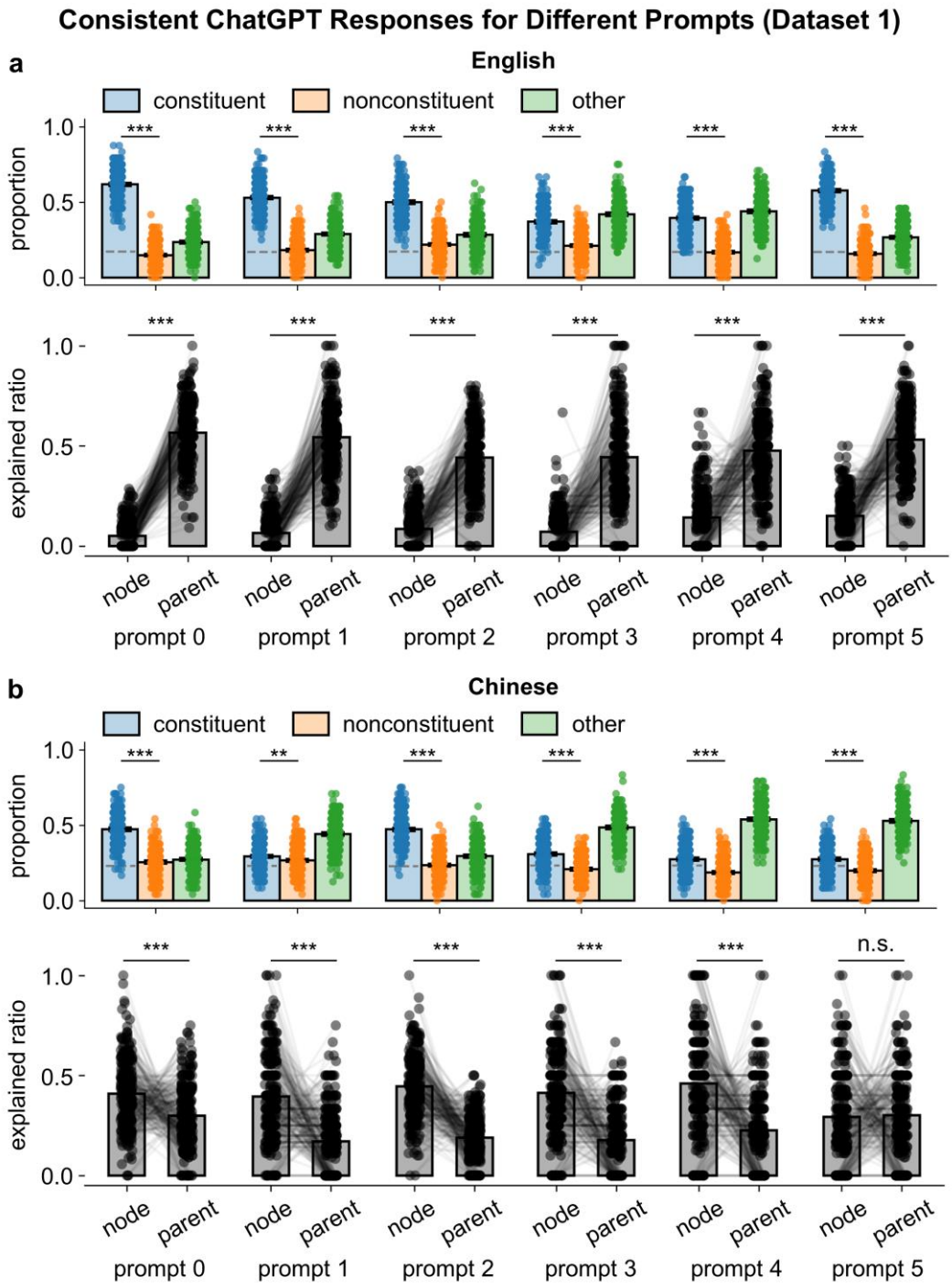

Figure S3. ChatGPT results on Dataset 1 under different prompts. Results of prompt 0 are reported in the main text. **ab.** Properties of the deleted word string and the explained ratio for Dataset 1. Each dot on the bar graph represents data from a single run of ChatGPT (N = 300 for each prompt). Error bars represent 95% CIs of the mean across runs, estimated using bootstrap. The prompt does not strongly affect the results. **p < 0.01, ***p < 0.001, paired two-sided bootstrap, FDR corrected.

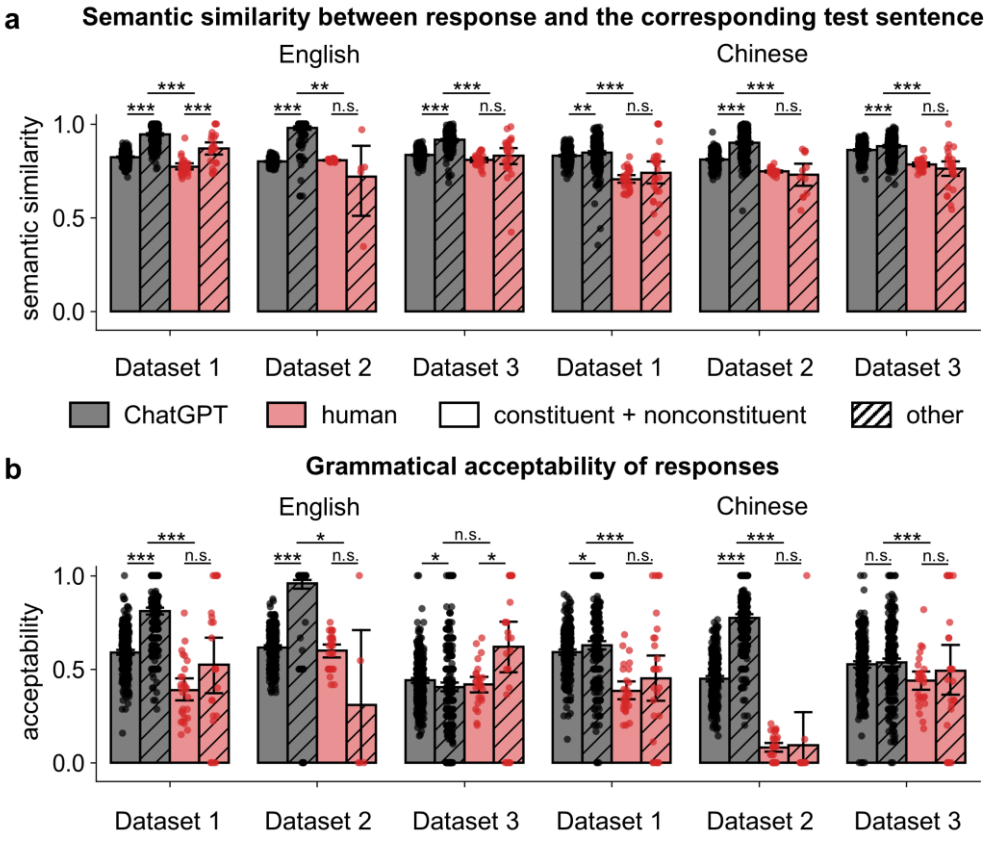

Figure S4. Semantic similarity and grammatical acceptability in Datasets 1-3. **a**. Semantic similarity between the response and the corresponding test sentence. **b**. Grammatical acceptability of responses. ChatGPT tends to preserve the semantics of the test sentence and generate more grammatically acceptable responses than human. Each dot on the bar graph represents data from a human participant or a single run of ChatGPT (N = 30 for humans and N = 300 for ChatGPT. Some data points are discarded because the responses from human/ChatGPT do not fall into the corresponding category.). Error bars represent 95% CIs of the mean across participants (for humans) or runs (for ChatGPT), estimated using bootstrap. *p < 0.05. **p < 0.01. ***p < 0.001, unpaired two-sided bootstrap, FDR corrected.

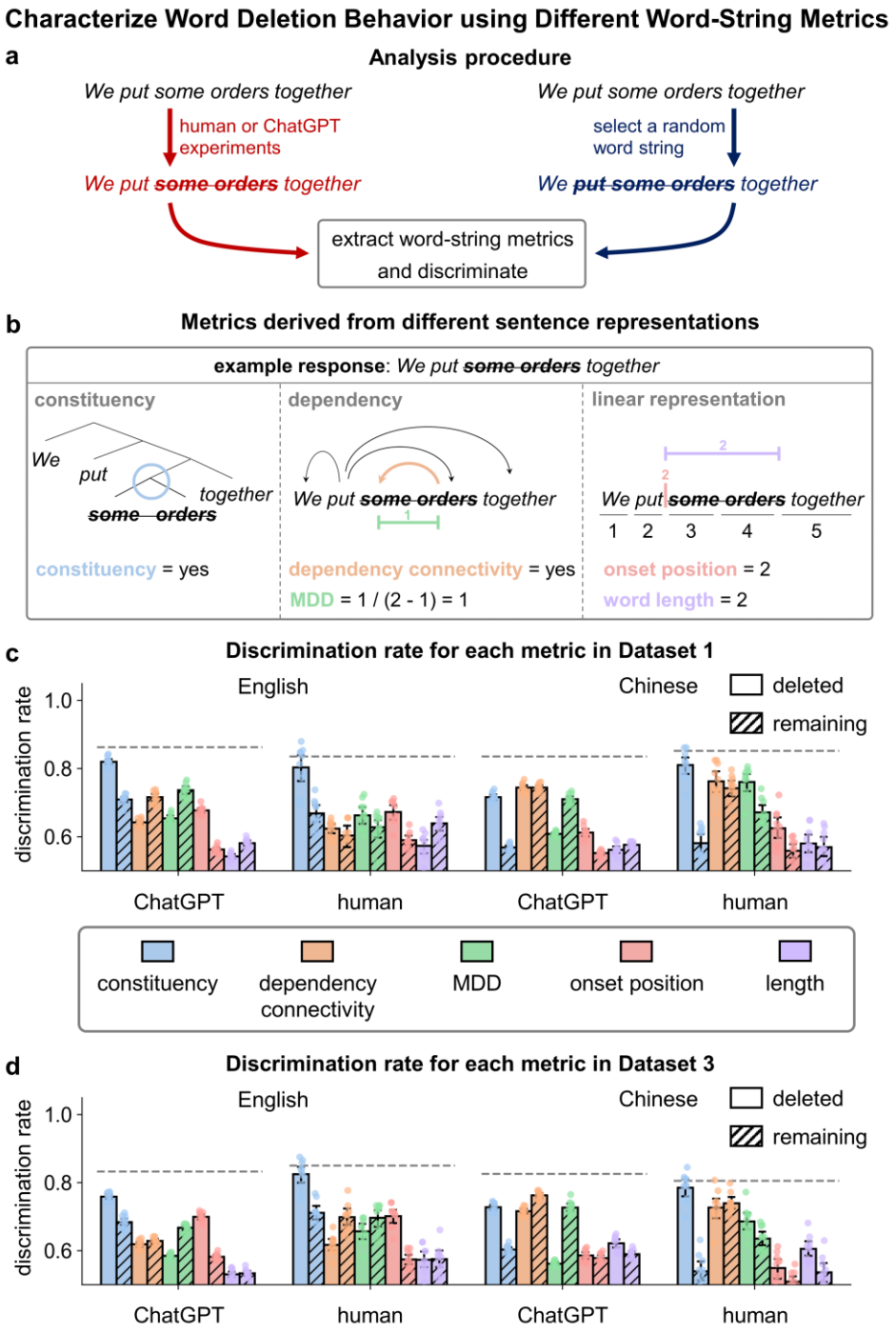

Figure S5. Characterize word deletion behavior using different word-string metrics. **a.** Pipeline of the discrimination analysis. **b.** Illustration of different word-string metrics. **cd.** Discrimination of real human/ChatGPT responses and random word strings based on different word string metrics. Three metrics are derived from a tree-based representation, i.e., constituency, dependency connectivity, and mean dependency distance (MDD), and two metrics are derived from a linear representation. Each bar shows the discrimination based on one metric. Each dot on the bar graph represents data from a repeated run (N = 10). Error bars represent 95% CIs of the mean across runs, estimated

using bootstrap. The gray dashed line represents the discrimination rate when all metrics are jointly considered. The classifier is an SVM.

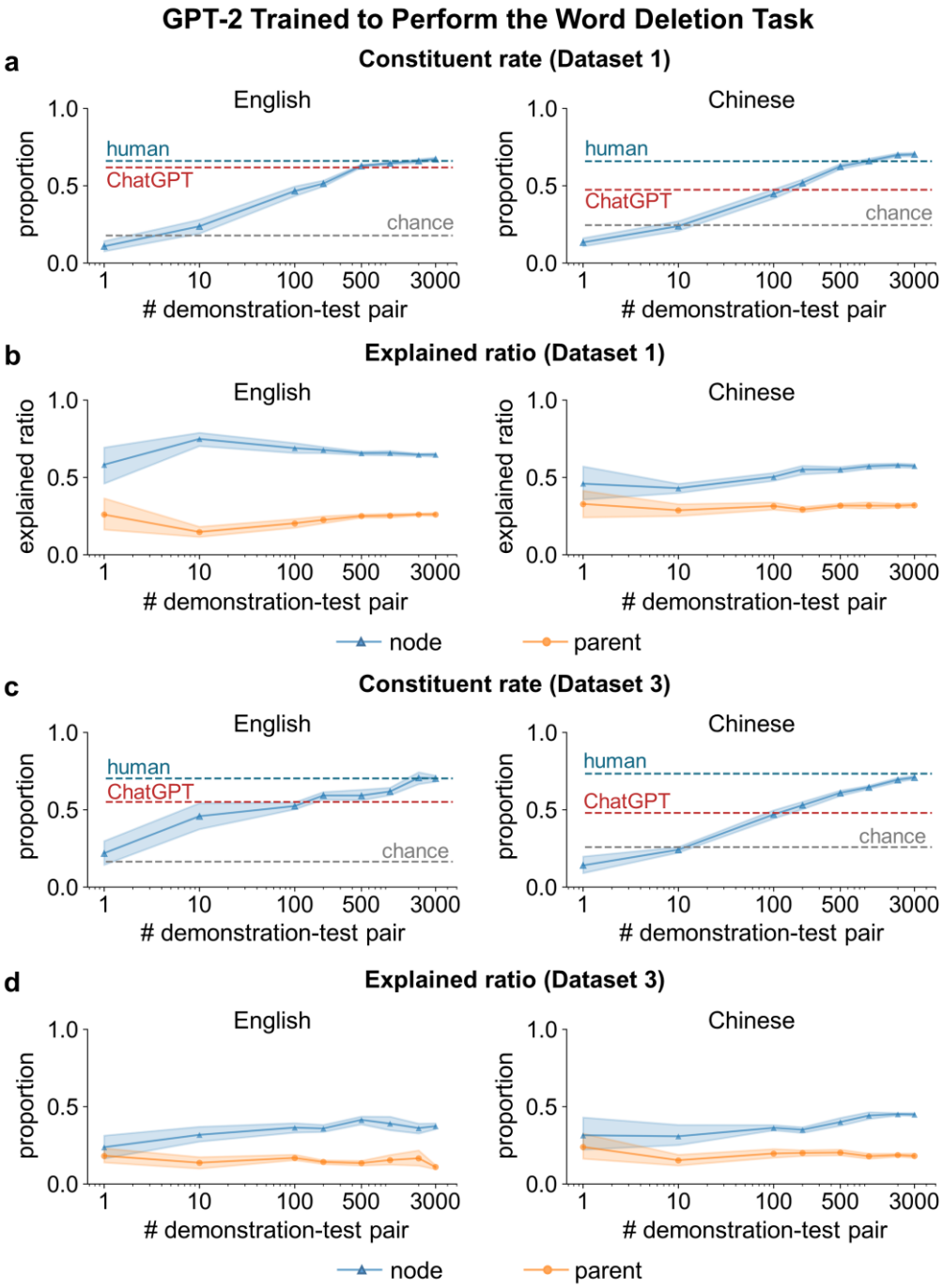

Figure S6. Performance of a baseline model, i.e., a naive GPT-2 model, trained to perform the word deletion task. The naive GPT-2 model is not pretrained, and its input includes word position embeddings and pretrained word embeddings. **ac.** Constituent rate of the GPT-2 model as a function of the number of demonstration-test pairs used for training, for Datasets 1 and 3. The model was trained 10 times based on different samples and the shaded areas cover 95% CIs across runs, estimated using bootstrap. Mean performance of humans and ChatGPT is shown by the dashed lines. **bd.** The explained ratio for the node- and parent-category rules for the GPT-2 model.

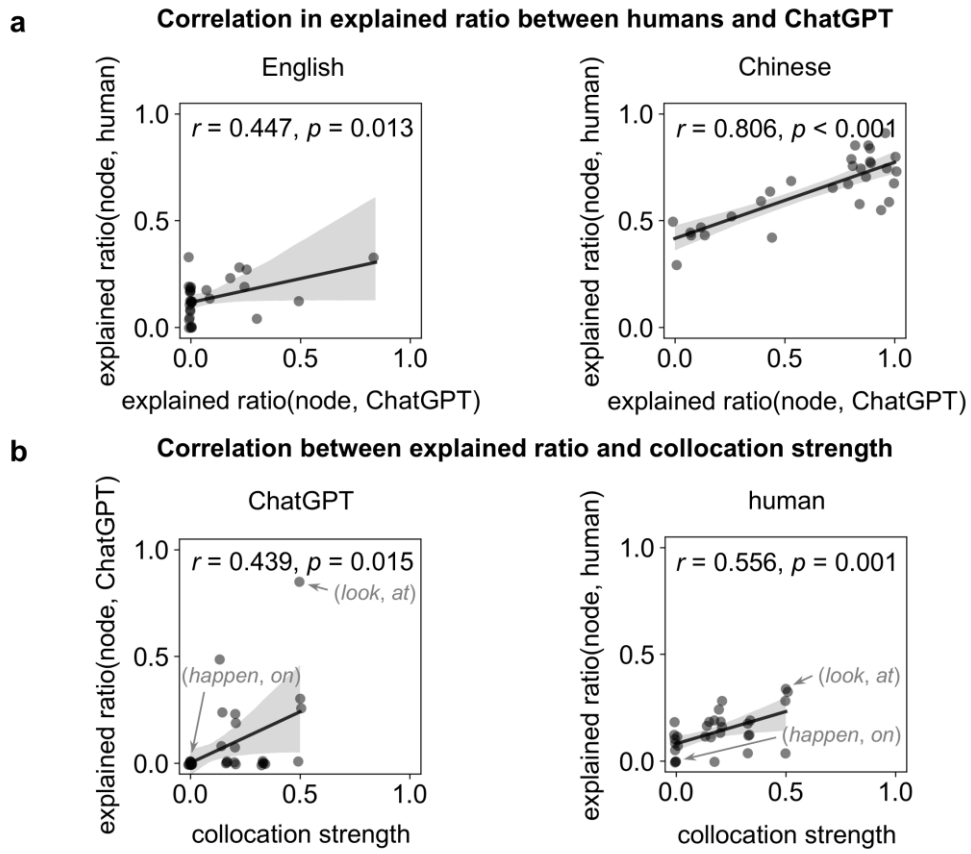

Figure S7. Analysis of lexical effects on word deletion task in Dataset 2. **a**. The explained ratio for the node-category rule is correlated between humans and ChatGPT. Each dot represents a sentence (N = 30). The shaded areas cover 95% CIs across sentences, estimated using bootstrap. **b**. The explained ratio for the node-category rule is correlated with the strength of collocation between verb and preposition. Collocation strength is extracted based on the Oxford Collocations Dictionary[1] using the following method – If the verb-preposition combination, e.g., “look at”, was one of the K verb-preposition collocations for the verb in the dictionary, the collocation strength was coded as 1/K. If a combination was not listed as a collocation, the collocation strength was coded as 0.

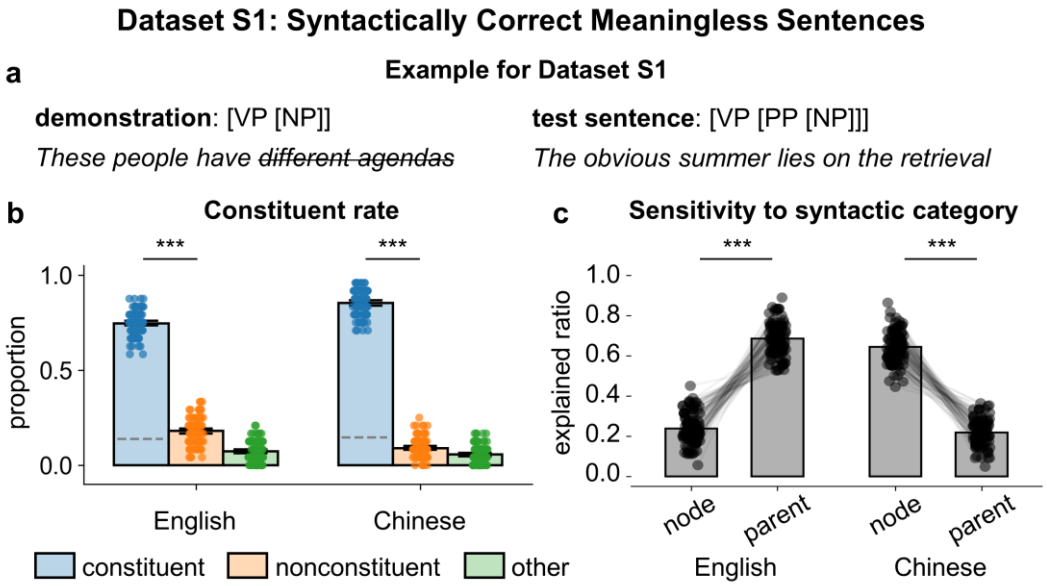

Figure S8. Results of Dataset S1. The test sentences were syntactically correct but meaningless sentences that were constructed based on the test sentences of Dataset 2. The demonstration sentence was the same as Dataset 2. **a**. Example sentences. **b**. Constituent rate of word strings deleted by ChatGPT. **c.** Explained ratio for the node- and parent-category rules. Each dot on the bar graph represents data from a single run of ChatGPT (N = 100). Error bars represent 95% CIs of the mean across runs, estimated using bootstrap. Results are consistent with the results of Dataset 2. ***p < 0.001, paired two-sided bootstrap, FDR corrected.

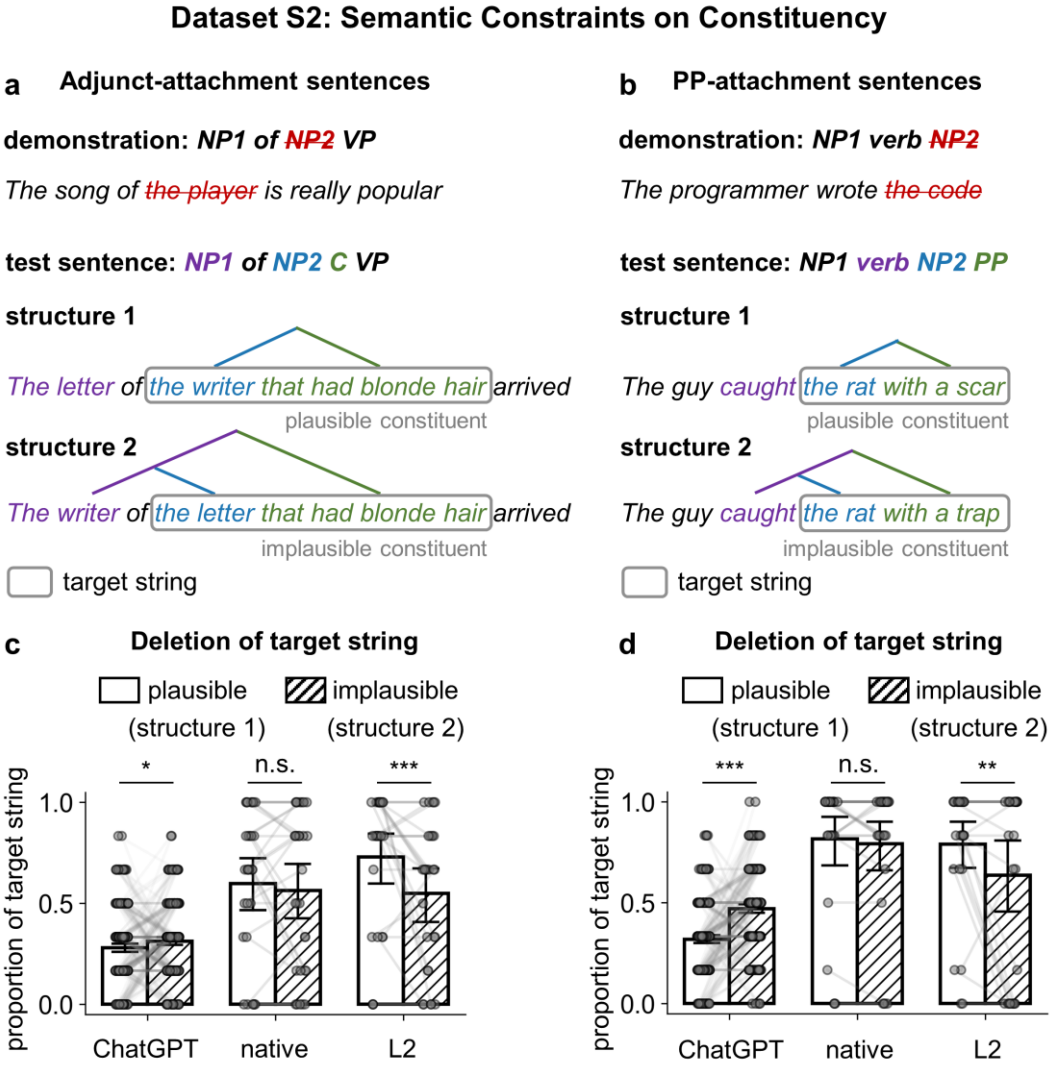

Figure S9. Influence of semantic plausibility on the word deletion task. The test sentence is always a syntactically ambiguous sentence, and only one of the syntactic structures corresponds to a semantically plausible sentence. **a.** In the first type of syntactically ambiguous sentence, the clause *C* (green) could either modify the *NP1* (purple) or *NP2* (blue), forming two potential syntactic structures. The words are selected so that either structure 1 or structure 2 is semantically plausible. The demonstration induces the participants to delete the target string (boxed by the gray line), which is only a semantically plausible constituent in structure 1. **b.** In the second type of syntactically ambiguous sentence, the *PP* (green) could either modify *NP2* (blue) or the verb (purple). **cd.** The proportion of tests in which the target string is deleted for the two types of syntactically ambiguous sentences. Each data point represents the result from a participant or a run of ChatGPT (N = 30 for humans and N = 300 for ChatGPT). Error bars represent 95% CIs of the mean across participants (for humans) or runs (for ChatGPT), estimated using bootstrap. The target string is

more frequently deleted when structure 1 is semantically plausible for L2 speakers, but not for native speakers and ChatGPT. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, paired two-sided bootstrap, FDR corrected.

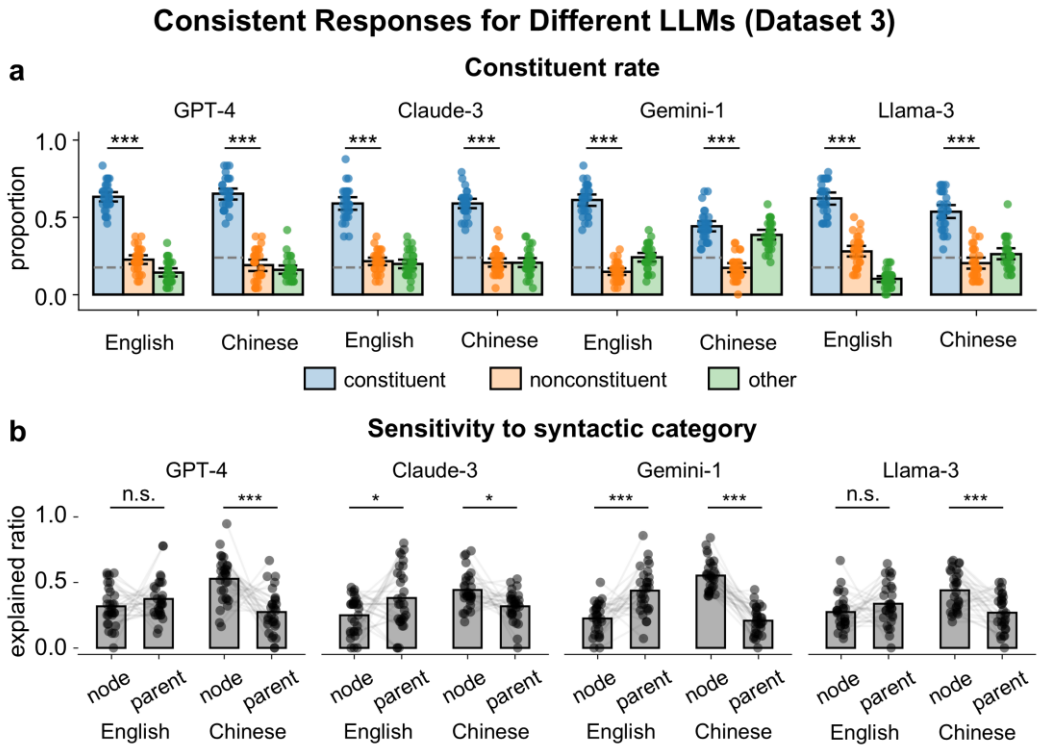

Figure S10. Results on Dataset 3 for different LLMs, including GPT-4, Claude-3, Gemini-1, and Llama-3. Each model receives 720 independent tests, which are divided into 30 runs and each run has the same number of tests (N = 24) as in the human experiment. **a.** Properties of the deleted word string. **b.** The explained ratio for the node- and parent-category rules. Each dot on the bar graph represents data from a single run of LLM (N = 30). Error bars represent 95% CIs of the mean across runs, estimated using bootstrap. The results are generally consistent with the results of ChatGPT. *p < 0.05, **p < 0.01, ***p < 0.001, paired two-sided bootstrap, FDR corrected.

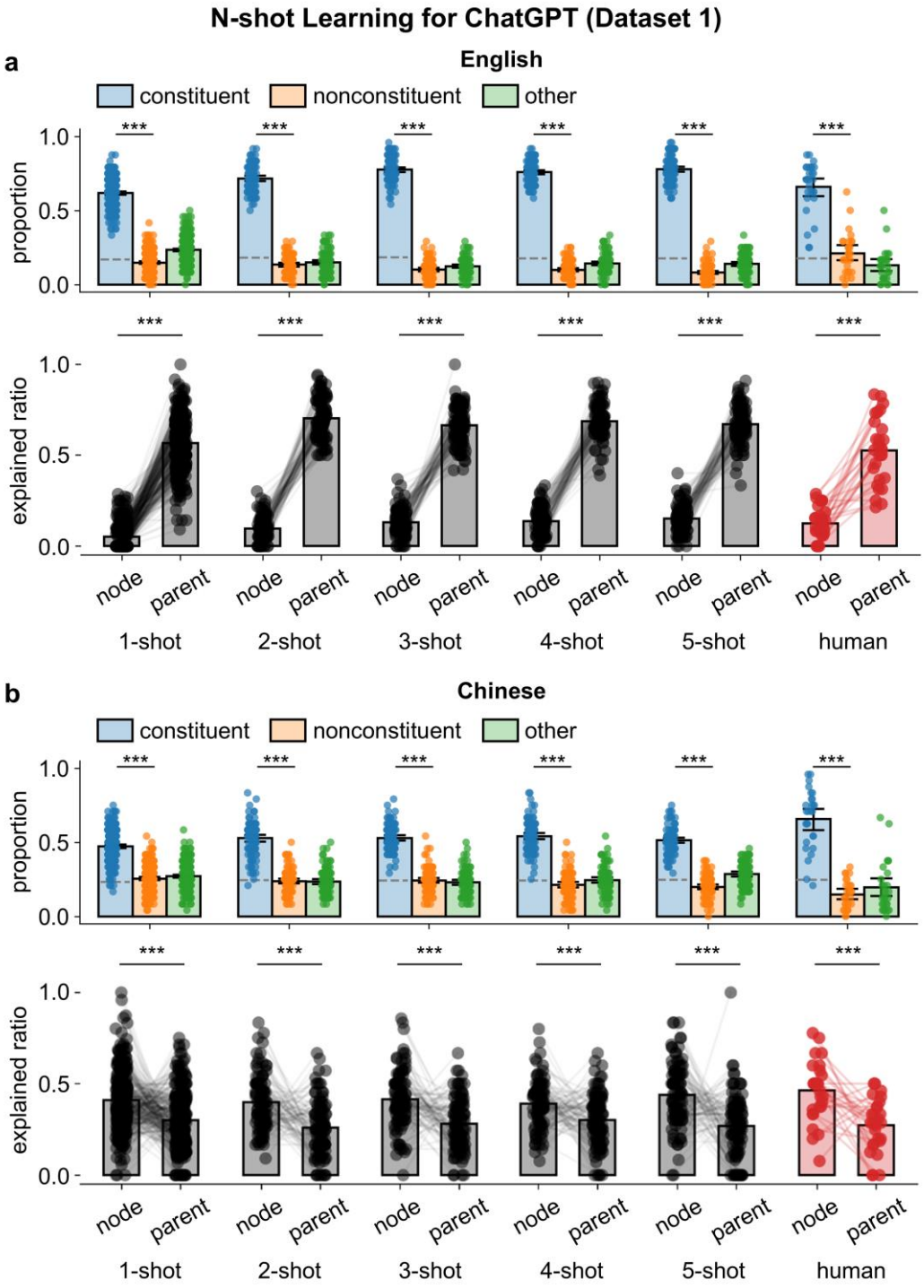

Figure S11. ChatGPT N-shot learning for Dataset 3. **ab.** Constitute rate slightly improves with more demonstrations. The explained ratio for the node- and parent-category rules do not clearly change with more demonstrations. Each dot on the bar graph represents data from a single run of LLM (N = 300 for 1-shot, N = 100 for 2- to 4-shot, and N = 30 for humans). Error bars represent 95% CIs of the mean across participants (for humans) or runs (for ChatGPT), estimated using bootstrap. ***p < 0.001, paired two-sided bootstrap, FDR corrected.

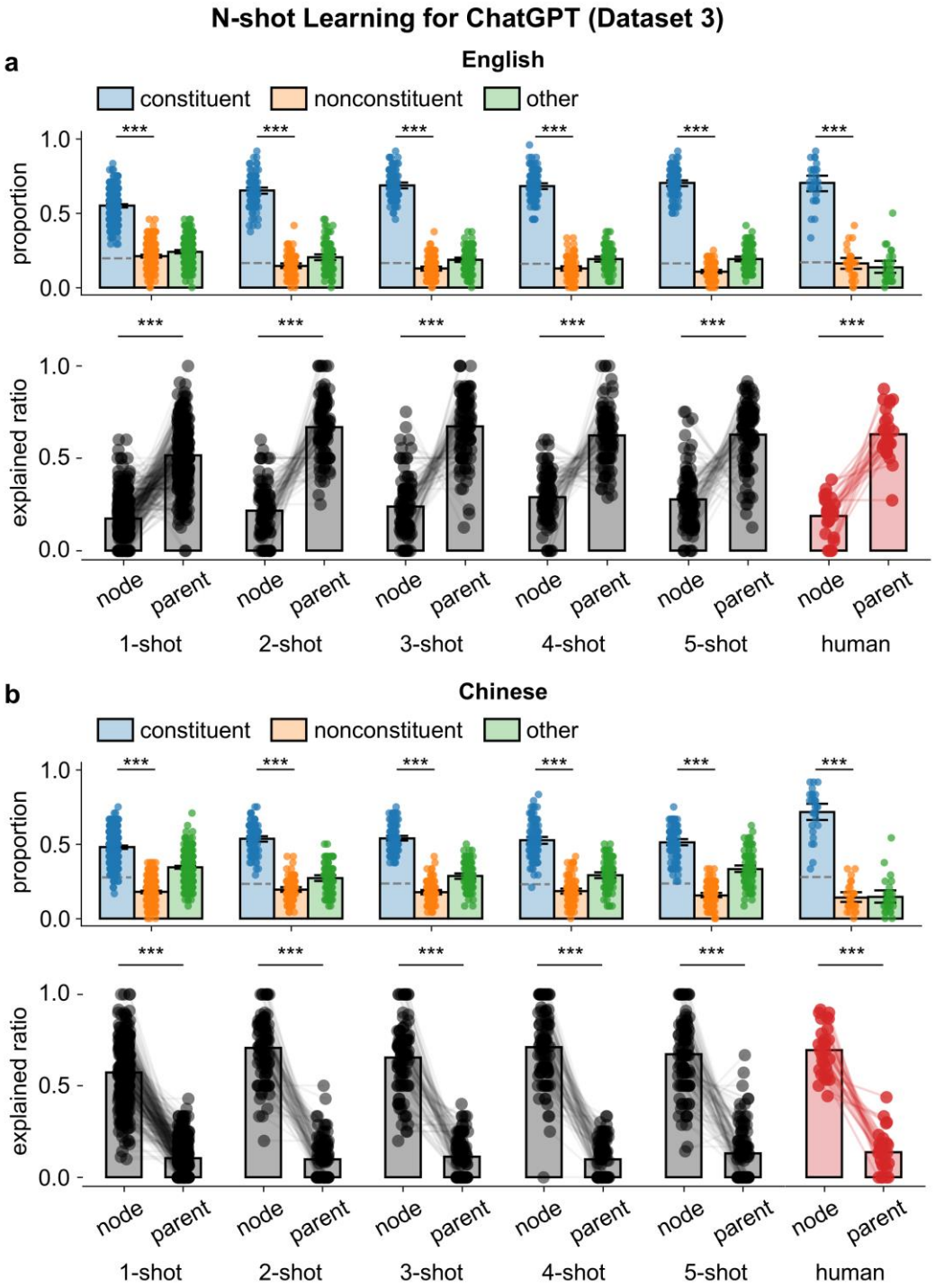

Figure S12. ChatGPT N-shot learning for Dataset 3. **ab.** Constitute rate slightly improves with more demonstrations. The explained ratio for the node- and parent-category rules do not clearly change with more demonstrations. Each dot on the bar graph represents data from a single run of LLM (N = 300 for 1-shot, N = 100 for 2- to 4-shot, and N = 30 for humans). Error bars represent 95% CIs of the mean across participants (for humans) or runs (for ChatGPT), estimated using bootstrap. ***p < 0.001, paired two-sided bootstrap, FDR corrected.

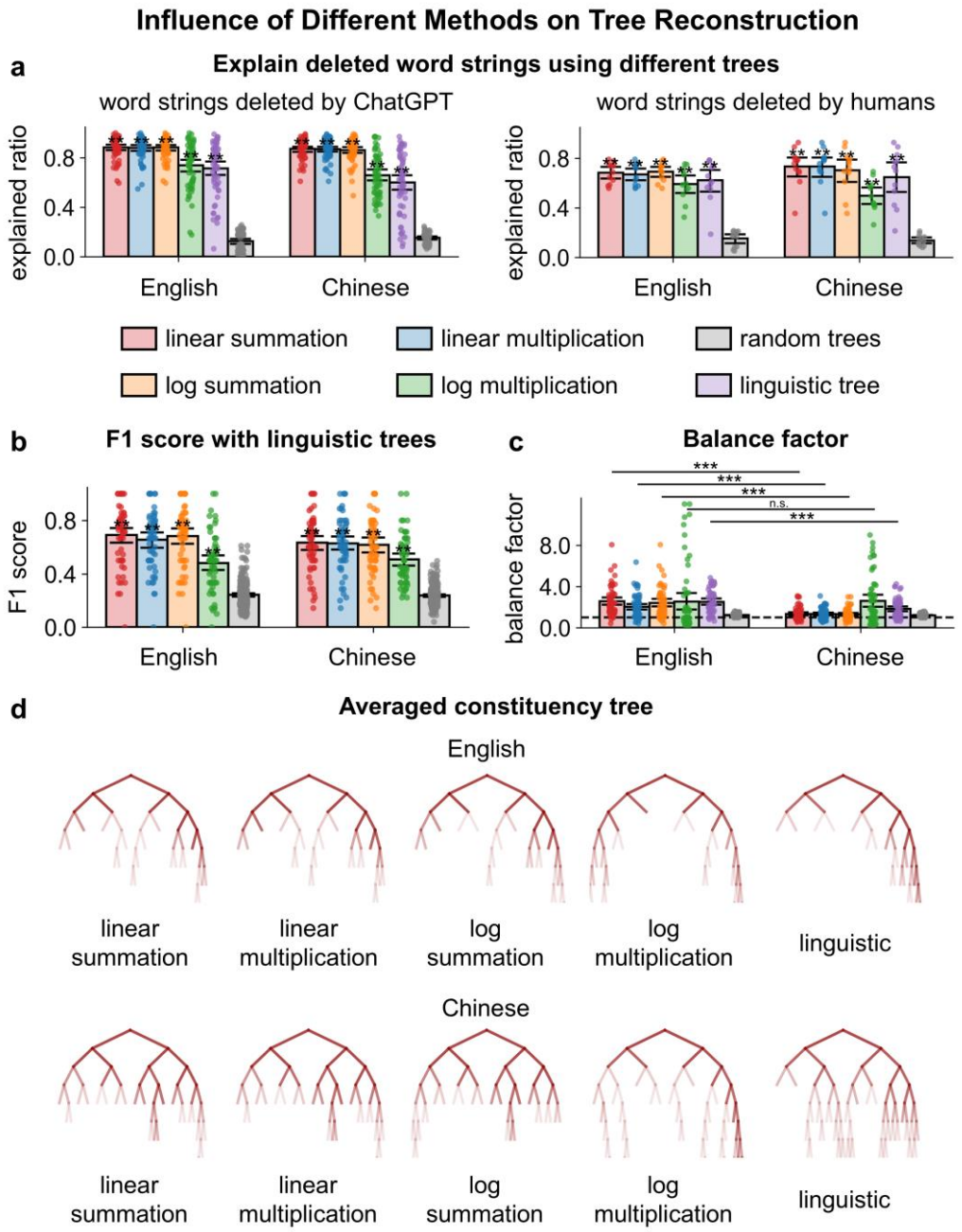

Figure S13. Constituency tree reconstructed using 4 different methods differing in how the explained ratio is integrated, i.e., linear summation (reported in the main text), linear multiplication, log summation and log multiplication. The results are generally consistent except for log multiplication. **a**. Explained ratios for different trees. **b**. F1 score with respect to the linguistic tree. **c**. Balance factor of each constituency tree. Each dot represents a constituency tree (N = 60). Error bars represent 95% CIs of the mean across constituency trees, estimated using bootstrap. **d**. Visualization of the constituency trees averaged over all test sentences. **p < 0.01, ***p < 0.001, two-sided Monte Carlo test for comparisons with the chance level, unpaired two-sided bootstrap for other comparisons, FDR corrected.

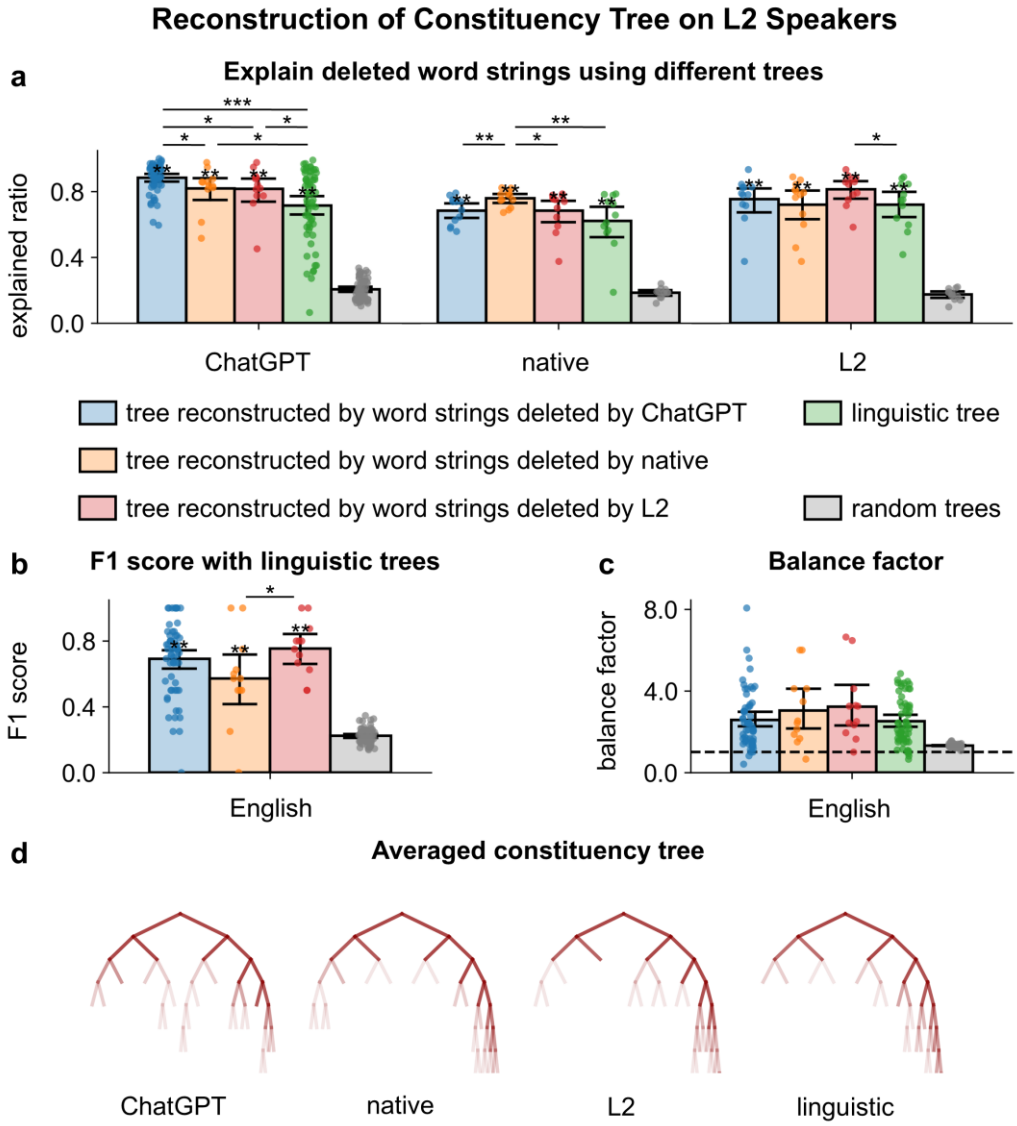

Figure S14. Reconstruction of constituency tree based on word deletion behavior for L2 English speakers (*N* = 34, 19-29 years old, mean 23 years old; 22 females), following the same procedure for native speakers in Dataset 4. The results are generally consistent to ChatGPT and native speakers. **a**. Explained ratios for different trees. The trees reconstructed by the L2 speakers explain nearly 80% of the deleted word strings (red bar in the right panel of Fig. S14), significantly higher than chance (p = 0.002, two-sided Monte Carlo test, FDR corrected). **b**. F1 score with respect to the linguistic tree. **c**. Balance factor of each constituency tree. Each dot represents a constituency tree (*N* = 12 for native and L2; *N* = 60 for ChatGPT, linguistic tree, and random trees). Error bars represent 95% CIs of the mean across constituency trees, estimated using bootstrap. **d**. Visualization of the constituency trees averaged over all test

sentences. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, two-sided Monte Carlo test for comparisons with the chance level; unpaired two-sided bootstrap for other comparisons, FDR corrected.

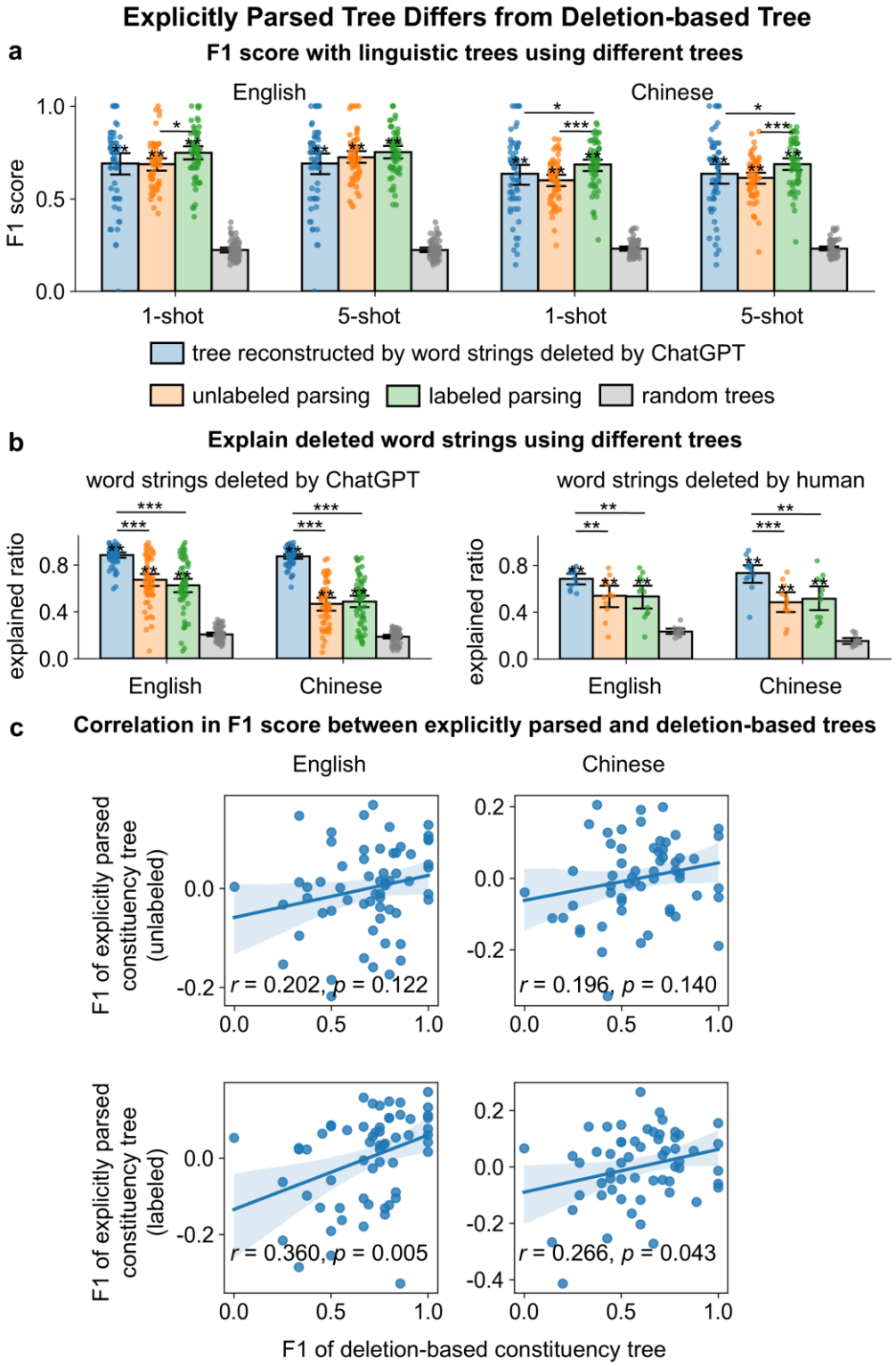

Figure S15. Comparison between syntactic tree explicitly parsed by ChatGPT and deletion-based tree. Explicit parsing is achieved using prompts[2] using either 1-shot or 5-shots learning. **a**. F1 score with respect to the linguistic tree. One-shot and 5-shot parsing obtain similar performance and we use 1-shot parsing for further analysis. **b**. Explained ratios for different trees. The deletion-

based trees explain more word-deletion responses than the explicitly parsed trees. Each dot represents a constituency (N = 60). Error bars represent 95% CIs of the mean across constituency trees, estimated using bootstrap. **c**. F1 score with treebank annotation is not strongly correlated between explicitly parsed and deletion-based trees, after regressing out metrics related to parsing difficulty (i.e., the syntactic depth, MDD of the sentences, as well as the F1 score obtained by the Stanford parser). Without regressing out these metrics, the correlation is still below 0.422 in all conditions. *p < 0.05, **p < 0.01, ***p < 0.001, two-sided Monte Carlo test for comparisons with the chance level; unpaired two-sided bootstrap for other comparisons, FDR corrected.

Supplementary Tables

Table S1. Detailed classification of deleted word string in Datasets 1 and 2.

a **Dataset 1**

English sentences	constituent	nonconstituent		other				
		multiple words	single word	multiple strings	reorder words	add words	no change	refuse to reply
ChatGPT	0.618	0.042	0.106	0.052	0.018	0.064	0.097	0.003
human	0.659	0.111	0.100	0.079	0.005	0.015	0.027	0.004

Chinese sentences	constituent	nonconstituent		other					
		multiple words	single word	multiple strings	delete characters	reorder words	add words	no change	refuse to reply
ChatGPT	0.473	0.097	0.158	0.092	0.025	0.020	0.104	0.031	0.0
human	0.658	0.107	0.040	0.100	0.036	0.004	0.033	0.010	0.012

b **Dataset 2**

English sentences	constituent	nonconstituent		other				
		multiple words	single word	multiple strings	reorder words	add words	no change	refuse to reply
ChatGPT	0.933	0.0	0.018	0.003	0.0	0.001	0.045	0.0
human	0.990	0.0	0.0	0.004	0.002	0.004	0.0	0.0

Chinese sentences	constituent	nonconstituent		other					
		multiple words	single word	multiple strings	delete characters	reorder words	add words	no change	refuse to reply
ChatGPT	0.429	0.009	0.265	0.064	0.009	0.012	0.086	0.125	0.001
human	0.971	0.0	0.007	0.008	0.002	0.001	0.010	0.0	0.001

Table S2. Detailed classification of deleted word string in Datasets 3 and S1.

a **Dataset 3**

English sentences	constituent	nonconstituent		other				
		multiple words	single word	multiple strings	reorder words	add words	no change	refuse to reply
ChatGPT	0.550	0.048	0.162	0.167	0.008	0.013	0.046	0.006
human	0.702	0.116	0.046	0.085	0.003	0.039	0.005	0.004

Chinese sentences	constituent	nonconstituent		other					
		multiple words	single word	multiple strings	delete characters	reorder words	add words	no change	refuse to reply
ChatGPT	0.479	0.072	0.106	0.121	0.033	0.048	0.091	0.048	0.002
human	0.716	0.103	0.037	0.075	0.031	0.007	0.025	0.003	0.003

b **Dataset S1**

English sentences	constituent	nonconstituent		other				
		multiple words	single word	multiple strings	reorder words	add words	no change	refuse to reply
ChatGPT	0.746	0.0	0.181	0.002	0.0	0.001	0.07	0.0

Chinese sentences	constituent	nonconstituent		other					
		multiple words	single word	multiple strings	delete characters	reorder words	add words	no change	refuse to reply
ChatGPT	0.853	0.066	0.024	0.005	0.0	0.0	0.005	0.047	0.0

Table S3. Constituent rate and explained ratio in Dataset 3 for each type of constituents.

a

English	type	# test	constituent	parent	node
ChatGPT	VP-NP	1792	0.717	0.776	0.041
	PP-NP	1828	0.527	0.500	0.0
	VP-PP	793	0.549	0.335	0.384
	NP-PP	501	0.289	0.0	0.500
	VP-ADJP	149	0.832	0.440	0.240
	VP-ADVP	69	0.725	1.0	0.0
human	VP-NP	191	0.759	0.766	0.101
	PP-NP	174	0.713	0.526	0.140
	VP-PP	111	0.694	0.514	0.446
	NP-PP	98	0.602	0.382	0.206
	VP-ADJP	26	0.769	0.947	0.053
	VP-ADVP	11	0.455	0.800	0.200

b

Chinese	type	# test	constituent	parent	node
ChatGPT	VP-NP	2440	0.570	0.217	0.620
	VP-PP	1360	0.393	0.033	0.738
	NP-DNP	1240	0.520	0.0	0.447
	NP-DP	620	0.397	0.0	0.636
	NP-QP	200	0.435	0.0	0.625
	PP-NP	180	0.511	0.447	0.053
human	VP-NP	271	0.727	0.212	0.703
	VP-PP	136	0.765	0.0	0.902
	NP-DNP	102	0.745	0.015	0.585
	NP-DP	53	0.811	0.032	0.806
	NP-QP	23	0.957	0.227	0.727
	PP-NP	12	0.667	0.250	0.500

Supplementary Results

Effects of Learning and Instruction

For ChatGPT, each test was processed independently, and the human participants were also instructed that all tests were independent. Nevertheless, it remained possible that the human participants adapted their rules as they were exposed to more tests, i.e., learning based on multiple demonstrations. Therefore, we also analyzed the response to each of the 24 tests that the human participants accomplished (Fig. S1) and found that the constituent rate did not significantly vary across tests (English: $F(23, 621) = 0.881$, $p = 0.626$, partial eta-square = 0.032; Chinese: $F(23, 621) = 1.236$, $p = 0.206$, partial eta-square = 0.042. One-way repeated measures ANOVA).

The task was also tested on human participants and ChatGPT using other instructions/prompts (Figs. S2 & S3). The results showed that both humans and ChatGPT preferred to delete constituents, but the performance of ChatGPT was more variable across different prompts. Some prompts elicited role-playing behavior in ChatGPT[41], e.g., by adding filler words like "*uh*" and "*um*" when the prompt attributed the transformed sentence to "*a man after a stroke*". Furthermore, compared with humans, ChatGPT showed a higher tendency to produce a grammatical word string that was semantically consistent with the test sentence (see Fig. S4).

Word-string Metrics

In the following, we employed several alternative metrics to characterize the deleted word string and the remaining word string, and analyzed which metric could most reliably discriminate the word strings deleted in the human/LLM experiment and random word strings (Fig. S5a). Among these alternatives, we considered candidate metrics derived from the tree-structured representations, e.g., mean dependency distance (MDD) of the deleted word string, as well as metrics derived from a linear sentence representation, e.g., length of the deleted word string (Fig. S5b, see Methods for the full list of metrics). We employed each metric to discriminate word strings deleted by humans/ChatGPT and randomly selected word strings using a support vector machine (SVM).

The results showed that, among individual metrics, constituency of the deleted string could best discriminate real human/ChatGPT results and random word strings, except for the Chinese ChatGPT test, in which the three tree-structure-based metrics, i.e., constituency, dependency connectivity, and MDD, reached similar performance (Figs. S5c & S5d). Overall, dependency-based metrics performed better on Chinese than English, possibly because constituency parsing is more challenging in Chinese due to the lack of case markers[3,4] or that dependency parsing is more suitable for languages with more flexible word order[5,6]. Furthermore, in terms of parsing strategies, participants appear to reply more on the constituency analysis for the deleted string than for the remaining string. This might be triggered by the fact that the deleted strings were consistently constituents. In contrast, dependency-based metrics of the remaining string outweighed the dependency-based metrics of the deleted string – A possible reason is that participants primarily adopted the dependency-based analysis for the remaining string since the constituency analysis is not applicable. When all metrics were jointly

considered, the discrimination rate was further boosted compared with when constituency was considered alone. Therefore, constituency of the deleted word string best characterized human/ChatGPT responses while other metrics also had some contributions.

Baseline Model

The discrimination analysis examined in which aspects real word deletion responses differed from random word strings. The analysis, however, only tested a few classic word-string metrics while humans/ChatGPT might exploit more fine-grained word properties to perform the task. Furthermore, the discrimination analysis was descriptive and could not directly predict the response based on the demonstration. Therefore, in the following, we further tested how well a baseline model could learn the word deletion task when it only had access to the ordinal position and fine-grained semantic/syntactic properties of each word (through position embedding and pretrained word embeddings). The baseline model was a naive GPT-2 model that was not pretrained based on large corpora and should be unaware of the constituency structure of language. To explore how well the baseline model could learn to delete a constituent based on the properties of individual words, we trained the model in a supervised manner by providing an answer to each test sentence (see Methods). The constituent rate of GPT-2 after learning one demonstration-test pair was treated as a baseline for the word deletion task, which was not higher than chance in both Dataset 1 (Fig. S6a, $p = 0.002$ for English and $p = 0.004$ for Chinese, two-sided Monte Carlo test, FDR corrected) and Dataset 3 (Fig. S6c, $p = 0.166$ for English and $p = 0.004$ for Chinese, two-sided Monte Carlo test, FDR corrected). After training with more samples, the constituent rate of GPT-2 gradually increased and reached the constituent rate of ChatGPT/humans after about 100/500 samples in Dataset 1 and about 100/3000 samples in Dataset 3. These results showed that, even with supervised learning, a baseline model unaware of the constituency

structure of a language needed hundreds of samples to reach the 1-shot constituent rate of ChatGPT and humans. We also analyzed the rule inferred by the baseline GPT-2 model, and found that the node-category rule consistently explained more responses than the parent-category rule for both Chinese and English (Figs. S6b & S6d).

Influence of Lexical Property

Since each sentence in Dataset 2 was tested on about 20 participants, we could analyze whether the rule preference differed across sentences that all had the same syntactic structure. We found the preference for node- vs. parent-category rule greatly varied across sentences (explained ratio of node-category rule varied from 0.0 to 0.333 in English and 0.292 to 0.905 in Chinese across sentences), suggesting that detailed word properties, not just the part of speech, could influence the participants' word deletion behavior. We also found the preference for the node-category rule reliably correlated between humans and ChatGPT across sentences (Fig. S7a, $r = 0.447$, $p = 0.013$ for English and $r = 0.806$, $p < 0.001$ for Chinese). Furthermore, human participants more strongly preferred the node-category rule if the verb-preposition combination in the sentence formed a stronger collocation (Fig. S7b, $r = 0.439$, $p = 0.015$ for humans and $r = 0.556$, $p = 0.001$ for ChatGPT). Note that the node-category rule was to delete the NP while the parent-category rule was to delete the preposition and the NP. Therefore, when the verb-preposition formed a stronger collocation, the participants avoided breaking up the collocation and therefore preferred the node-category rule.

Meaningless Sentences

We further tested whether ChatGPT could still perform the word deletion task when semantic information is largely absent by constructing syntactically well-formed but meaningless sentences. Another purpose was to test how

ChatGPT responded to sentences that were unlikely to have appeared in its training set. Since the training sets of ChatGPT and many other LLMs were not disclosed, it was possible that they included the treebank sentences. In Dataset S1, we constructed syntactically correct but meaningless sentences based on the sentence structure in Dataset 2, i.e., "*NP1 verb preposition NP2"*. An example of the constructed sentence was "*The obvious summer lies on the retrieval*". For these meaningless sentences, ChatGPT still deleted constituents much more frequently than chance (Fig. S8b, $p = 0.002$ for both Chinese and English, two-sided Monte Carlo test, FDR corrected).

Syntactic Ambiguity

Parsing a sentence into structurally interconnected constituents is subject to not only syntactic constraints, but also semantic ones, especially for sentences with ambiguous syntactic structures. Therefore, we analyzed whether humans and ChatGPT considered semantic constraints when analyzing syntactically ambiguous constituents in Dataset S2. We tested two types of constructions involving structural ambiguity, both of which could be syntactically parsed into two potential constituency trees, referred to as structure 1 and structure 2 (Fig. S9ab). By manipulating the words in the sentence, we constrained that only one structure was associated with a plausible meaning.

For ChatGPT, the constituent rate in Dataset S2 was 0.66 ± 0.013 (95% CI) and 0.99 ± 0.0015 for adjunct-attachment and PP-attachment sentences. For humans, the constituent rate was 0.76 ± 0.085 and 0.95 ± 0.040 for adjunct-attachment and PP-attachment sentences. The analysis focused on what we called the target string, i.e., a word string that was a semantically plausible constituent in structure 1 but not structure 2 (see Fig. S9ab). If a participant deleted the target string more frequently when structure 1 was associated with a plausible meaning, i.e., when the target string was a semantically plausible

constituent, it suggested that the participant integrated semantic information to determine which constituent to delete. Nevertheless, the results showed that ChatGPT deleted the target string slightly but significantly more frequently when structure 2 was associated with plausible meaning (Fig. S9cd, $p = 0.027$ for adjunct attachment and $p < 0.001$ for PP attachment, paired two-sided bootstrap, FDR corrected), while the frequency of native English speakers deleting the target string was not significantly modulated by the plausibility of structure 1 or 2 ($p = 0.468$ for adjunct attachment and $p = 0.145$ for PP attachment, paired two-sided bootstrap, FDR corrected).

We additionally tested high-proficiency L2 English speakers, and found that L2 speakers deleted the target string more frequently when structure 1 was associated with a plausible meaning ($p < 0.001$ for adjunct attachment and $p = 0.008$ for PP attachment, paired two-sided bootstrap, FDR corrected). The constituent rate for L2 speakers was 0.79 ± 0.114 and 0.98 ± 0.015 for adjunct-attachment and PP-attachment sentences. These results indicated that native speakers and ChatGPT were not strongly influenced by sentence-level semantic plausibility when performing the word deletion task, but high-proficiency non-native speakers were significantly, although not strongly, influenced by semantic plausibility.

Properties of the Deletion-based Trees

To further examine the overall characteristics of the deletion-based trees, we averaged the deletion-based trees, as well as the linguistic trees, over all test sentences (Fig. 6d). In general, the deletion-based constituency trees had similar depth and width with the linguistic constituency trees for each language. The linguistic tree tended to have more right descendant nodes in

English than Chinese, and a similar trend was observed for the deletion-based tree. We quantified this trend using balance factors, i.e., the total number of right descendant nodes divided by the total number of left descendant nodes (see Methods). The balance factor was significantly higher in English than Chinese for both the linguistic constituency tree (Fig. 6c, $p < 0.001$, unpaired two-sided bootstrap, FDR corrected) and the deletion-based constituency tree ($p < 0.001$ for both humans and ChatGPT, unpaired two-sided bootstrap, FDR corrected).

Supplementary Discussion

Language-dependent rule inference

We observe that both human participants and ChatGPT prefer different rules to determine which constituent should be deleted when processing different languages, i.e., Chinese and English. A likely reason is that English is more of a right-branching language, in which the syntactic tree tends to have more right descendent nodes[7,8]. It has been proposed that top-down parsing strategies, which first consider parent nodes, are more memory efficient than bottom-up parsing strategies, which first consider child nodes, for right-branching languages, and the opposite is true for left-branching languages[9,10]. Therefore, it is possible that both humans and LLMs analyze different languages through different constituency parsing strategies, and such strategies are deployed for the word deletion task, resulting in the language-dependent rule inference. Importantly, a recent fMRI study has demonstrated that the neural responses recorded during Chinese and English sentence processing tasks separately fit the predictions of bottom-up and top-down parsers[8]. The idea that a person or model can utilize different parsing strategies for different languages is consistent with the principles and parameters framework in linguistics[11,12] – The structure of different languages shares a number of common principles but differs in its parameters, e.g., head-initial or head-final. Our data and previous results[8] suggest that humans and LLMs may flexibly switch their analysis parameters based on the input language.

Variability on the word deletion task

The word deletion task here is a highly ambiguous one-shot learning task – Many rules can possibly explain the single demonstration and we aim to test whether the participants randomly sample from possible rules or show preference to some rules. Overall, both the human participants and ChatGPT

prefer to delete a constituent but variations are observed across participants, trials, and experiments. In Dataset 2, which tests relatively simple sentences, the constituent rate is consistently high, i.e., >0.9, for human participants. In experiments testing treebank sentences, which generally have more complicated syntactic structures than the sentences in Dataset 2, the constituent rate is often below 0.75. Therefore, sentence complexity is a factor that increases response variability.

Two factors may lead to response variability for more complex sentences. First, multiple rules that can explain the demonstration (e.g., rules in Fig. 1b) may compete when participants perform the task. A complex sentence has many constituents that can all be candidates to delete – Choosing from many candidates is a challenging task and therefore some participants may resort to inferring rules based on shallower cues, e.g., the linear word order. Second, there may be individual variability in the internal sentence representation[13] and the internal sentence representations in humans and LLMs may not perfectly align with the annotations in treebanks. If a participant did not perceive the deleted word string in the demonstration as a constituent, he/she would not infer that the rule was to delete a constituent. Furthermore, even if participant intended to delete a constituent, the deleted word string would not be scored as a constituent if it was not annotated as a constituent in the treebank.

Semantic influence on the word deletion task

Although lexical information affects the word deletion task (Fig. S7), sentence-level semantic plausibility barely influences how native speakers perform the word deletion task (Datasets S1 and S2). A possible explanation for Dataset S2 is that the word deletion task does not explicitly require sentence-level semantic comprehension so that the participants do not actively resolve the two ambiguous interpretations[14–17]. In contrast, human participants who are

proficient but not native speakers of the testing language show evidence of integrating semantic information in the word deletion task, consistent with the idea that L2 speakers focus more on lexico-semantic cues and underuse syntactic analysis[18]. Different from the behavior of human participants, ChatGPT shows a weak but significant influence of semantics but tended to delete a semantically implausible constituent. This behavior potentially reflects incorrect integration of semantic and syntactic information, which also echoes previous demonstrations that LLMs trained only with text data often show deficits in world knowledge[19–21].

Reference

1. Oxford University Press. *Oxford Collocation Dictionary (Bilingual Version)*. (Foreign Language Teaching and Research Press Pub, 1991).
2. Tian, Y., Xia, F. & Song, Y. Large Language Models Are No Longer Shallow Parsers. in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds. Ku, L.-W., Martins, A. & Srikumar, V.) 7131–7142 (Association for Computational Linguistics, Bangkok, Thailand, 2024). doi:10.18653/v1/2024.acl-long.384.
3. Li, C. N. & Thompson, S. A. *Mandarin Chinese: A Functional Reference Grammar*. (Univ of California Press, 1989).
4. Wu, F. & He, Y. Some Typological Characteristics of Mandarin Chinese Syntax. in *The Oxford Handbook of Chinese Linguistics* (Oxford University Press, 2015). doi:10.1093/oxfordhb/9780199856336.013.0020.
5. Huang, C.-T. J. *Logical Relations in Chinese and the Theory of Grammar*. (Taylor & Francis, 1998).
6. Kübler, S., McDonald, R. & Nivre, J. Dependency Parsing. in *Dependency Parsing* 11–20 (Springer International Publishing, Cham, 2009). doi:10.1007/978-3-031-02131-2_2.
7. Levy, R. & Manning, C. D. Is it Harder to Parse Chinese, or the Chinese Treebank? in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* 439–446 (Association for Computational Linguistics, Sapporo, Japan, 2003). doi:10.3115/1075096.1075152.

8. Zhang, X., Wang, S., Lin, N. & Zong, C. Is the Brain Mechanism for Hierarchical Structure Building Universal Across Languages? An fMRI Study of Chinese and English. in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (eds. Goldberg, Y., Kozareva, Z. & Zhang, Y.) 7852–7861 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022). doi:10.18653/v1/2022.emnlp-main.535.

9. Mazuka, R. & Lust, B. On Parameter Setting and Parsing: Predictions for Cross-Linguistic Differences in Adult and Child Processing. in *Language Processing and Language Acquisition* (eds. Frazier, L. & De Villiers, J.) 163–205 (Springer Netherlands, Dordrecht, 1990). doi:10.1007/978-94-011-3808-6_7.

10. Resnik, P. Left-corner parsing and psychological plausibility. in *Proceedings of the 14th Conference on Computational Linguistics - Volume 1* 191–197 (Association for Computational Linguistics, USA, 1992). doi:10.3115/992066.992098.

11. Chomsky, N. *A Minimalist Program for Linguistic Theory*. (The MIT Press, 1993).

12. Chomsky, N. & Lasnik, H. The Theory of Principles and Parameters. in *An International Handbook of Contemporary Research* (eds. Jacobs, J., Stechow, A. von, Sternefeld, W. & Vennemann, T.) 506–569 (De Gruyter Mouton, Berlin • New York, 1993). doi:doi:10.1515/9783110095869.1.9.506.

13. Kidd, E., Donnelly, S. & Christiansen, M. H. Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences* **22**, 154–169 (2018).

14. Sanford, A. J. & Sturt, P. Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Sciences* **6**, 382–386 (2002).

15. Ferreira, F. & Patson, N. D. The ‘Good Enough’ Approach to Language Comprehension. *Language and Linguistics Compass* **1**, 71–83 (2007).

16. Townsend, D. J. & Bever, T. G. *Sentence Comprehension: The Integration of Habits and Rules*. (MIT Press, Cambridge, MA, 2001).

17. Ferreira, F. The misinterpretation of noncanonical sentences. *Cognitive Psychology* **47**, 164–203 (2003).

18. Clahsen, H. & Felser, C. Grammatical processing in language learners. *Applied Psycholinguistics* **27**, 3–42 (2006).

19. Mooney, R. J. Learning to connect language and perception. in *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3* 1598–1601 (AAAI Press, Chicago, Illinois, 2008).

20. Bisk, Y. *et al.* Experience Grounds Language. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds. Webber, B., Cohn, T., He, Y. & Liu, Y.) 8718–8735 (Association for Computational Linguistics, Online, 2020). doi:10.18653/v1/2020.emnlp-main.703.

21. Marcus, G. The next decade in AI: four steps towards robust artificial intelligence. *ArXiv* **abs/2002.06177**, (2020).

List of parallel sentences

Parallel sentences with the structure of [S [NP] [VP [NP]]]:

Accounting problems raise more knotty issues
会计问题引发更多棘手事件

Prospective competition is one problem
预期的竞争是一个问题

These people have different agendas
这些人有不同的议程

Pension reform was its main thrust
养老金改革是其主要推动力

Rising operating expenses are another problem
上升的运营费用是另一个问题

Some institutional traders loved the wild ride
一些机构交易员喜欢这种疯狂的旅程

The plan lacked a withdrawal timetable
该计划缺乏撤回时间表

Two election commission members opposed the matching plans
两名选举委员会成员反对配对计划

The suit seeks unspecified damages
该诉讼寻求未具体说明的损害赔偿

Many law firms sponsor their own programs
许多律师事务所赞助自己的项目

Hostile takeovers are quite a new phenomenon
恶意收购是一个相当新的现象

His father ran a construction company
他的父亲经营一家建筑公司

Potential side effects include high blood pressure
可能的副作用包括高血压

The federal judge granted a temporary restraining order
联邦法官授予临时限制令

This boy needs your help
这个男孩需要你的帮助

The young girl loves her new toy
那个年轻的女孩爱她的新玩具

The old man feeds the stray cat
那个老人喂那只流浪猫

The diligent student learns her lessons
勤奋的学生学习她的课程

My little brother enjoys this interesting game
我的小弟弟享受这个有趣的游戏

Local residents appreciate the beautiful park
当地居民欣赏美丽的公园

Your father deserves a vacation
你的父亲值得一个假期

Our friendly neighbor shares his delicious food
我们友好的邻居分享他美味的食物

Their enthusiastic coach trains the young team
他们热情的教练训练年轻的团队

The local community supports our fundraising event
当地社区支持我们的筹款活动

Her father protects the family
她的父亲保护家人

The urban children receive free education
城市的孩子接收免费的教育

City dwellers crave a quiet countryside life
城市居民渴望安静的乡村生活

Most people respect this great leader
大部分人尊敬这位伟大的领导者

My friends cherish our high school memories
我的朋友们珍惜我们的高中回忆

This beautiful painting evokes strong emotions
这幅美丽的画引发强烈的情感

Parallel sentences with the structure of [S [NP] [VP [PP [NP]]]]:
His father died of a heart attack
他的父亲死于一次心脏病

The accident happened on Saturday night
这场意外发生在周六晚上

His family lived in an old farmhouse
他们一家生活在一个老旧农舍里

Most people live in rural areas
大部分人生活在农村地区

the rest students sat on the ground
剩余学生坐在地上

the old man looked at the window
这个老人看向窗户

A surprising revelation happened on their wedding anniversary
一件惊讶的事情发生在他们的结婚纪念日上

A full moon appeared in the dark night sky
一轮圆月出现在漆黑的夜空

The students look at the complex equation
学生们看向复杂的方程式

The bird rests on the sturdy branch
小鸟们停在粗树枝上

A gentleman walks towards the bustling cafe
一位绅士走向繁忙的咖啡厅

The thick essay lies on the shelf
厚厚的论文摆放在书架上

A picture hangs at his study
一张照片挂在他的书房

The fresh linens lie in the laundry basket
干净的床单放在洗衣篮中

Many squirrels live in the oak tree
许多松鼠生活在橡树上

The road sign points to the nearest exit
路标指向最近的出口

The fragile vase stays on the top shelf
易碎的花瓶放在顶层书架

The little girl walks towards the living room
小女孩走向卧室

the old guests sat at the dinner table
年老的客人们坐在晚餐桌旁

The great artist died from a serious illness
这位伟大的艺术家死于一场重病

His glasses lie on the bedside table
他的眼镜放在床头柜上

The golden treasure lies in the dense jungle
黄金宝藏位于茂密的丛林

The old map points to a hidden treasure
旧地图指向隐藏的宝藏

The small boat stays in the calm lake
小船停在平静的湖面上

The young boy looks at the toy store window
小男孩看向玩具店的橱窗

A lost puppy walks towards the city park
一只迷路的小狗走向城市公园

The secret letter lies on the old wooden desk
这封秘密信放在旧木桌上

The rare artifact stays in the secured vault
这件稀有文物保存在安全保险库

His parents rushed to his side
他的父母冲到他身旁

The story begins in a small town
这个故事始于一个小镇

List of syntactically correct but meaningless sentences

English meaningless sentences:

His enthusiasm died of a price sight
The construction happened on giant economy
His chance lived in a light radio
Most data live in steep data
The opening funds sat on the bid
The common effort looked at the trading
A disappointing recognition happened on their session week
A unexplained niche appeared in the high argument meat
The cultures look at the substantial reason
The budget rests on the global telephone
A activity walks towards the slowing year
The obvious summer lies on the retrieval
A year hangs at his urgency
The thin banks lie in the new industry
Windy sales live in the sleepy hallmark
The source business points to the best deficiency
The similar house stays on the rich culture
The same furor walks towards the mail loss
The touchy machines sat at the process business
The several company died from a great space
His facts lie on the alert war
The similar portfolio lies in the high touchdown
The real panic points to an renewed mistrust
The several demand stays in the consecutive industry
The other board looks at the sense guide window
A specialized share walks towards the forest office
The strong commission lies on the big blue cost
The intimate bill stays in the invented amount
His convulsions rushed to his taste
The director begins in a first noise

Chinese meaningless sentences:

什么接口死于这老公
这屏幕发生在驾照精神
别的户口生活在其他宏伟护士里
这经理生活在耐心公园

很多安排坐在事情上
这些英语看向显卡
多少低矮的意思发生在你的电梯上
两个价位出现在严肃的历史上
暑假医生看向新的幸福
想法护照停在根本交换机上
一家酒店走向繁忙的眼霜
大型的达人摆放在网线中
两个教育挂在你的科技
纯粹的师傅放在房东里
不少家庭生活在天线里
公园指向行政的天线
男的结果放在友情科技
小客套走向常识
活的电子会计坐在公安时间上
这些个大的男人死于好多高中
俺的势头放在游击队里
机会水平位于所谓的学期
胖标书指向原版的老乡
光头标书停在夜行的电子版上
老工作看向价格的对话
一条新的时间走向人品学前班
哪个大厕所放在科技中
这同质化策划保存在同事样子上
我的燃料冲到工号旁边
那个电池始于不少工号

List of syntactically ambiguous sentences

Sentences with adjunct attachment:

The letter of the writer that had blonde hair arrived this morning
The writer of the letter that had blonde hair arrived this morning

The flowers of the valley that had the old castle excited the tourists
The valley of the flowers that had the old castle excited the tourists

The thesis of the editor that had the big nose made a lot of sense
The editor of the thesis that had the big nose made a lot of sense

The machine of the inventor that had the goatee was amazing
The inventor of the machine that had the goatee was amazing

The church of the bishop that had the funny eyebrows looked odd
The bishop of the church that had the funny eyebrows looked odd

The car of the driver that had the moustache was pretty cool
The driver of the car that had the moustache was pretty cool

The chef of the restaurant that had the blue tiles pleased us
The restaurant of the chef that had the blue tiles pleased us

The painter of the house that had the small windows looked odd
The house of the painter that had the small windows looked odd

The supplier of the drugs that had a nasty effect hurt everyone
The drugs of the supplier that had a nasty effect hurt everyone

The gang of the criminal that had a long scar disappeared last Monday
The criminal of the gang that had a long scar disappeared last Monday

The song of the singer that had long eyelashes was very smart
The singer of the song that had long eyelashes was very smart

The miner of the gold that had the impurities was worthless
The gold of the miner that had the impurities was worthless

Sentences with PP attachment:

The guy caught the rat with a scar
The guy caught the rat with a trap

Bill cut the paper with a very bright colour
Bill cut the paper with a very sharp knife

Dan entered the room with a window
Dan entered the room with a key

Charles runs a photo studio with two floors
Charles runs a photo studio with two friends

The police investigated the house with a garden
The police investigated the house with the permission

Jasmine sang the song with the romantic lyrics
Jasmine sang the song with the lovely students

Queen Elizabeth II ate the cake with a red rose
Queen Elizabeth II ate the cake with a silver fork

Michelle watched the movie with English subtitles
Michelle watched the movie with her boyfriend

The basketball player won the game with lots of audience
The basketball player won the game with his team

The professor recruited the students with black eyes
The professor challenged the students with his projects

The doctor noticed the boy with big headphones
The doctor saved the boy with a pacemaker

Van Gogh drew the paintings with starry night
Van Gogh sold the paintings with a broken heart

The cook ate the meal with some cauliflowers
The cook made the meal with a cleaver

The magician thanked the participant with a big nose
The magician amazed the participant with a fancy scene

The coach led the team with three members
The coach motivated the team with inspiring words

The comedian saw the crowd with fresh flowers
The comedian entertained the crowd with witty jokes

Mom ate the chocolate with chopped nuts
Mom melted the chocolate with an iron pot

Bruce likes the rock band with different instruments
Bruce leads the rock band with his voice

Demonstration sentences for adjunct attachment:
The collection of the store fascinated the students
The house of the painter looked odd
The plains of the tribe looked strange
The attorney of the company bothered me
The forest of the animals pleased us
The manager of the factory was efficient
The king of the mountain impressed Arthur
The director of the orchestra played well
The dog of the officer came back this evening
The designer of the dress took a bow
The photographer of the exhibition sold his favorite piece
The artist of the mural completed his work
The sailor of the ship arrived this afternoon
The dancer of the troupe dazzled the audience
The library of the school attracted many scholarse
The exhibit of the museum intrigued the historians
The park of the city delighted the children
The song of the player is really popular
The beach of the island mesmerized the couples
The lake of the area enchanted the swimmers
The festival of the village entertained the guests
The castle of the kingdom awed the tourists
The bakery of the district tempted the locals

The plaza of the metropolis energized the crowd
The gallery of the center impressed the critics
The alley of the town fascinated art lovers
The orchard of the farm tempted the foodies
The market of the city attracted the collectors
The observatory of the institute amazed the astronomers
The mansion of the celebrity excited the fans

Demonstration sentences for PP attachment:
The photographer adjusted the lens
The bird sang a melody
The chef cooked the meal
The teacher explained the concept
The gardener watered the plants
The fisherman caught the trout
The student solved the equation
The carpenter built the house
The programmer wrote the code
The dancer performed the routine
The singer composed a song
The pilot flew the airplane
The detective solved the mystery
The baker baked the cake
The journalist reported the news
The hiker climbed the mountain
The lawyer argued the case
The baker kneaded the dough
The musician tuned the guitar
The mechanic fixed the engine
The architect designed the building
The librarian sorted the books
The jeweler polished the diamond
The sculptor chiseled the statue
The chef seasoned the soup
The tailor altered the coat
The student read the book
The gamer played the game
The driver steered the car
The actor memorized the script

Prompts for LLMs

Prompt 0 was employed in the main text, and the experiments conducted with GPT-4, Claude-3, Gemini-1, Llama-3 were also employed prompt 0. The remaining prompts 1-5 were employed on Dataset 1 with ChatGPT (see results on Fig. S2). For Chinese experiments, the prompts were the translated version of the English prompts.

Prompt 0
John developed a very special way of speaking. For '{demonstration sentence}', he would say '{deleted sentence}'.

Please carefully examine John's speaking style, and guess what he would say for the sentence: '{test sentence}'.

Prompt 1
A man speaks in a really weird way after a stroke. For '{demonstration sentence}', he would say '{deleted sentence}'.

What will he say for the sentence: '{test sentence}'? Do not add any extra word. Please think step by step and give out the most possible answer you think.

Prompt 2
An alien is learning our language, but he tends to omit some words in his sentences. For '{demonstration sentence}', he would say '{deleted sentence}'.

Please guess how the alien would express the following sentence: '{test sentence}'.

Prompt 3
A young child is facing difficulties in language acquisition, often forgetting to say certain words. When he tries to say '{demonstration sentence}', he ends up saying '{deleted sentence}'.

Can you anticipate what he would say for the sentence: '{test sentence}'?

Prompt 4

A robot has been programmed to communicate in a unique way. When it tries to say '{demonstration sentence}', it outputs '{deleted sentence}'.

Can you predict how the robot would express the sentence: '{test sentence}'?

Prompt 5

In a parallel universe, the language construct is slightly different. The sentence '{demonstration sentence}' would be communicated as '{deleted sentence}'.

How would the sentence '{test sentence}' be conveyed in this universe?

Prompt 0

小张有一种非常奇特的说话方式。对于"{demonstration sentence}"，他会说成"{deleted sentence}"。

请你认真分析小张的说话方式，然后猜测他会如何表达"{test sentence}"这个句子。

Prompt 1

这个人中风后的讲话方式真的很奇怪。对于"{demonstration sentence}"，他会说成"{deleted sentence}"。

那么对于句子"{test sentence}"，这个人会怎么说呢？请不要添加额外的新词，给出你觉得最有可能的答案。

Prompt 2

一个外星人正在学习我们的语言，但他总是在句子中删掉一些词。对于"{demonstration sentence}"，他会说成"{deleted sentence}"。

请你猜测这个外星人会如何表达"{test sentence}"这个句子。

Prompt 3

一个小孩在学习语言的时候遇到了困难，经常忘记说某些词。当他试图说"{demonstration sentence}"时，他却说成了"{deleted sentence}"。

你能预测他会如何表达"{test sentence}"这个句子吗?

Prompt 4

一个机器人被编程为用一种独特的方式进行交流。当它尝试说“{demonstration sentence}”时，它会输出“{deleted sentence}”。

你能预测这个机器人如何表达“{test sentence}”这句话吗？

Prompt 5

在一个平行宇宙中，语言结构有些不同。“{demonstration sentence}”这个句子会被表达为 “{deleted sentence}”。

在这个宇宙中“{test sentence}”应该如何表达呢？

Instruction for human participants

Human participants were provided with a task instruction before experiments.

All English versions of experiments employed the instructions below:

In this experiment, you will be presented with some texts followed by questions. Your task is to answer the questions based on the information provided in the text. Each question should only be answered based on the current text, please do not refer to the previous text or questions. Some trials may feel very complex or difficult. Just try your best.

All Chinese versions of experiments employed the instructions below:

在实验中，您需要回答一些以中文形式呈现的问题。实验中的每个问题都是独立的，与之前或之后的问题无关。您对当前问题的回答不应受到先前或后续问题的影响。即使有些单词或表达方式您不熟悉，也请尽力而为。大多数情况下，不认识题目中所有单词并不影响回答问题。

During the human experiments, instruction 0 (the same as prompt 0 for LLMs) was employed in the main text, and instruction 1 (the same as prompt 1 for LLMs) was employed on Dataset 1 (see results on Fig. S3).

Instruction for explicit constituency parsing

ChatGPT was asked to annotate the linguistic constituent based on the following instruction:

The following is a sentence that has already been tokenized. Words are separated by white spaces. Please parse the sentence with given words.
[sentence]: {demonstration sentence}
[parse tree]: {parsed demonstration sentence}
[sentence]: {test sentence}
[parse tree]:

For Chinese experiments, the instruction was the translated version of the English instruction:

以下是一个中文句子。请对句子进行解析。
[句子] : {demonstration sentence}
[解析树] : {parsed demonstration sentence}
[句子] : {test sentence}
[解析树] :