nature portfolio

# Peer Review File

# Active Use of Latent Tree-structured Sentence Representation in Humans and Large Language Models

Corresponding Author: Dr Nai Ding

Version 0:

Decision Letter:

25th September 2024

Dear Dr Ding,

Thank you once again for your manuscript, entitled "Active Use of Latent Constituency Representation in both Humans and Large Language Models". Please accept my sincere apologies for the delay in contacting you with a decision on your manuscript and thank you for your patience during the review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to (1) identifying key priorities that should be addressed in revision and (2) overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritised set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

In particular, we ask you to address the following (as well as all other reviewers' concerns):

1) improve the literature review, synthesise findings in the field, in particular from neuroscience of language. This is important to show the robustness of outcomes in response to various theories because you challenge the existing results.

2) carry out additional analyses in response to methodological concerns raised by Reviewers 2 and 3, including running GPT-2 as baseline and multiple shots given the sensitivity of LLMs

3) address the reviewers' concerns that the training data maybe including phrase annotations (Reviewers 1 and 3).

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript, we would hope to receive it within 4 months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

• Include a “Response to the editors and reviewers” document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim

followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

**Note:** This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,

[redacted]
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: neurolinguistics, language constituents

Reviewer #2: LLMs, language representation, computational neuroscience

Reviewer #3: parsing algorithm, machine translation

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

This paper presents a set of experiments testing the latent grammatical structure in the representations formed by pre-trained LLMs and human language users. A key innovation is a one-shot learning task given to both kinds of language agents in which they are prompted to remove words from a target sentence. The principle finding is that both humans and LLMs, in both English and Chinese, tend to remove word sequences that form a phrase rather than sequences that do not. This is interesting because (a) phrases are typically not explicitly marked in the input data (for either humans or LLMs); these patterns must be abstracted during learning, and (b) the novel one-shot task minimizes potential overlap with statistical patterns in the training data. A series of experiments demonstrated that results from the task, both human and LLM, can be used to reconstruct the syntactic prhase-structure for the target sentences with reasonable accuracy, this accuracy generalizes across English and Chinese, and across different current LLMs.

Overall this is an elegent and well-executed series of experiments. The paper is clearly written and the contributions are nicely situated in comparison to existing work on the linguistic/syntactic abilities of LLMs and also the linguistic methodologies for constituency tests which are adapted here.

I really only have a clarification question. The authors note clearly that a central challenge is designing a test that avoids overlap with the training data. The chosen task is well adapted for this (e.g. it avoids using standard constituency-testing methods that may be explicitly present in the training data.) Still, I do wonder whether phrase annotations for English and Chinese may be present in the training data in some form. The training data used by ChatGPT and other models is not open for scrutiny, but scientific articles (including linguistics) make a major contribution to the MassiveText dataset that was used to train Gopher (sciencedirect is the top high-level URL contributing text; pubmed is in the top 3). If such annotations are present (e.g. the bigram "the remark" in Fig 1 or a bigram with a similar enough embedding is explicitly annotated as "NP"), then it seems plausible that such explicit information might nudge the LLM towards the observed behavior explicitly, rather than implicitly as argued. Is there any information about whether the training data contains linguistic analysis, or perhaps annotated corpora (even portions of the English and Chinese penn treebanks used in these studies)?

It may be that these are not answerable questions given the proprietary nature of some of the key LLMs tested here (though I would hope models that promote open access, like Llama, would make this informaiton available). The upshot is that it might be helpful to discuss what is and isn't known about the training data in terms of explicit syntactic annotations.

Reviewer #2 (Remarks to the Author):

This work combines experiments involving both humans and state-of-the-art large language models to demonstrate that, behaviorally, both access and utilize linguistic knowledge in the form of latent constituency representations to perform the word deletion task. The word deletion task is a novel approach, and while it does not directly target linguistic behavior, it allows for the exploration of linguistic knowledge. The results would be of significant interest to the field if the authors can expand on the range of hypotheses that could account for the behavioral data.

My main critique of the paper focuses on how the behavioral responses to the task are evaluated. In the design of the experimental materials for Experiment 1, the only condition considered is deletion based on constituency. The authors show in Figure 1C that human behavior aligns with deletion on constituents. However, there is a large variability in both the constituent and non-constituent conditions. According to the methods, non-constituents are defined as anything not identified as a NP constituent by the constituency tree (Chomsky, 1957). This is not a strong alternative hypothesis, as it ignores other formalizations of sentence structure and does not provide a robust null hypothesis.

Several questions arise: Are participants not behaving in a linguistically informed way or instead their behavior could be explained better by an alternative account, say Dependency length minimization? For trials that they don't not follow the constituency, is there a consistency in their response? Are there specific sentences that elicit non-constituency behavior across participants? A more principled approach to characterizing human behavior would involve developing a set of theories based on constituency and other linguistic or null theories that aim to capture human behavior across all trials and show that overall constituency explained behavioral patterns better.

This points to a more general issue with the paper's presentation and how it synthesizes existing results in the field to motivate the research. Work in the neuroscience of language, specifically (Fedorenko et al., 2016, 2024), has challenged the notion of syntax as a core driver or separate module in language processing as evidenced by activity in the the brain regions responsible for language processing, (see also Goodman, 1997; Linebarger et al., 1983).

Additionally, the authors need to clarify why they consider their task a form of active language use. Why should human behavior in a task that is meta-linguistic and not representative of everyday language use be taken as evidence that they actively use constituency to represent language?

Other comments

The LSTM baseline is not particularly informative, as it primarily controls for results based purely on learning. The fact that it only requires about 50 demonstrations to learn the task raises questions about the level of linguistic knowledge required to perform the task. For example, if the subjects identify that the task is about grammar and assume they have received some grammar training during their education, they could solve the task by simply referring to their acquired knowledge of grammar as taught in school. This approach contrasts with how a preschool child learns and uses language actively.

In the last paragraph before the section "Sensitivity to Syntactic Category," the authors argue that the phenomena cannot be explained by statistical cues in word properties. However, I am uncertain how this conclusion is drawn from the LSTM results.

If the demonstration did not follow a constituency rule, the authors' notion suggests that humans would still be biased toward using constituency-based parsing to modify the test. I believe that demonstrating this bias would enhance the results by providing additional evidence of humans' reliance on constituency structures, even in non-standard scenarios.

Another point of concern is the sensitivity of large language models (LLMs) to prompts. If this is a one-shot learning task that does not depend on context beyond the demonstration sentence, why does Figure S2 show that the prompt can substantially affect the model's behavior (as seen with prompt 2, prompt 3, and prompt 4)? This observation echoes the earlier comment regarding the non-constituent and other responses that the models provide. This is in contrast to humans, who show consistency in their response patterns (Figure S3).

The authors need to do more work in both the introduction and discussion sections to highlight theories and experimental results that challenge the original constituency parsing hypothesis. The first paragraph of the introduction lacks a comprehensive review of the various approaches that exist in the field to address the main question. While the authors refer to the constituency as an influential hypothesis supported by research in psychology and neuroscience, they fail to cover alternative accounts that challenge this hypothesis. Empirical work by researchers such as Fedorenko, among others, highlights significant observations that contradict the dominance of unique syntactic parsing for explaining activity in regions in the human brain that process language, for a review see (Fedorenko et al., 2024).

In figure 5C, it is unclear to me why the points depicting the results are in bands. Can authors show the results per-subject basis, to see whether the effect is present in every subject.
In figure 5C the baseline of chatGPT proportions is much lower compared to other figures, why is this the case?

References

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. Nature Reviews. Neuroscience. https://doi.org/10.1038/s41583-024-00802-4

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the

construction of sentence meaning. Proceedings of the National Academy of Sciences of the United States of America, 113(41), E6256–E6262.

Goodman, E. B. J. C. (1997). On the Inseparability of Grammar and the Lexicon: Evidence from Acquisition, Aphasia and Real-time Processing. Language and Cognitive Processes, 12(5-6), 507–584.
Linebarger, M. C., Schwartz, M. F., & Saffran, E. M. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics. Cognition,

Reviewer #3 (Remarks to the Author):

This work studies whether a large language model (LLM) can understand the concept of constituency of English and Chinese in the same manner as done by humans. Basic ideas is to ask them to delete a word sting by following an example of deleting a valid constituent, e.g., a noun phrase, to see if they can correctly identify a similar phrase in a sentence. Four experiments are conducted in this work: 1) verifying the sensitivity of identifying a noun phrase and an impact by its parent category, 2) measuring the overall accuracies of all kinds of constituency, 3) syntactic tree reconstruction by using deleted phrases, and 4) measuring the semantic plausibility using prepositional phrase attachment. The series of experiments shows that ChatGPT behaves similarly to humans in both English and Chinese.

The idea of identifying a constituent by deletion is interesting and it has not been investigated at least in computational linguistics. The experiments for both an LLM and humans clearly show some correlation between them, and the findings are of interest to the research community. I have some concerns and suggestions to this work as follows:

- The tree reconstruction using CKY parsing is very interesting but the current setting sounds not apt to me. Given Figure 3, the probabilities associated with all strings in a derivation are treated as scores and they are simply summed to compute the score of a derivation. Since each probability could be regarded as a probability of a constituent given a sentence, I think they should be multiplied, or summed in log-domain, to represent the probability of deletion in a recursive manner. If summing is appropriate, I'd expect a justification for the approach.

- Experiment 2 carries out experiments on arbitrary constituent categories, but I'd expect more fine-grained analysis to see if there exist any different tendencies by the categories, e.g., NP, VP and PP. In particular, it is of interest in measure the impact of PP attachment.

- Penn-Treebank and Chinese Treebank are very popular data sources and have been used in many prior studies. I'm wondering if the data leakage might have an impact to this studies since the dataset, at least non-bracket texts, are found elsewhere mainly intended for language modeling. Also note that ChatGPT is capable to show Penn-Treebank like bracketing without any few-shot examples. It is better to measure how accurately GhatGPT can generate the bracket structure that is consistent with the gold annotation, and how that will be related to the accuracies of the deletion task.

Minor comment and suggestions:

- The first experiments comprises two sub-experiments, one for measuring the accuracies of a noun phrase deletion and the other for investigating the impact of parent constituency. It is not clear why these experiments are treated as a single experiment given that they have different theme.

- It is not clear whether an LSTM is a good baseline, and I think GPT-2 will be a better baseline since that follows Transformer architecture which is also used in ChatGPT. If possible, it is better to run an experiment on GPT-2 as an additional baseline to test the capacity.

- If possible, it is interesting to measure the impact of L2 for Experiment 3 to see if any biases in the tree structures.

- The current experiments are based on one-shot example, i.e., a single demonstration, for each prompt to an LLM. I think it is of interest to see if more shots, e.g., 3-shots, would significant impact the performance or not, in the same manner as observed in LSTM experiments.

- Constituency test has been investigated for unsupervised parsing and it seems to be marginally related: https://aclanthology.org/2020.emnlp-main.389/

Version 1:

Decision Letter:

6th February 2025

Dear Dr Ding,

Thank you for submitting your manuscript entitled "Active Use of Latent Tree-structured Sentence Representation in Humans and Large Language Models". We have given the paper our careful consideration and find it of potential interest. However, due to certain shortcomings we are concerned that sending the current manuscript out to review could lead to unnecessary delays and quite possibly an undesirable outcome of the review process.

In particular, it is journal's policy that we can accept one-tailed tests only when the work has been preregistered. Otherwise, we require two-tailed tests. Therefore, we would request that you repeat analyses using two tailed tests (e.g., Monte Carlo analysis). When submitting your revision, please include an updated cover letter, explaining whether there are any qualitative differences in the results when using two-tailed tests.

We would therefore like to invite you to revise your manuscript to address these concerns before we make a final determination on whether to send your manuscript for external review.

We shall hope to receive your revised version as soon as you are able to complete the suggested revisions. If something similar is published in the interim we will have to consider the impact it has on the novelty of a revised manuscript.

If you anticipate a delay of more than four weeks, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Human Behaviour or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date may be that of the revised, not the original, version.

Nature Human Behaviour is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as ‘corresponding author’ on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on ‘Modify my Springer Nature account’. For more information please visit please visit <a href="http://www.springernature.com/orcid">www.springernature.com/orcid</a>.

If you are not interested in submitting a suitably revised manuscript in the future please let me know immediately so we can close your file. If you have any questions, please contact me.

Please use the link below when you are prepared to resubmit.
Link Redacted

Thank you for your interest in Nature Human Behaviour.

Sincerely,

[redacted]
Nature Human Behaviour

Version 2:

Decision Letter:

23rd March 2025

Dear Dr Ding,

Thank you once again for your revised manuscript, entitled "Active Use of Latent Tree-structured Sentence Representation in Humans and Large Language Models," and for your patience during the re-review process.

Your manuscript has now been evaluated by the same reviewers who evaluated your original manuscript. All reviewer feedback is included at the end of this letter. Although the reviewers found your manuscript to have improved during revision, Reviewer #2 also raises some important outstanding concerns. We remain interested in the possibility of publishing your study in Nature Human Behaviour, but would like to consider your response to these outstanding concerns in the form of a revised manuscript before we make a decision on publication.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting,

please don't hesitate to contact me.

In sum, we invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We hope to receive your revised manuscript within 4-8 weeks. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

• Include a “Response to the editors and reviewers” document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors and reviewers to evaluate your revision.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

**Note:** This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Sincerely,

[redacted]
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: neurolinguistics, language constituents

Reviewer #2: LLMs, language representation, computational neuroscience

Reviewer #3: parsing algorithm, machine translation

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The authors have thoughtfully addressed my comments, thank you. I look forward to seeing this paper in print.

Reviewer #2 (Remarks to the Author):

I would like to begin by thanking the authors for the detailed investigation and follow-up analyses based on previous feedback. The paper is in a much stronger position now, and the results are more convincing as a consequence. Below, I outline a few remaining concerns and suggestions regarding the interpretation of the results.

2. Major Comments

2.1. Interpretation of Results in Figure 5
Lexical vs. syntactic information: The results in Figure 5 suggest that the absence of lexical information dramatically affects the “constituent rate” in models across both languages. This pattern, in my view, does not strongly support a primarily syntactic constituency-based deletion process. Indeed, the authors also note a similar pattern in Supplementary Figure S6 for Dataset 2. I would recommend adjusting the claim about the separability between lexical information and syntax; or if appropriate moving it to the supplementary section. The data as it stands suggest that lexical cues are intertwined with whatever syntactic pattern might be at play.

Data leakage discussion (Lines 533–535, 536–541, and 541–543).
The manuscript takes the results from Figure 5 as evidence against data leakage, uses Figure S6 to argue that LLMs are not relying on “only” syntactic information, and then concludes that these outcomes generalize to sentences not used for LLM training (and are not simply based on possible phrase annotations). However, it is unclear if these observations alone can justify the conclusion that there is no data leakage or that the behavior generalizes beyond the training set. The current evidence may not be sufficient to rule out other explanations (e.g., partial memorization of corpus, or alternative strategies models employ).

2.2. Differences Between Humans and LLMs
There are places where model behaviors clearly diverge from humans. For instance, in Supplementary Figure S2, LLM responses are sensitive to different prompts, whereas human behavior is relatively robust.
In addition, the text sometimes implies that if LLMs do not delete the word, they do not understand the task. However, an alternative view is that the LLM’s behavior reflects its inference over the instruction plus sentence modification example. This is not necessarily a failure to understand, but rather a different type of reasoning or strategy.
I think it would strengthen the paper if the authors could speculate further on why LLM behavior might differ from human behavior. For instance, do LLMs rely more on surface-level or distributional cues? Are they more literal or less flexible with instructions?

3. Minor Comments and Clarifications
1.Figure S4 (SVM Results for Remaining Strings). For Dataset 1 (English/ChatGPT), constituency, dependency connectivity, and MDD are equally good at discriminating the remaining string. For Chinese/ChatGPT and Chinese/humans, dependency connectivity seems to outperform constituencies. This is mostly true for Dataset 3 as well.
Please clarify in the text what you conclude from these differences. Why might dependency connectivity perform better in Chinese? How should we interpret “remaining string” results in the broader context of language processing?

2.Lines 62–78: Clarity of Argument. The paragraph is somewhat confusing regarding the argument about LLMs not explicitly encoding constituents or dependencies, and how that “raises the possibility” that these structures are not necessary for language understanding. Recommendation: Please clarify how one leads to the other. The absence of explicit encoding does not necessarily imply these constructs are unnecessary—it might only imply they are not represented in the same way as symbolic syntactic structures.

3.Line 367: CKY Reference. The statement about CKY is described as “not informative.” Please elaborate what CKY is.

4.Dependency Grammar vs. Constituency Grammar. The authors argue that the dependency grammar approach is still inferior to constituency-based analysis (Supplementary figure S4). Follow-up Question: If demonstration sentences were explicitly designed to illustrate dependency-based deletions, would the same advantages of constituent-based analysis hold?

Overall, the revised manuscript is much clearer and stronger than before. Nevertheless, additional clarifications on the points above would help solidify the central claims—particularly around the interplay of lexical vs. syntactic cues, the possibility of data leakage, and the differences between LLM and human deletion patterns. Moreover, addressing ambiguities in certain datasets (especially Dataset 5) would ensure that readers clearly understand how these results factor into the main conclusions.

I hope these comments and suggestions are helpful in refining the final version of the paper. Thank you again for the careful revisions and for engaging with the feedback so thoroughly.

Reviewer #3 (Remarks to the Author):

This work studies whether a large language model (LLM) can understand the concept of constituency of English and Chinese in the same manner as done by humans. Basic ideas is to ask them to delete a word sting by following an example of deleting a valid constituent, e.g., a noun phrase, to see if they can correctly identify a similar phrase in a sentence.

I had three major concerns for the initial manuscript:

- Justification for deletion probabilities for parsing.
- Fine-grained analysis for categories.
- Potential memorization issue.

The revised one addresses those concerns and also has made several updates for the suggested rewrites. I think this work is complete enough, and I have no further comments.

Version 3:

Decision Letter:

Our ref: NATHUMBEHAV-24051968C

28th April 2025

Dear Dr. Ding,

Thank you for submitting your revised manuscript "Active Use of Latent Tree-structured Sentence Representation in Humans and Large Language Models" (NATHUMBEHAV-24051968C).

After careful discussion, We are happy in principle to publish it in Nature Human Behaviour, pending minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within two weeks. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

[redacted]
Nature Human Behaviour

**EDITORIAL COMMENTS:**
In particular, we ask you to address the following (as well as all other reviewers' concerns):

**We would like to thank the editors and reviewers for their highly constructive feedback and thoughtful suggestions. We have addressed all the concerns and we believe the paper has been significantly strengthened. Additionally, since the revised manuscript is limited to 5000 words, some of the original analyses and new analyses are now moved into Supplementary Materials.**

**In the following, we first respond to the three general editorial comments and then respond to each question raised by the reviewers.**

1) improve the literature review, synthesise findings in the field, in particular from neuroscience of language. This is important to show the robustness of outcomes in response to various theories because you challenge the existing results.

**General Response 1:**
**The previous manuscript only discussed the constituency representation of sentences, which might have misled the readers to think that we used the results to argue for syntactic modularity and argue against other types of structural constructs such as dependency relations. Neither of these was our claims.**

**To avoid the potential confusion, we have extensively changed the Introduction and Discussion, as well as the title. The current study aims to investigate whether sentences are internally represented by a structured representation (either a constituency or dependency tree) or a linear representation (see Response 2.2). In the revised manuscript, we have employed multiple metrics derived from different representation theories, e.g., mean dependency distance (MDD), to explain the word deletion behavior (see Response 2.4). It is emphasized that the constituency and dependency representations can be isomorphic (Miller 1999) and a constituent generally corresponds to a cluster in the dependency representation (Klein and Manning 2004). See Fig. R1 for an example.**

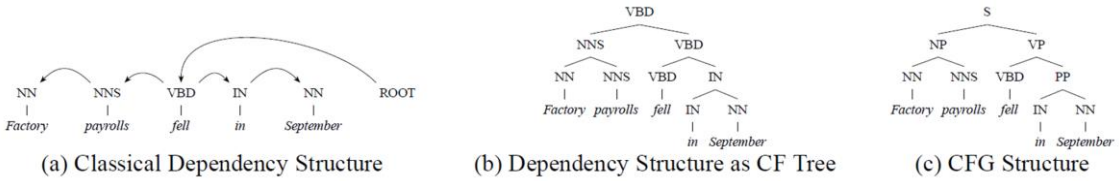


Figure R1. Isomorphic dependency and constituency representations (Klein and Manning 2004).

2) carry out additional analyses in response to methodological concerns raised by Reviewers 2 and 3, including running GPT-2 as baseline and multiple shots given the sensitivity of LLMs

**General Response 2:**
**Following the thoughtful suggestions, we have now used GPT-2 as the baseline model and have also tested multi-shot learning in LLMs. The GPT-2 results were similar to the previous LSTM results (see the new Figs. S5a & S5c), and multi-shot learning slightly improved the results (see the new Figs. S8 and S9). Additionally, a new analysis of the GPT-2 baseline model showed that the model consistently preferred the node-category rule for both Chinese and English even with 3000 training samples (see the new Figs. S5b & S5d), i.e., not showing human/ChatGPT-like language-dependent rule preference. See Response 3.5 and Response 3.7 for details.**

3) address the reviewers' concerns that the training data maybe including phrase annotations (Reviewers 1 and 3).

**General Response 3:**
**The reviewers raised an important concern about whether ChatGPT was trained on the treebank sentences used in our experiment or other data that contained phrase annotations. To address this concern, we have now applied three new analyses to demonstrate that the current results could generalize to sentences not involved in model training and did not reflect direct application of rules learned from syntactically annotated corpora.**

**One concern was that the test-set treebank sentences were included in the training set of ChatGPT. Therefore, ChatGPT might perform particularly well on the treebank sentences. To address this concern, we tested syntactically correct but meaningless sentences, e.g., "*The obvious summer lies on the retrieval*". These meaningless sentences were constructed for the experiment and not possibly used for ChatGPT**

**training. Nevertheless, ChatGPT still tended to delete a complete constituent (Fig. 5a, $p < 0.001$ for both Chinese and English, one-sided Monte Carlo test, false discovery rate (FDR) corrected). Language-dependent preference for node-/parent-category rule was also observed (Fig. 5b). Therefore, our conclusions based on treebank sentences could apply to sentences that ChatGPT had never seen. See Response 1.1 for details.**

**Another concern was whether ChatGPT learned syntactic rules directly from syntactically annotated corpora, and applied these rules to perform the word deletion task. To test this possibility, we analyzed whether the human/LLM responses were fully explained by syntactic rules extracted from the annotations in treebanks or grammar books or there were variabilities depending on lexical properties that were not explicitly labeled in treebanks or grammar books. We focused on the sentences that had a *[verb [preposition […]$_{NP}$]$_{PP}$]$_{VP}$* structure, and found that both humans and ChatGPT preferred to delete the NP (and therefore keep the verb-preposition combination intact) if the verb-preposition combination formed a stronger collocation, e.g., "*look at*" vs. "*happen on*". Such lexical-based structure variabilities were not present in treebank annotations. Therefore, the word deletion behavior could not be fully explained by just learning the syntactic annotations in treebanks. See Response 1.1 or Response 2.3 for details.**

**Furthermore, as suggested by the reviewers, we also asked ChatGPT to directly parse the syntactic structure of sentences. In terms of the F1 score with the tree annotated in the treebank, the explicitly parsed constituency tree was not strongly correlated with the constituency tree reconstructed from the word deletion task for individual sentences (Fig. S12c). Furthermore, human word deletion behavior was better explained by the constituency tree reconstructed based on ChatGPT word deletion responses than the constituency tree explicitly parsed by ChatGPT (Fig. S12b). These results suggested a dissociation between the ability to explicitly parse a sentence and the ability to implicitly extract constituents in the word deletion task. See Response 3.3 for details.**

**In summary, the three new analyses have demonstrated that (1) the current conclusions can generalize to novel sentences not used for ChatGPT training; (2) models do not just rely on syntactic rules extracted from annotations in treebanks to perform the word deletion task; and (3) model's performance on explicit sentence parsing dissociates from the performance on the word deletion task.**

**REVIEWER COMMENTS:**

**Reviewer #1 (Remarks to the Author):**

I really only have a clarification question. The authors note clearly that a central challenge is designing a test that avoids overlap with the training data. The chosen task is well adapted for this (e.g. it avoids using standard constituency-testing methods that may be explicitly present in the training data.) Still, I do wonder whether phrase annotations for English and Chinese may be present in the training data in some form. The training data used by ChatGPT and other models is not open for scrutiny, but scientific articles (including linguistics) make a major contribution to the MassiveText dataset that was used to train Gopher (sciencedirect is the top high-level URL contributing text; pubmed is in the top 3). If such annotations are present (e.g. the bigram "the remark" in Fig 1 or a bigram with a similar enough embedding is explicitly annotated as "NP"), then it seems plausible that such explicit information might nudge the LLM towards the observed behavior explicitly, rather than implicitly as argued. Is there any information about whether the training data contains linguistic analysis, or perhaps annotated corpora (even portions of the English and Chinese penn treebanks used in these studies)?

It may be that these are not answerable questions given the proprietary nature of some of the key LLMs tested here (though I would hope models that promote open access, like Llama, would make this informaiton available). The upshot is that it might be helpful to discuss what is and isn't known about the training data in terms of explicit syntactic annotations.

**Response 1.1.**

**We thank the reviewer for his/her positive and constructive feedback. We fully agree that data leakage is a potential problem that needs to be addressed. As is outlined in General Response 3, we have applied three new analyses to test whether the current results can be explained by data leakage. Here, we describe the first two analyses. The first analysis tested syntactically correct but meaningless sentences, and the second analysis analyzed whether the word deletion behavior was influenced by lexical effects not available in phrase annotations.**

**Analysis 1: Syntactically correct but meaningless sentences**

**We constructed syntactically correct but meaningless sentences based on the sentences in Experiment 1b (which is now referred to as Dataset 2), by replacing each content word by a randomly selected word of the same part of speech. Since the word-replaced sentences could be meaningful or ungrammatical, a volunteer was recruited to select 60 sentences that were meaningless but still have a clear syntactic structure for both Chinese and English (listed in SI). For the meaningless sentences, the constituent rate was still significantly higher than the nonconstituent rate (Fig. 5a, $p < 0.001$ for both Chinese and English, paired two-sided bootstrap, FDR corrected). However, since ChatGPT often refused to respond to meaningless sentences, the ratio of "other" response sharply increased (see Table S2 for details). Since the meaningless sentences were not possibly included in the training data for ChatGPT, the results confirmed that the preference to delete a constituent generalized to the sentences that ChatGPT had never observed.**

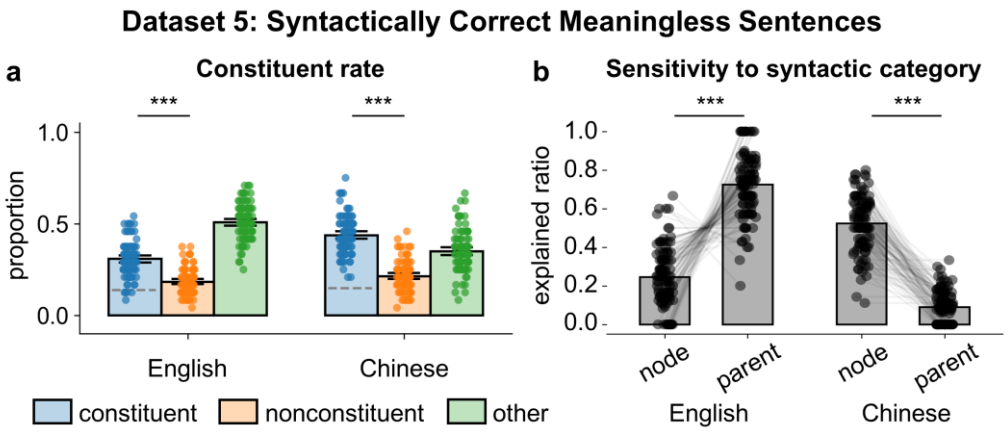


Figure 5. Results on syntactically correct meaningless sentences that are constructed based on the parallel sentences in Dataset 2 (see Methods). **a**. Constituent rate of word strings deleted by ChatGPT. The constituent rate is significantly higher than the nonconstituent rate. The ratio of "other" response also increases compared with other dataset since ChatGPT often refuses to respond to meaningless sentences (see Table S2 for details). **b.** The explained ratio for the node- and parent-category rules, which is consistent with the results on Dataset 2. ***$p < 0.001$.

**Analysis 2: Lexical influence on word deletion behavior**

**We analyzed whether there were fine-grained patterns in the human/ChatGPT responses that could not be fully explained by syntactic rules directly extracted from syntactically annotated corpora. In Dataset 2, all test sentences had the same syntactic structure, i.e., *[verb [preposition […]$_{NP}$]$_{PP}$]$_{VP}$*, and participants preferred to delete the PP for English and preferred to delete the NP for Chinese. A detailed analysis of the data, however, revealed more complex patterns.**

**We found the chance of deleting a NP consistently varied across sentences. For human participants, the mean rate of deleting a NP ranged from 0.0 to 0.333 in English and 0.292 to 0.905 in Chinese. Importantly, the chance of deleting a NP was correlated between humans and ChatGPT across sentences (Fig. S6a), suggesting that the variation was not random. Therefore, the deletion behavior might rely on fine-grained lexical properties, instead of just part of speech information that was consistent across all the sentences. For the English sentences, we further analyzed whether the chance of deleting a NP depended on whether the verb and the preposition formed a stronger collocation based on the Oxford Collocations Dictionary (Oxford University Press 1991). We defined collocation strength in the following way – 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. The collocation strength significantly correlated with how often the human participants deleted a NP (Fig. S6b). These analyses have now been added to Supplementary Results. The analysis was not applied to Chinese due to the lack of collocation dictionaries.**

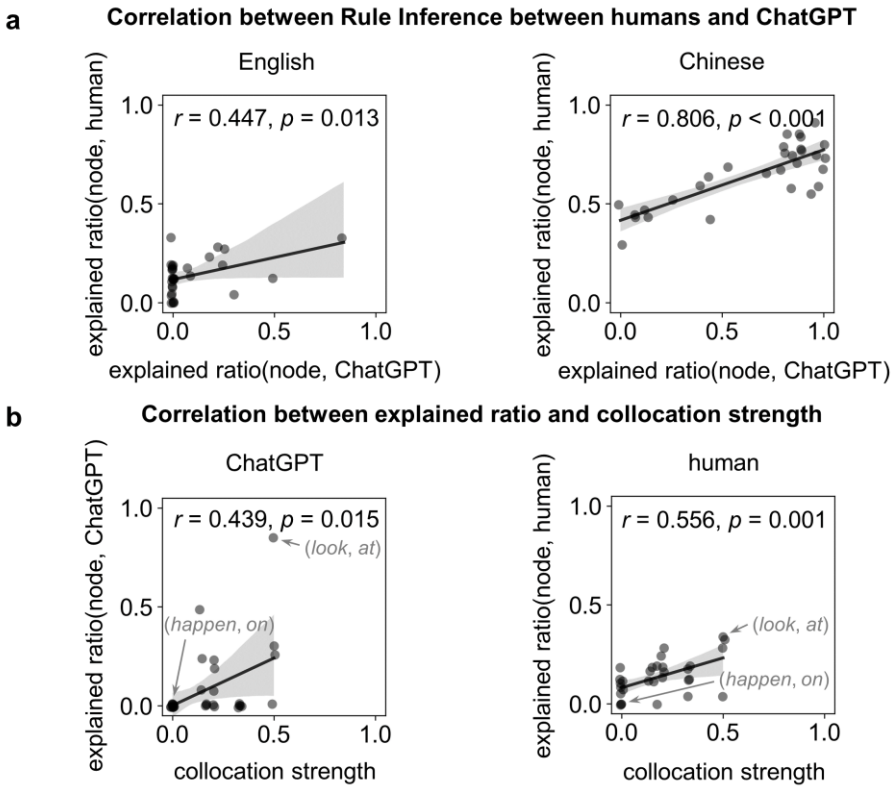


Figure S6. Analysis of lexical effects on word deletion task in Dataset 2. **a.** The explained ratio for the node-category rule is correlated between humans and ChatGPT. Each dot represents a sentence (*N* = 30). **b.** The explained ratio for the node-category rule is correlated with the strength of collocation between verb and preposition. Collocation strength is extracted based on the Oxford Collocations Dictionary 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.

**Reviewer #2 (Remarks to the Author):**

My main critique of the paper focuses on how the behavioral responses to the task are evaluated. In the design of the experimental materials for Experiment 1, the only condition considered is deletion based on constituency. The authors show in Figure 1C that human behavior aligns with deletion on constituents. However, there is a large variability in both the constituent and non-constituent conditions. According to the methods, non-constituents are defined as anything not identified as a NP constituent by the constituency tree (Chomsky, 1957). This is not a strong alternative hypothesis, as it ignores other formalizations of sentence structure and does not provide a robust null hypothesis.

**We thank the reviewer for his/her very constructive feedback. We explained the response variability in Response 2.1 and the construction of stronger alternative hypotheses in Responses 2.2 and 2.4. A clarification was that in the analysis of Experiment 1 (Dataset 1 in the revised version), we did not restrict the analysis to NPs. Instead, as the results showed, English speakers preferred to delete a child node of the VP, whether the node was an NP or not. However, the reviewer was correct that the target word string generally fell into a narrow syntactic class.**

**Response 2.1: Individual Variability**
**The large individual variability may arise from at least two sources. First, the task was an inference task that was intrinsically ambiguous – Many rules could possibly explain the demonstration. The study aimed to test whether participants preferred some rules, but individual variability was still expected. The instruction did not even mention to delete words and, in about 17.2% of the trials in Dataset 1, the participants did not delete words – Instead, they added words, reorder words, or did not change the test sentence (Table S1a). In Dataset 1, only in 7.3% of trials, the participants deleted a multiword word string that was not a constituent (Table S1a). Even when the participants inferred that they should delete words, many rules remained possible to explain the demonstration, some of which were now illustrated in Fig. 1b. A new analysis suggested by the reviewer (see Response 2.4) further showed that a number of metrics could influence the word deletion task.**

**Second, human participants and ChatGPT may have internal sentence representations differing from the constituency annotation in the**

**treebanks (see Response 2.3 for a case analysis), especially for sentences with a complicated structure. For example, if the word string deleted in a demonstration was not perceived as a constituent, the participant or ChatGPT would fail to infer that the rule was to delete a constituent. Similarly, if their internal sentence representation of the test sentence differed from the treebank annotation, they might delete what they perceived as a constituent (or, equivalently, a cluster in the dependency representation), which might not be a constituent in the treebank annotation. When the sentences had very simple and clear syntactic structure, e.g., the artificially constructed sentences in Dataset 2, the constituent rate of humans was as high as 0.99 (see *Language-dependent Rule Preference and Lexical Influence (Dataset 2)*).**

**In response to the insightful comment, we have (1) added a new table (Tables S1 and S2) that breaks up "nonconstituent" and "other" responses into finer-grained categories, (2) adapted Fig. 1 to further illustrate that the task was intrinsically ambiguous, and (3) added a Supplementary Discussion section entitled *Variability on the word deletion task* (see below).**

**"*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 representation8 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.*"**

1. Several questions arise: Are participants not behaving in a linguistically informed way or instead their behavior could be explained better by an alternative account, say Dependency length minimization?

**Response 2.2**
**Thank you for the very insightful comment. We apologize that the previous manuscript only discussed the constituency representation. In fact, we view constituency and dependency representations as the same category of representations in the current context (see Fig. R1 for an example), in contrast to the representation based on linear word order. The dependency representation, however, can indeed derive more diverse metrics, e.g., the mean dependency distance (MDD), which are now used to quantify the word deletion behavior (see Response 2.4 for details).**

**In the revised manuscript, we considered both constituency and dependency representations, and have changed "constituency representation" in the title to "tree-structured representation". We emphasized the commonality between constituency and dependency representations and contrasted them with a linear representation of word sequences. For example, in the first paragraph of the Introduction, we now mentioned:**

**"*The basic forms of dependency and constituency representations are isomorphic[5] and a constituent generally corresponds to a cluster in the dependency representation[6]. Such structured representations are in contrast to a representation based on the linear ordering of words, such as the input to transformer-based language models[7].*"**

For trials that they don't not follow the constituency, is there a consistency in their response? Are there specific sentences that elicit non-constituency behavior across participants?

**Response 2.3**
**Thank you for the very insightful comment. In Datasets 1 and 3, each participant was tested with a different set of sentences to increase**

**sentence diversity, and therefore we could not analyze whether the response consistency across participants. In Dataset 2, however, each parallel sentence was tested on about 24 participants, allowing the analyses suggested by the reviewer. In Dataset 2, all test sentences had the same syntactic structure, i.e., *[verb [preposition […]$_{NP}$]$_{PP}$]$_{VP}$*, and participants preferred to delete the PP for English and preferred to delete the NP for Chinese. A detailed analysis of the data, however, revealed more fine-grained patterns in the human/ChatGPT responses that could not be fully explained by syntactic rules that can be extracted from syntactically annotated corpora.**

**We found the chance of deleting a NP consistently varied across sentences. For human participants, the mean rate of deleting a NP ranged from 0.0 to 0.333 in English and 0.292 to 0.905 in Chinese. Importantly, the chance of deleting a NP was correlated between humans and ChatGPT across sentences (Fig. S6a), suggesting that the variation was not random. Therefore, the deletion behavior might rely on fine-grained lexical properties, instead of just part of speech information that was consistent across all the sentences. For the English sentences, we further analyzed whether the chance of deleting a NP depended on whether the verb and the preposition formed a stronger collocation based on the Oxford Collocations Dictionary (Oxford University Press 1991). We defined collocation strength in the following way – 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. The collocation strength significantly correlated with how often the human participants deleted a NP (Fig. S6b). In other words, when the verb and preposition formed strong collocation, the participants might perceive it as a chunk, in contrast to what classic syntactic theories predict. These analyses have now been added to Supplementary Results.**

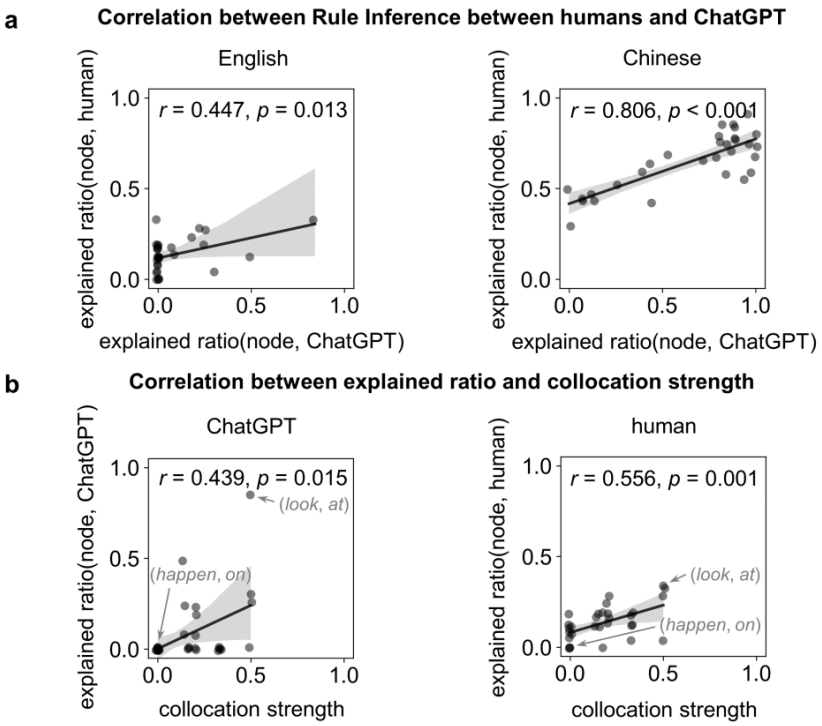


Figure S6. Analysis of lexical effects on word deletion task in Dataset 2. **a.** The explained ratio for the node-category rule is correlated between humans and ChatGPT. Each dot represents a sentence (*N* = 30). **b.** The explained ratio for the node-category rule is correlated with the strength of 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.

A more principled approach to characterizing human behavior would involve developing a set of theories based on constituency and other linguistic or null theories that aim to capture human behavior across all trials and show that overall constituency explained behavioral patterns better.

**Response 2.4**

**Thank you for the insightful comments and valuable suggestions. Following the suggestions, we derived a number of metrics based on the tree-structured (i.e., constituency or dependency representation) and linear sentence representations, and tested which metric could best explain the word deletion behavior.**

**Specifically, we tested three metrics derived from a tree-structured representation, i.e.,**
**(1) constituency;**
**(2) whether the word string formed a connected graph in the dependency representation;**
**(3) mean dependency distance (MDD);**
**and two metrics derived from a linear representation, i.e.,**
**(4) string length, number of words in the string;**
**(5) onset position, number of words prior to the string.**
**Furthermore, each metric was calculated based on both the deleted word string and the remaining word string, resulting in 10 metrics per test.**

**We employed these metrics to discriminate word strings deleted by humans/ChatGPT and randomly selected word strings, based on an SVM classifier. The results (Fig. S4) showed that constituency of the deleted word string was the metric achieving highest discrimination rate, except for Chinese ChatGPT experiment (in which the three metrics derived from tree-structure representations, i.e., constituency, dependency connectivity, and MDD, reached similar performance). The discrimination analysis has been added to Supplementary Results due to the word limit.**

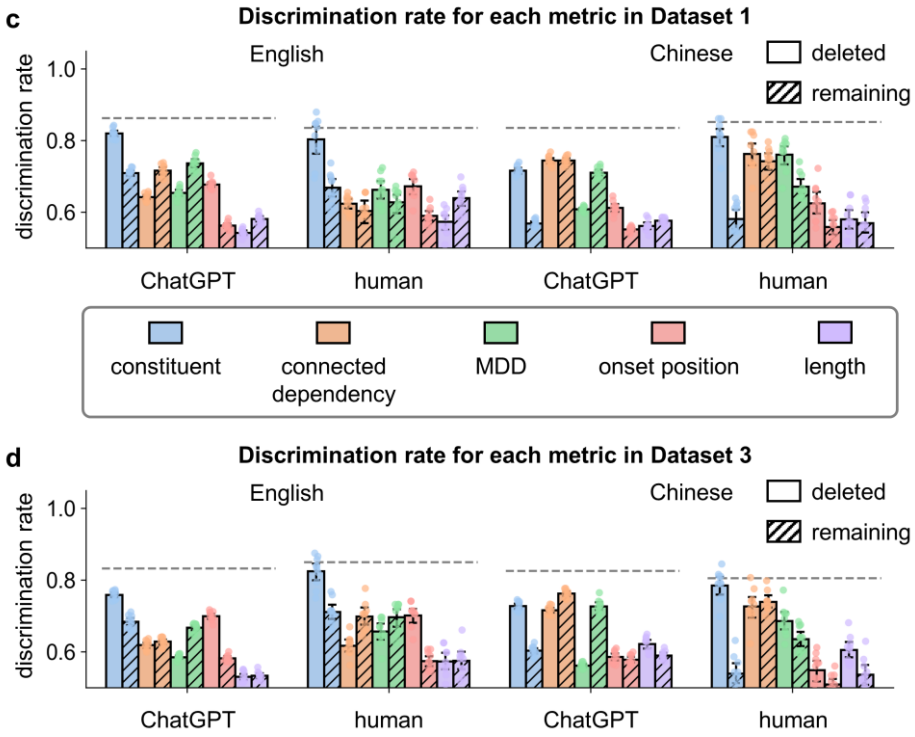


Figure S4. ... **cd.** Discrimination of real human/ChatGPT responses and random word strings based on different word string metrics. Three metrics are

derived from a tree-based representation, i.e., constituency, dependency connectivity, and MDD, and two metrics are derived from a linear representation. Each bar shows the discrimination based on one metric. The gray dashed line represents the discrimination rate when all metrics are jointly considered. The classifier is an SVM.

2. This points to a more general issue with the paper's presentation and how it synthesizes existing results in the field to motivate the research. Work in the neuroscience of language, specifically (Fedorenko et al., 2016, 2024), has challenged the notion of syntax as a core driver or separate module in language processing as evidenced by activity in the the brain regions responsible for language processing, (see also Goodman, 1997; Linebarger et al., 1983).

**Response 2.5**
**Thank you for the insightful comment. In fact, we did not claim that syntax was a separate module and the word "module" never appeared in the manuscript. Since structural analysis of a sentence may integrate syntactic, semantic, and prosodic information, we analyzed how lexical properties (see Response 2.3) and semantic plausibility (Dataset 6) influence the word deletion behavior.**

**To avoid confusion, we have rewritten many parts of the paper. For example, when discussing the lexical influence on behavior, we mentioned that "*detailed word properties, not just the part of speech, could influence the participants' word deletion behavior*". In Introduction, we mentioned:**

"***sentences across languages show tendency to reduce the mean dependency distance[8,9]. Both behavioral[10,11] and brain responses[12-16] are sensitive to the onset and offset of linguistic constituents. Details of the structured representations, however, remain debated, e.g., the specific representation format and granularity of constituents[17-19].***"

3. Additionally, the authors need to clarify why they consider their task a form of active language use. Why should human behavior in a task that is meta-linguistic and not representative of everyday language use be taken as evidence that they actively use constituency to represent language?

**Response 2.6**

**Thank you for the important question. We claimed that the participants actively used a tree-structured sentence representation when solving a rule inference task. The actively used representation is a representation that can directly influence behavior and can be inferred from behavior. It is in contrast to a representation that can be decoded from either brain activity or model hidden layers (Brennan et al. 2012; Hewitt and Manning 2019) but is in fact not actively accessed by the brain or LLM when solving a task. For example, previous studies have found that the dependency tree of a sentence could be decoded based on the word embeddings in a LLM, but it was concerned that such a representation was artificially decoded and were not actively read out by deeper-layer neurons in the model (Belinkov 2022; van Dijk et al. 2023). Here, we demonstrated that the tree-structured representation could be directly inferred from behavior.**

**To avoid confusion, we have (1) changed "*active language use*" to "*actively contribute to behavior*"; (2) changed "*analyzing language-use behavior*" to "*analyzing rule inferencing behavior*"; (3) explicitly discussed that our deletion task demonstrated "active use" of a tree-structured representation in an inference task while the probe studies demonstrated the tree-structured representation in "everyday" language use. Therefore, these two approaches are complementary: "*Many previous studies have also demonstrated that an internal constituency representation emerges in language models or in the human brain by directly analyzing the internal representation of models, i.e., activation of hidden-layer neurons, or human neural recordings…. the analysis of internal representation is highly complementary to the analysis of word deletion behavior, since the former can be achieved during arbitrary language processing task while the latter can probe whether a constituency representation can actively contribute to behavior.*".**

Other comments

The LSTM baseline is not particularly informative, as it primarily controls for results based purely on learning. The fact that it only requires about 50 demonstrations to learn the task raises questions about the level of linguistic knowledge required to perform the task. For example, if the subjects identify that the task is about grammar and assume they have received some grammar training during their education, they could solve the task by simply referring to their acquired knowledge of

grammar as taught in school. This approach contrasts with how a preschool child learns and uses language actively.

**Response 2.7**

**Thank you for the insightful comment. The concern was that the participants might start to use school grammar to solve the task after a few trials, just like how the baseline LSTM model learned to perform the task. To address this concern, we first analyzed whether the constituent rate significantly increased after a few trials, but this phenomenon was not observed (Fig. S1). Why did the LSTM model, but not the human participants and ChatGPT, show a strong learning effect? The difference lay in how the models were trained – The former LSTM model and the current GPT-2 model were trained in a supervised manner using the expected answer to each question, while the expected answer was not provided during human/ChatGPT experiments. We have now emphasized the difference in training: "*we also trained a baseline model, i.e., an initialized GPT-2 model, to learn the word deletion task based on the ordinal position and lexical properties of each word. When the expected answer was provided to allow supervised learning, such a model still needed hundreds of samples to reach the constituent rate of humans or ChatGPT (see Supplementary Results and Fig. S5).*".**

In the last paragraph before the section "Sensitivity to Syntactic Category," the authors argue that the phenomena cannot be explained by statistical cues in word properties. However, I am uncertain how this conclusion is drawn from the LSTM results.

**Response 2.8**

**We apologize for the unclear statement. The LSTM had access to word properties through pretrained word vectors, but failed to explain the word deletion behavior with just a few demonstrations. Therefore, we meant to conclude that the properties of individual words were not sufficient to explain word deletion behavior – For example, if the participants intended to delete words that were semantically related to the words deleted in the demonstration, such a trend would have been easily captured by the LSTM. The LSTM analysis had been replaced with the GPT-2 analysis and the sentence has been removed.**

If the demonstration did not follow a constituency rule, the authors' notion suggests that humans would still be biased toward using constituency-based parsing to modify the test. I believe that demonstrating this bias would enhance the results by providing additional evidence of

humans' reliance on constituency structures, even in non-standard scenarios.

**Response 2.9**

**Thank you for the insightful question. However, we do not assume that humans or ChatGPT will delete a constituent if what is deleted in the demonstration is not a constituent. In fact, we assume the opposite – As is discussed in the paper, we assume the default strategy of both humans and ChatGPT is to output a grammatical and meaningful word string, which is generally a constituent. Therefore, participants who do not carefully examine the demonstration may tend to type in a constituent as a default tendency for language production. To avoid that this tendency is taken as evidence for a tree-structured sentence representation, we choose to delete a constitute and leave a word string that is generally not a constituent – We hypothesize that if the participants cannot discover a clear pattern in the remaining word string, they start to analyze the deleted word string. We have made these points clear in the Discussion. For example, "*In the demonstration, a grammatical and meaningful sentence was transformed into a word string that was generally not grammatical or meaningful. Therefore, the participants had to avoid their natural tendency to produce grammatical and meaningful expressions and infer the uncommon rule underlying the transformation.*".**

**To directly answer the question here, if demonstration keeps a constituent and removes a nonconstituent, we hypothesize that the participants will learn to keep a constituent, since it is consistent with their default strategy.**

Another point of concern is the sensitivity of large language models (LLMs) to prompts. If this is a one-shot learning task that does not depend on context beyond the demonstration sentence, why does Figure S2 show that the prompt can substantially affect the model's behavior (as seen with prompt 2, prompt 3, and prompt 4)? This observation echoes the earlier comment regarding the non-constituent and other responses that the models provide. This is in contrast to humans, who show consistency in their response patterns (Figure S3).

**Response 2.10**

**The reviewer correctly pointed out that the LLM results partly depended on the prompt – Different prompts do not affect the general conclusion but have quantitative influences on the results. Reliance on the prompt,**

**however, is commonly observed in LLMs (Sclar et al. 2024). For example, by just adding "Let's think step by step" to a prompt, the LLM performance on an inference task can boost by about 40% (Kojima et al. 2022) and different LLMs prefer different prompts in different tasks (Chang et al. 2024). In this experiment, the difference between prompts was quantitative instead of qualitative and all prompts supported the same conclusion – The constituent rate was higher than the nonconstituent rate and the preference for node-/parent-category rule differs between languages.**

The authors need to do more work in both the introduction and discussion sections to highlight theories and experimental results that challenge the original constituency parsing hypothesis. The first paragraph of the introduction lacks a comprehensive review of the various approaches that exist in the field to address the main question. While the authors refer to the constituency as an influential hypothesis supported by research in psychology and neuroscience, they fail to cover alternative accounts that challenge this hypothesis. Empirical work by researchers such as Fedorenko, among others, highlights significant observations that contradict the dominance of unique syntactic parsing for explaining activity in regions in the human brain that process language, for a review see (Fedorenko et al., 2024).

**Response 2.11**

**We fully agree that we should not only discuss the constituency parsing hypothesis, and we have thoroughly changed the title, Introduction, and Discussion. We actually agree with Fedorenko et al. that there may not be a specific syntactic region in the brain, but we did not include detailed discussions on this topic since the paper does not attempt to address the neural substrate for language processing. To address this concern, we have (1) rewritten most part of the Introduction; (2) changed the discussion section title from "Syntactic capacity of LLMs" to "Sentence representation in LLMs"; (3) briefly mentioned that "*In the human brain, a distributed language network contributes to sentence representation*[73-75]*".***

In figure 5C, it is unclear to me why the points depicting the results are in bands. Can authors show the results per-subject basis, to see whether the effect is present in every subject.

**Response 2.12**

**We apologize for the confusion. In fact, data points in Figs. 5cd (Figs. S13cd in the revised manuscript) were the results from individual participants. We have now clarified this in the figure caption: "*… Each data point represents the result from a participant or a run of ChatGPT.*". Since each participant (or run of ChatGPT) only received 6 tests in each condition, the proportion of target string could only take 7 possible values, which was why the data points concentrate on 7 values.**

In figure 5C the baseline of chatGPT proportions is much lower compared to other figures, why is this the case?

**Response 2.13**

**Fig. 5c (Fig. S13c in the revised version) shows the proportion of the target strings, instead of the total constituent rate. ChatGPT often deleted constituents that were not the target strings. In this experiment, the constituent rate was 0.66 ± 0.013 (95% CI) for ChatGPT (see Supplementary Results).**

**Reviewer #3 (Remarks to the Author):**

- The tree reconstruction using CKY parsing is very interesting but the current setting sounds not apt to me. Given Figure 3, the probabilities associated with all strings in a derivation are treated as scores and they are simply summed to compute the score of a derivation. Since each probability could be regarded as a probability of a constituent given a sentence, I think they should be multiplied, or summed in log-domain, to represent the probability of deletion in a recursive manner. If summing is appropriate, I'd expect a justification for the approach.

**We thank the reviewer for his/her very constructive feedback.**

**Response 3.1**

**We agree with the reviewer that it is more reasonable to sum log probability and we have now analyzed how different operations, i.e., linear summation, linear multiplication, log summation, and log multiplication, influenced tree reconstruction. All operations, except for log multiplication, yielded similar results (Fig. S10). In the main text, we chose to report the result for linear summation since the linearly summed probability can be interpreted as the ratio of tests explained by a tree. However, we mentioned that "*In the example here, the explained ratio of a tree was the sum of explained ratio of its nodes. Since the sum of likelihood was often calculated on the log scale, we also tried log-scale integration of explained ratios and obtained similar results (see Fig. S10).*".**

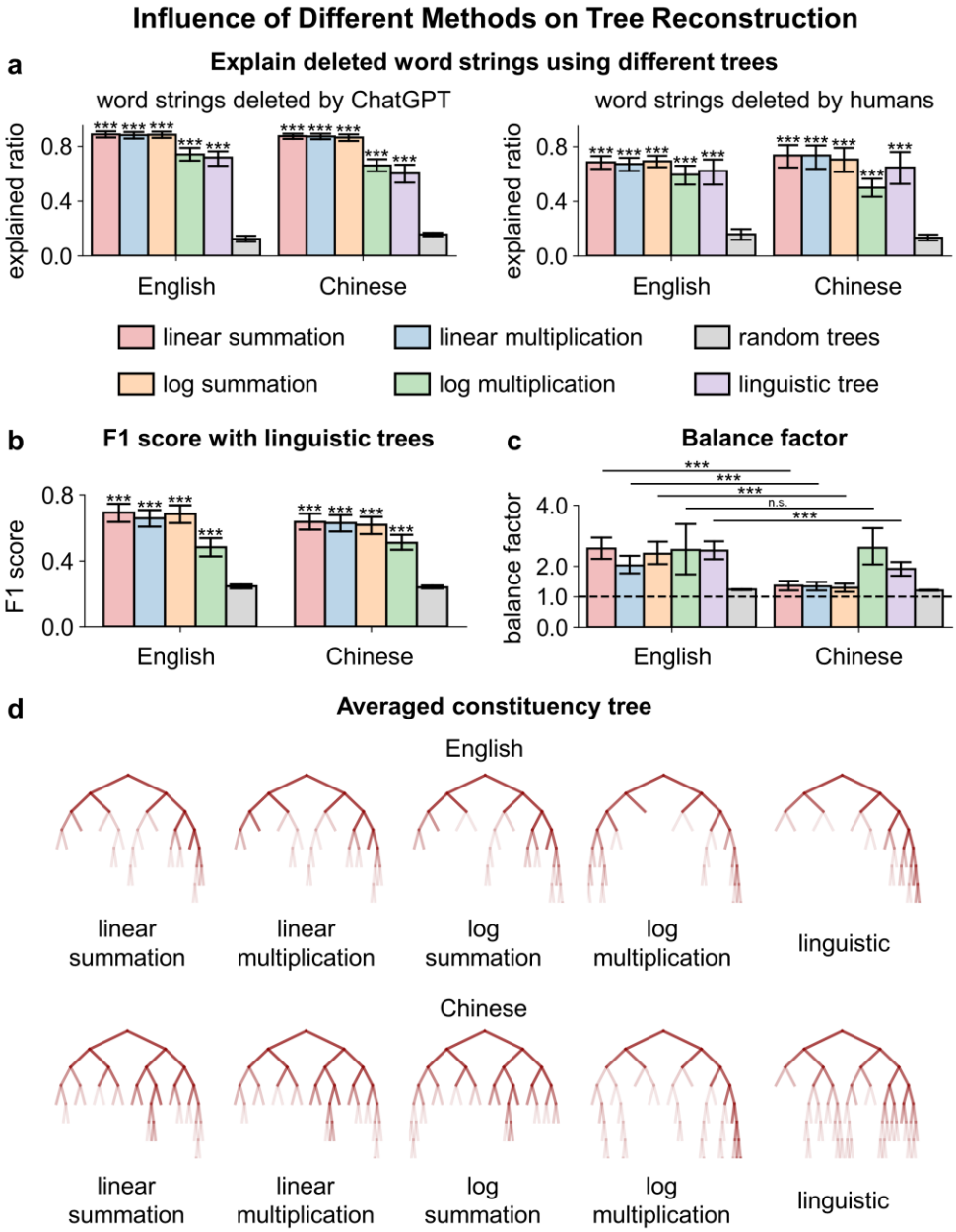


Figure S10. Constituency tree reconstructed using 4 different methods differing in how the explained ratio is integrated, i.e., linear summation (reported in the main text), linear multiplication, log summation and log multiplication. The results are generally consistent except for log multiplication. **a**. Explained ratios for different trees. **b**. F1 score with respect to the linguistic tree. **c**. Balance factor of each constituency tree. **d**. Visualization of the constituency trees averaged over all test sentences. ***$p < 0.001$.

- Experiment 2 carries out experiments on arbitrary constituent categories, but I'd expect more fine-grained analysis to see if there exist any different tendencies by the categories, e.g., NP, VP and PP. In particular, it is of interest in measure the impact of PP attachment.

**Response 3.2**

**Thank you for the highly constructive suggestion. We have now separately analyzed the constituent rate and explained ratio for different constituent categories in Experiment 2 (Dataset 3 in the revised version). The results were reported in Table S3 for constituent categories with more than 10 samples in the human experiment. In general, the results are consistent across constituent categories – Participants preferred to delete constituents, and the responses to English and Chinese sentences are separately better explained by parent- and node-category rules. There are, however, a few exceptions – For ChatGPT, the constituent rate was below 0.3 for the NP[PP] structure in English and ChatGPT deleted multiple strings in 41.1% trials (Table S3). An example of the NP[PP] category demonstration was "*Some found it on the screen ~~of a personal computer~~*". ChatGPT might infer that the task was to shorten the sentence. An example ChatGPT response was "*Competition in ~~the~~ sale ~~of complete bikes is~~ heating up too*", which captured the gist of the sentence but did not delete a constituent.**

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 |

- Penn-Treebank and Chinese Treebank are very popular data sources and have been used in many prior studies. I'm wondering if the data leakage might have an impact to this studies since the dataset, at least non-bracket texts, are found elsewhere mainly intended for language modeling. Also note that ChatGPT is capable to show Penn-

Treebank like bracketing without any few-shot examples. It is better to measure how accurately GhatGPT can generate the bracket structure that is consistent with the gold annotation, and how that will be related to the accuracies of the deletion task.

**Response 3.3**

**Following the constructive suggestion, we have asked ChatGPT to explicitly parse the sentence structure, using the prompt described in a recent study (Tian et al. 2024). The explicitly parsed constituency tree reached F1 scores comparable to that of the deletion-based constituency tree (Fig. S12a). However, the explicitly parsed constituency tree differed from the deletion-based constituency tree, and performed significantly worse at explaining the word deletion behavior (Fig. S12b). Furthermore, for individual sentences, the F1 score of explicitly parsed trees was not strongly correlated with the F1 score of deletion-based trees (Fig. S12c), suggesting that the capacity to explicitly parse a sentence dissociated from the capacity to implicitly extract linguistic constituents in the word deletion task.**

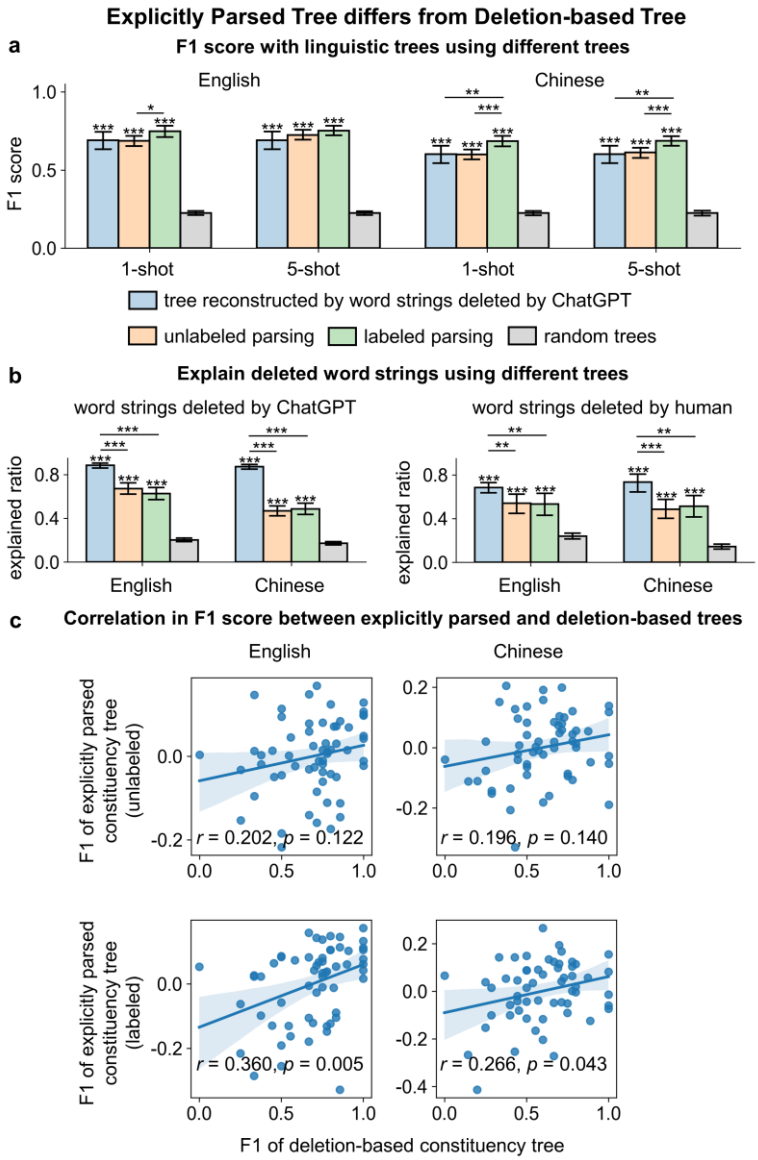


Figure S12. Comparison between syntactic tree explicitly parsed by ChatGPT and deletion-based tree. Explicit parsing is achieved using prompts[2] using either 1-shot or 5-shots learning. **a**. F1 score with respect to the linguistic tree. One-shot and 5-shot parsing obtain similar performance and we 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. **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.

**Please also see Response 1.1 for other analyses to address the data leakage problem. The comparison between the explicitly parsed trees and deleted-based trees have been added to Fig. S12.**

Minor comment and suggestions:

- The first experiments comprises two sub-experiments, one for measuring the accuracies of a noun phrase deletion and the other for investigating the impact of parent constituency. It is not clear why these experiments are treated as a single experiment given that they have different theme.

**Response 3.4**
**Experiments in the previous manuscript referred to behavioral testing on humans/ChatGPT, but we agree it could be misleading. Therefore, the data collected during Experiment 1 is now referred to Dataset 1 and there are two separate analyses on the same dataset.**

- It is not clear whether an LSTM is a good baseline, and I think GPT-2 will be a better baseline since that follows Transformer architecture which is also used in ChatGPT. If possible, it is better to run an experiment on GPT-2 as an additional baseline to test the capacity.

**Response 3.5**
**Following the valuable suggestion, we have now used GPT-2 to construct baseline models. With a single demonstration, the constituent rate of GPT-2 was not significantly above chance. In Dataset 1 (Figs. S5a & S5b), after trained with about 100-500 demonstration-test pairs, the GPT-2 model reached the performance of ChatGPT and its performance further increased with more training samples. However, GPT-2 consistently preferred the node-category rule for both Chinese and English, not showing the language-dependent preference as humans and ChatGPT. A similar pattern is observed for Dataset 3 (Figs. S5c & S5d). The results of GPT-2 have been added to Fig. S5.**

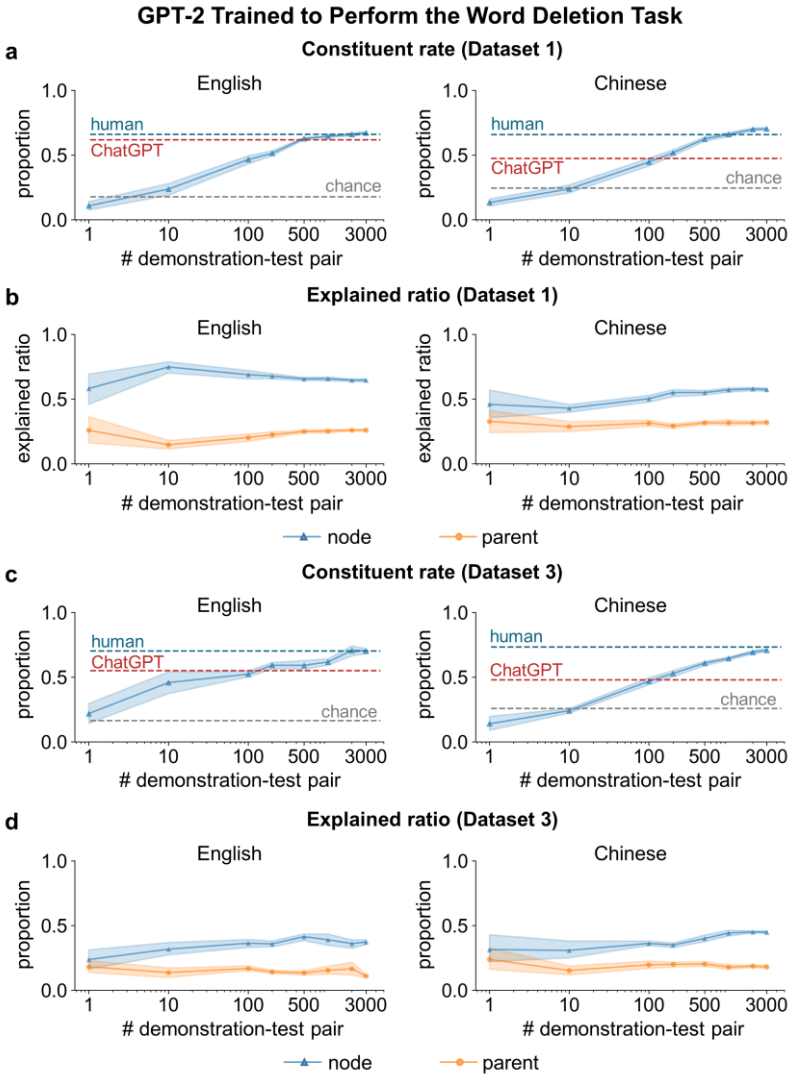


Figure S5. Performance of a baseline model, i.e., a naive GPT-2 model, trained to perform the word deletion task. The naive GPT-2 model is not pretrained, and its input includes word position embeddings and pretrained word embeddings. **ac.** Constituent rate of the GPT-2 model as a function of the number 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.

- If possible, it is interesting to measure the impact of L2 for Experiment 3 to see if any biases in the tree structures.

**Response 3.6**

**Following the valuable suggestion, we have now tested L2 speakers on Experiment 3 (Dataset 4 in the revised version), following the same**

**procedure for native speakers. The results are generally consistent between L2 and native speakers. The deletion-based constituency tree reconstructed by L2 speakers exhibited a slightly higher F1 score than ChatGPT and native speakers (Fig. S11b), and explained about 80% of the strings deleted in the L2 experiment (Fig. S11a).**

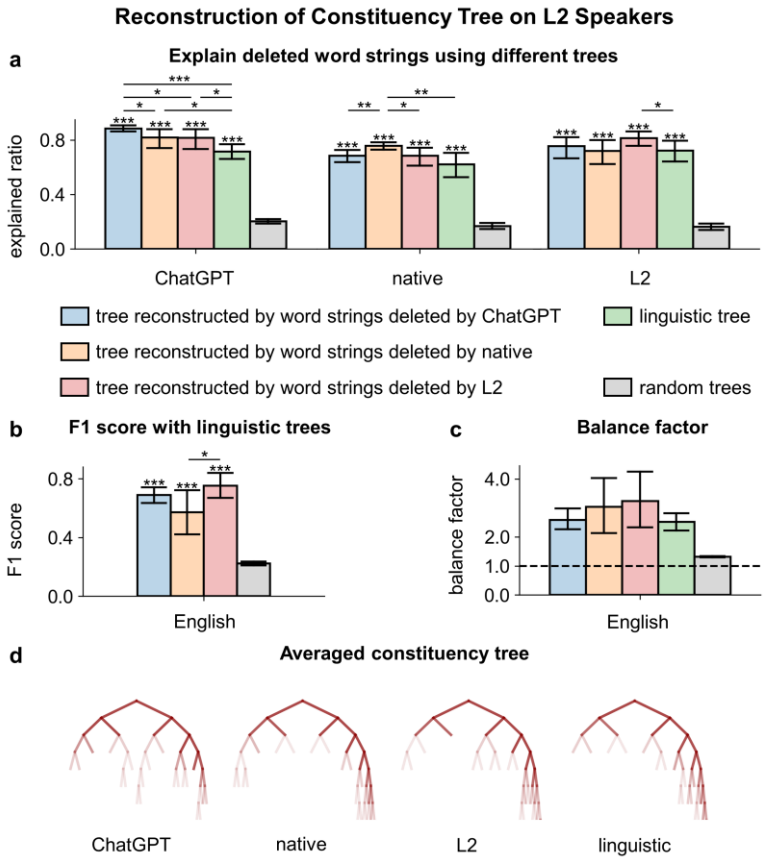


Figure S11. Reconstruction of constituency tree based on word deletion behavior for L2 English speakers ($N$ = 34, 19-29 years old, mean 23 years old; 22 females), following the same procedure for native speakers in Dataset 4. The results are generally consistent to ChatGPT and native speakers. **a**. 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. S11), significantly higher than chance ($p < 0.001$, one-sided Monte Carlo test, FDR corrected). **b**. F1 score with respect to the linguistic tree. **c**. Balance factor of each constituency tree. **d**. Visualization of the constituency trees averaged over all test sentences. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

- The current experiments are based on one-shot example, i.e., a single demonstration, for each prompt to an LLM. I think it is of interest to see if more shots, e.g., 3-shots, would significant impact the performance or not, in the same manner as observed in LSTM experiments.

**Response 3.7**

**Following the valuable suggestion, we have now tested N-shot learning on ChatGPT (N ≤ 5). The constituent rate slightly improved with more demonstrations (Dataset 1: from 0.473 to 0.515 for Chinese and 0.618 to 0.780 for English. Dataset 3: from 0.479 to 0.513 for Chinese and 0.550 to 0.702 for English), while the explained ratio of the node- and parent-category rules was not clearly influenced by the number of shots. The results of N-shot learning have been added to Figs. S8 and S9.**

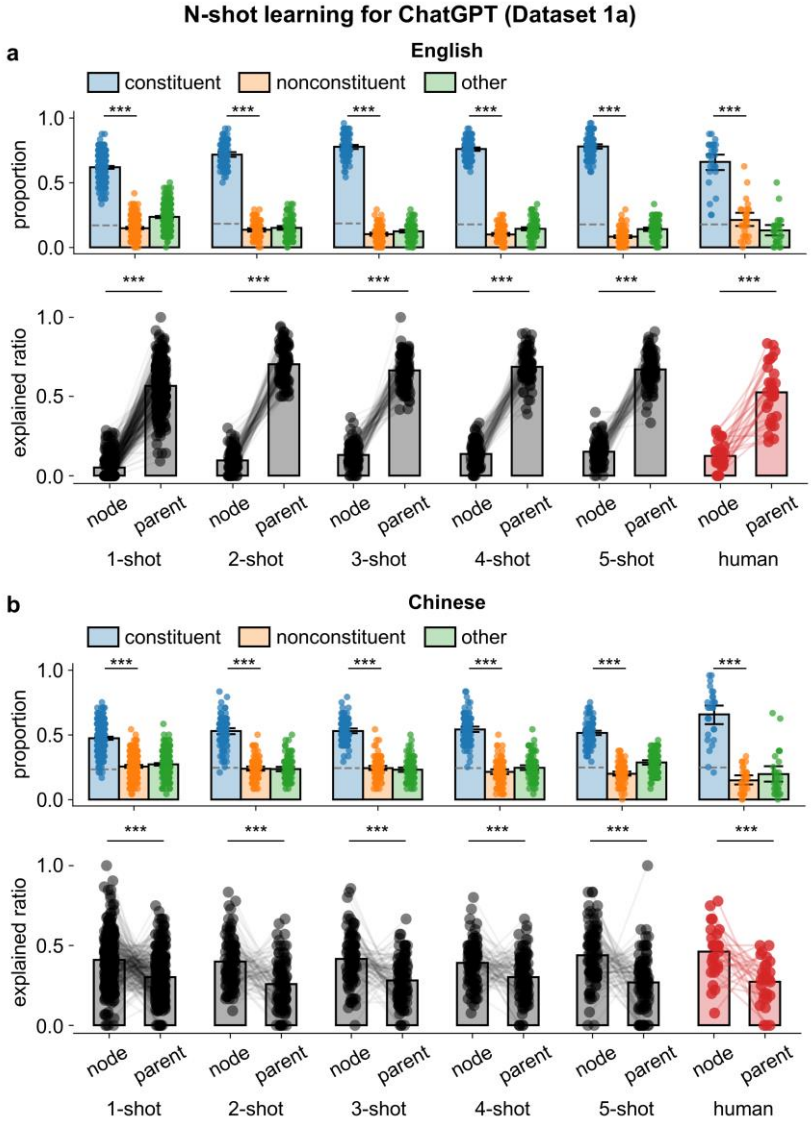


Figure S8. ChatGPT N-shot learning for Dataset 1. **a.** Constitute rate slightly improves with more demonstrations. **b.** The explained ratio for the node- and parent-category rules do not clearly change with more demonstrations. ***$p$ < 0.001.

- Constituency test has been investigated for unsupervised parsing and it seems to be marginally related: https://aclanthology.org/2020.emnlp-main.389/

**Response 3.8**

**Thank you for pointing out this highly relevant paper and we have now cited it – "*Based on linguistic constituency tests, algorithms have been designed to parse the constituency structure of a sentence in an unsupervised manner*[53]."**.

**Reference**


Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and language*, *120*(2), 163–173.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., et al. (2024). A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, *15*(3). https://doi.org/10.1145/3641289

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=e2TBb5y0yFf

Oxford University Press. (1991). *Oxford Collocation Dictionary (Bilingual version)*. Foreign Language Teaching and Research Press Pub.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=RIu5lyNXjT

**EDITORIAL COMMENTS:**
In particular, we ask you to address the following (as well as all other reviewers' concerns):

**We would like to thank the editors and reviewers for their highly constructive feedback and thoughtful suggestions. We have addressed all the concerns and we believe the paper has been significantly strengthened. Additionally, since the revised manuscript is limited to 5000 words, some of the original analyses and new analyses are now moved into Supplementary Materials.**

**In the following, we first respond to the three general editorial comments and then respond to each question raised by the reviewers.**

1) improve the literature review, synthesise findings in the field, in particular from neuroscience of language. This is important to show the robustness of outcomes in response to various theories because you challenge the existing results.

**General Response 1:**
**The previous manuscript only discussed the constituency representation of sentences, which might have misled the readers to think that we used the results to argue for syntactic modularity and argue against other types of structural constructs such as dependency relations. Neither of these was our claims.**

**To avoid the potential confusion, we have extensively changed the Introduction and Discussion, as well as the title. The current study aims to investigate whether sentences are internally represented by a structured representation (either a constituency or dependency tree) or a linear representation (see Response 2.2). In the revised manuscript, we have employed multiple metrics derived from different representation theories, e.g., mean dependency distance (MDD), to explain the word deletion behavior (see Response 2.4). It is emphasized that the constituency and dependency representations can be isomorphic (Miller 1999) and a constituent generally corresponds to a cluster in the dependency representation (Klein and Manning 2004). See Fig. R1 for an example.**

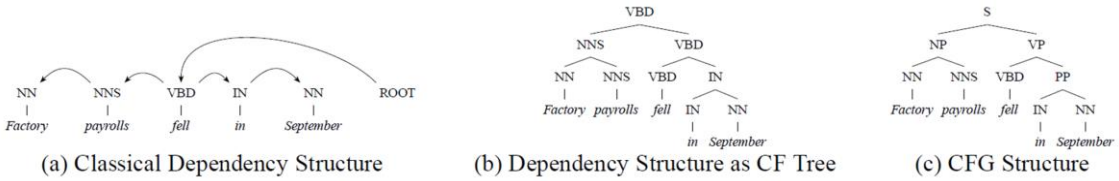


Figure R1. Isomorphic dependency and constituency representations (Klein and Manning 2004).

2) carry out additional analyses in response to methodological concerns raised by Reviewers 2 and 3, including running GPT-2 as baseline and multiple shots given the sensitivity of LLMs

**General Response 2:**
**Following the thoughtful suggestions, we have now used GPT-2 as the baseline model and have also tested multi-shot learning in LLMs. The GPT-2 results were similar to the previous LSTM results (see the new Figs. S5a & S5c), and multi-shot learning slightly improved the results (see the new Figs. S8 and S9). Additionally, a new analysis of the GPT-2 baseline model showed that the model consistently preferred the node-category rule for both Chinese and English even with 3000 training samples (see the new Figs. S5b & S5d), i.e., not showing human/ChatGPT-like language-dependent rule preference. See Response 3.5 and Response 3.7 for details.**

3) address the reviewers' concerns that the training data maybe including phrase annotations (Reviewers 1 and 3).

**General Response 3:**
**The reviewers raised an important concern about whether ChatGPT was trained on the treebank sentences used in our experiment or other data that contained phrase annotations. To address this concern, we have now applied three new analyses to demonstrate that the current results could generalize to sentences not involved in model training and did not reflect direct application of rules learned from syntactically annotated corpora.**

**One concern was that the test-set treebank sentences were included in the training set of ChatGPT. Therefore, ChatGPT might perform particularly well on the treebank sentences. To address this concern, we tested syntactically correct but meaningless sentences, e.g., "*The obvious summer lies on the retrieval*". These meaningless sentences were constructed for the experiment and not possibly used for ChatGPT**

**training. Nevertheless, ChatGPT still tended to delete a complete constituent (Fig. 5a, $p$ = 0.002 for both Chinese and English, two-sided Monte Carlo test, false discovery rate (FDR) corrected). Language-dependent preference for node-/parent-category rule was also observed (Fig. 5b). Therefore, our conclusions based on treebank sentences could apply to sentences that ChatGPT had never seen. See Response 1.1 for details.**

**Another concern was whether ChatGPT learned syntactic rules directly from syntactically annotated corpora, and applied these rules to perform the word deletion task. To test this possibility, we analyzed whether the human/LLM responses were fully explained by syntactic rules extracted from the annotations in treebanks or grammar books or there were variabilities depending on lexical properties that were not explicitly labeled in treebanks or grammar books. We focused on the sentences that had a *[verb [preposition […]$_{NP}$]$_{PP}$]$_{VP}$* structure, and found that both humans and ChatGPT preferred to delete the NP (and therefore keep the verb-preposition combination intact) if the verb-preposition combination formed a stronger collocation, e.g., "*look at*" vs. "*happen on*". Such lexical-based structure variabilities were not present in treebank annotations. Therefore, the word deletion behavior could not be fully explained by just learning the syntactic annotations in treebanks. See Response 1.1 or Response 2.3 for details.**

**Furthermore, as suggested by the reviewers, we also asked ChatGPT to directly parse the syntactic structure of sentences. In terms of the F1 score with the tree annotated in the treebank, the explicitly parsed constituency tree was not strongly correlated with the constituency tree reconstructed from the word deletion task for individual sentences (Fig. S12c). Furthermore, human word deletion behavior was better explained by the constituency tree reconstructed based on ChatGPT word deletion responses than the constituency tree explicitly parsed by ChatGPT (Fig. S12b). These results suggested a dissociation between the ability to explicitly parse a sentence and the ability to implicitly extract constituents in the word deletion task. See Response 3.3 for details.**

**In summary, the three new analyses have demonstrated that (1) the current conclusions can generalize to novel sentences not used for ChatGPT training; (2) models do not just rely on syntactic rules extracted from annotations in treebanks to perform the word deletion task; and (3) model's performance on explicit sentence parsing dissociates from the performance on the word deletion task.**

**REVIEWER COMMENTS:**

**Reviewer #1 (Remarks to the Author):**

I really only have a clarification question. The authors note clearly that a central challenge is designing a test that avoids overlap with the training data. The chosen task is well adapted for this (e.g. it avoids using standard constituency-testing methods that may be explicitly present in the training data.) Still, I do wonder whether phrase annotations for English and Chinese may be present in the training data in some form. The training data used by ChatGPT and other models is not open for scrutiny, but scientific articles (including linguistics) make a major contribution to the MassiveText dataset that was used to train Gopher (sciencedirect is the top high-level URL contributing text; pubmed is in the top 3). If such annotations are present (e.g. the bigram "the remark" in Fig 1 or a bigram with a similar enough embedding is explicitly annotated as "NP"), then it seems plausible that such explicit information might nudge the LLM towards the observed behavior explicitly, rather than implicitly as argued. Is there any information about whether the training data contains linguistic analysis, or perhaps annotated corpora (even portions of the English and Chinese penn treebanks used in these studies)?

It may be that these are not answerable questions given the proprietary nature of some of the key LLMs tested here (though I would hope models that promote open access, like Llama, would make this informaiton available). The upshot is that it might be helpful to discuss what is and isn't known about the training data in terms of explicit syntactic annotations.

**Response 1.1.**

**We thank the reviewer for his/her positive and constructive feedback. We fully agree that data leakage is a potential problem that needs to be addressed. As is outlined in General Response 3, we have applied three new analyses to test whether the current results can be explained by data leakage. Here, we describe the first two analyses. The first analysis tested syntactically correct but meaningless sentences, and the second analysis analyzed whether the word deletion behavior was influenced by lexical effects not available in phrase annotations.**

**Analysis 1: Syntactically correct but meaningless sentences**

**We constructed syntactically correct but meaningless sentences based on the sentences in Experiment 1b (which is now referred to as Dataset 2), by replacing each content word by a randomly selected word of the same part of speech. Since the word-replaced sentences could be meaningful or ungrammatical, a volunteer was recruited to select 60 sentences that were meaningless but still have a clear syntactic structure for both Chinese and English (listed in SI). For the meaningless sentences, the constituent rate was still significantly higher than the nonconstituent rate (Fig. 5a, $p < 0.001$ for both Chinese and English, paired two-sided bootstrap, FDR corrected). However, since ChatGPT often refused to respond to meaningless sentences, the ratio of "other" response sharply increased (see Table S2 for details). Since the meaningless sentences were not possibly included in the training data for ChatGPT, the results confirmed that the preference to delete a constituent generalized to the sentences that ChatGPT had never observed.**

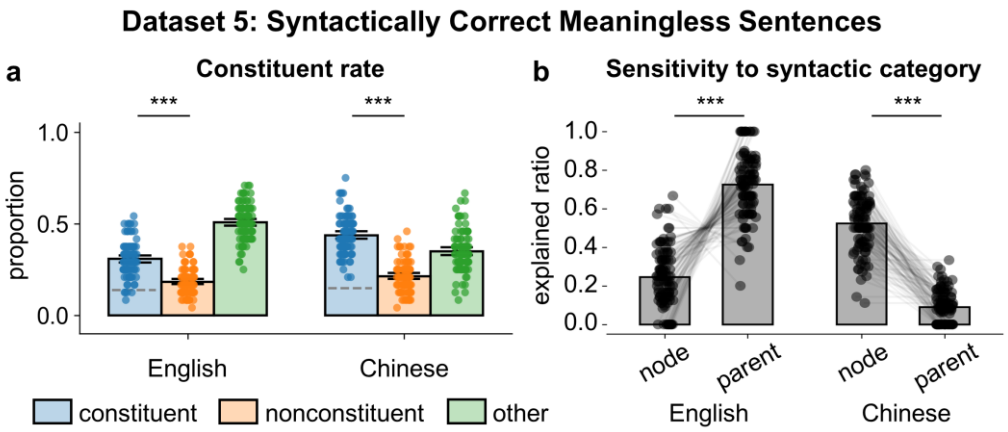


Figure 5. Results on syntactically correct meaningless sentences that are constructed based on the parallel sentences in Dataset 2 (see Methods). **a**. Constituent rate of word strings deleted by ChatGPT. The constituent rate is significantly higher than the nonconstituent rate. The ratio of "other" response also increases compared with other dataset since ChatGPT often refuses to respond to meaningless sentences (see Table S2 for details). **b.** The explained ratio for the node- and parent-category rules, which is consistent with the results on Dataset 2. ***$p < 0.001$.

**Analysis 2: Lexical influence on word deletion behavior**

**We analyzed whether there were fine-grained patterns in the human/ChatGPT responses that could not be fully explained by syntactic rules directly extracted from syntactically annotated corpora. In Dataset 2, all test sentences had the same syntactic structure, i.e., *[verb [preposition […]$_{NP}$]$_{PP}$]$_{VP}$*, and participants preferred to delete the PP for English and preferred to delete the NP for Chinese. A detailed analysis of the data, however, revealed more complex patterns.**

**We found the chance of deleting a NP consistently varied across sentences. For human participants, the mean rate of deleting a NP ranged from 0.0 to 0.333 in English and 0.292 to 0.905 in Chinese. Importantly, the chance of deleting a NP was correlated between humans and ChatGPT across sentences (Fig. S6a), suggesting that the variation was not random. Therefore, the deletion behavior might rely on fine-grained lexical properties, instead of just part of speech information that was consistent across all the sentences. For the English sentences, we further analyzed whether the chance of deleting a NP depended on whether the verb and the preposition formed a stronger collocation based on the Oxford Collocations Dictionary (Oxford University Press 1991). We defined collocation strength in the following way – 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. The collocation strength significantly correlated with how often the human participants deleted a NP (Fig. S6b). These analyses have now been added to Supplementary Results. The analysis was not applied to Chinese due to the lack of collocation dictionaries.**

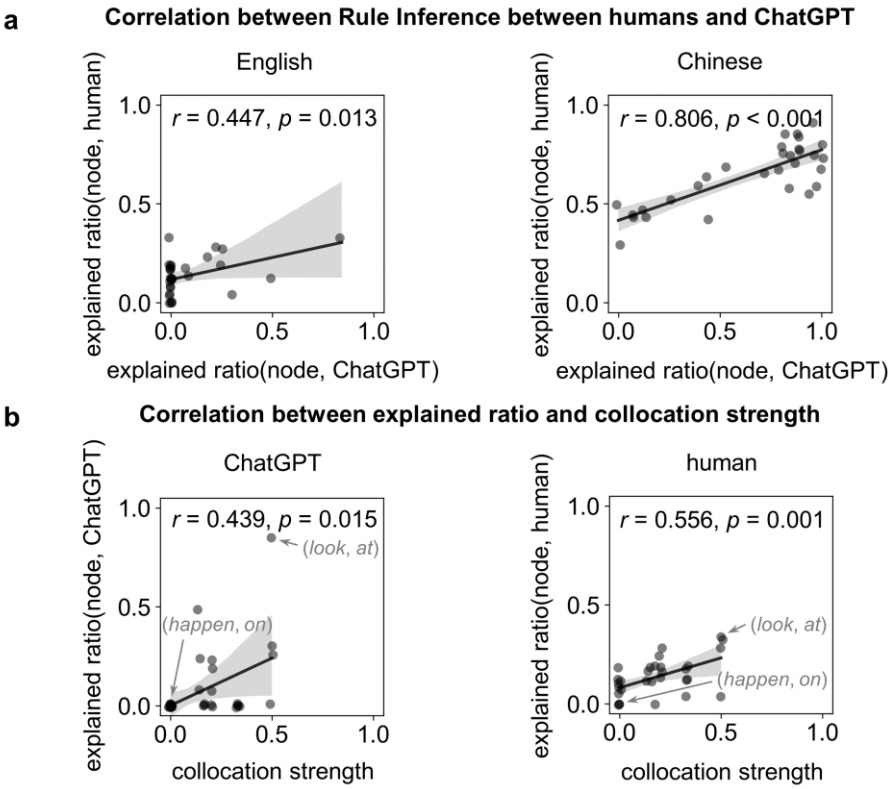


Figure S6. Analysis of lexical effects on word deletion task in Dataset 2. **a.** The explained ratio for the node-category rule is correlated between humans and ChatGPT. Each dot represents a sentence (*N* = 30). **b.** The explained ratio for the node-category rule is correlated with the strength of 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.

**Reviewer #2 (Remarks to the Author):**

My main critique of the paper focuses on how the behavioral responses to the task are evaluated. In the design of the experimental materials for Experiment 1, the only condition considered is deletion based on constituency. The authors show in Figure 1C that human behavior aligns with deletion on constituents. However, there is a large variability in both the constituent and non-constituent conditions. According to the methods, non-constituents are defined as anything not identified as a NP constituent by the constituency tree (Chomsky, 1957). This is not a strong alternative hypothesis, as it ignores other formalizations of sentence structure and does not provide a robust null hypothesis.

**We thank the reviewer for his/her very constructive feedback. We explained the response variability in Response 2.1 and the construction of stronger alternative hypotheses in Responses 2.2 and 2.4. A clarification was that in the analysis of Experiment 1 (Dataset 1 in the revised version), we did not restrict the analysis to NPs. Instead, as the results showed, English speakers preferred to delete a child node of the VP, whether the node was an NP or not. However, the reviewer was correct that the target word string generally fell into a narrow syntactic class.**

**Response 2.1: Individual Variability**
**The large individual variability may arise from at least two sources. First, the task was an inference task that was intrinsically ambiguous – Many rules could possibly explain the demonstration. The study aimed to test whether participants preferred some rules, but individual variability was still expected. The instruction did not even mention to delete words and, in about 17.2% of the trials in Dataset 1, the participants did not delete words – Instead, they added words, reorder words, or did not change the test sentence (Table S1a). In Dataset 1, only in 7.3% of trials, the participants deleted a multiword word string that was not a constituent (Table S1a). Even when the participants inferred that they should delete words, many rules remained possible to explain the demonstration, some of which were now illustrated in Fig. 1b. A new analysis suggested by the reviewer (see Response 2.4) further showed that a number of metrics could influence the word deletion task.**

**Second, human participants and ChatGPT may have internal sentence representations differing from the constituency annotation in the**

**treebanks (see Response 2.3 for a case analysis), especially for sentences with a complicated structure. For example, if the word string deleted in a demonstration was not perceived as a constituent, the participant or ChatGPT would fail to infer that the rule was to delete a constituent. Similarly, if their internal sentence representation of the test sentence differed from the treebank annotation, they might delete what they perceived as a constituent (or, equivalently, a cluster in the dependency representation), which might not be a constituent in the treebank annotation. When the sentences had very simple and clear syntactic structure, e.g., the artificially constructed sentences in Dataset 2, the constituent rate of humans was as high as 0.99 (see *Language-dependent Rule Preference and Lexical Influence (Dataset 2)*).**

**In response to the insightful comment, we have (1) added a new table (Tables S1 and S2) that breaks up "nonconstituent" and "other" responses into finer-grained categories, (2) adapted Fig. 1 to further illustrate that the task was intrinsically ambiguous, and (3) added a Supplementary Discussion section entitled *Variability on the word deletion task* (see below).**

**"*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 representation8 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.*"**

1. Several questions arise: Are participants not behaving in a linguistically informed way or instead their behavior could be explained better by an alternative account, say Dependency length minimization?

**Response 2.2**
**Thank you for the very insightful comment. We apologize that the previous manuscript only discussed the constituency representation. In fact, we view constituency and dependency representations as the same category of representations in the current context (see Fig. R1 for an example), in contrast to the representation based on linear word order. The dependency representation, however, can indeed derive more diverse metrics, e.g., the mean dependency distance (MDD), which are now used to quantify the word deletion behavior (see Response 2.4 for details).**

**In the revised manuscript, we considered both constituency and dependency representations, and have changed "constituency representation" in the title to "tree-structured representation". We emphasized the commonality between constituency and dependency representations and contrasted them with a linear representation of word sequences. For example, in the first paragraph of the Introduction, we now mentioned:**

**"*The basic forms of dependency and constituency representations are isomorphic[5] and a constituent generally corresponds to a cluster in the dependency representation[6]. Such structured representations are in contrast to a representation based on the linear ordering of words, such as the input to transformer-based language models[7].*"**

For trials that they don't not follow the constituency, is there a consistency in their response? Are there specific sentences that elicit non-constituency behavior across participants?

**Response 2.3**
**Thank you for the very insightful comment. In Datasets 1 and 3, each participant was tested with a different set of sentences to increase**

**sentence diversity, and therefore we could not analyze whether the response consistency across participants. In Dataset 2, however, each parallel sentence was tested on about 24 participants, allowing the analyses suggested by the reviewer. In Dataset 2, all test sentences had the same syntactic structure, i.e., *[verb [preposition […]$_{NP}$]$_{PP}$]$_{VP}$*, and participants preferred to delete the PP for English and preferred to delete the NP for Chinese. A detailed analysis of the data, however, revealed more fine-grained patterns in the human/ChatGPT responses that could not be fully explained by syntactic rules that can be extracted from syntactically annotated corpora.**

**We found the chance of deleting a NP consistently varied across sentences. For human participants, the mean rate of deleting a NP ranged from 0.0 to 0.333 in English and 0.292 to 0.905 in Chinese. Importantly, the chance of deleting a NP was correlated between humans and ChatGPT across sentences (Fig. S6a), suggesting that the variation was not random. Therefore, the deletion behavior might rely on fine-grained lexical properties, instead of just part of speech information that was consistent across all the sentences. For the English sentences, we further analyzed whether the chance of deleting a NP depended on whether the verb and the preposition formed a stronger collocation based on the Oxford Collocations Dictionary (Oxford University Press 1991). We defined collocation strength in the following way – 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. The collocation strength significantly correlated with how often the human participants deleted a NP (Fig. S6b). In other words, when the verb and preposition formed strong collocation, the participants might perceive it as a chunk, in contrast to what classic syntactic theories predict. These analyses have now been added to Supplementary Results.**

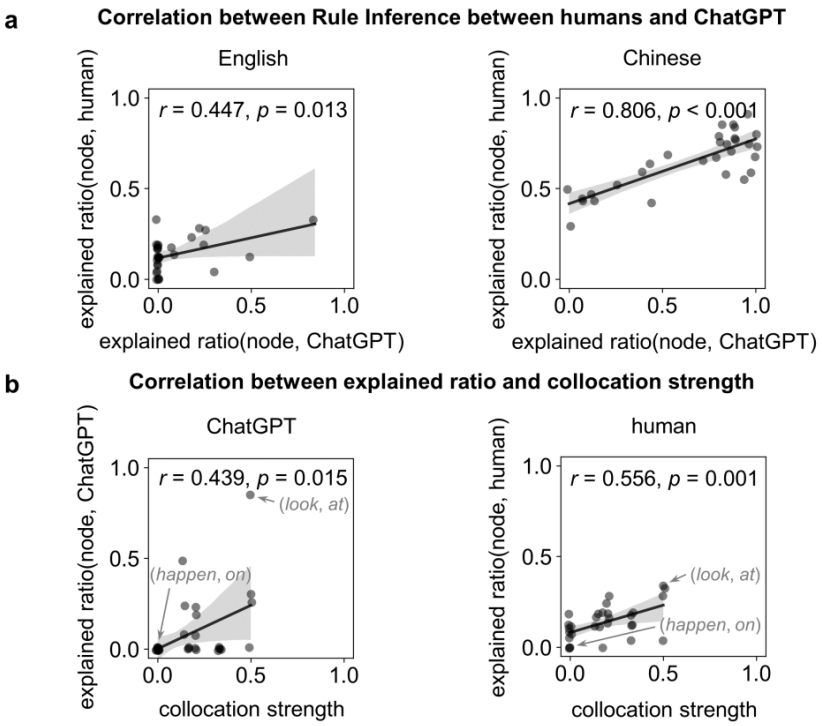


Figure S6. Analysis of lexical effects on word deletion task in Dataset 2. **a.** The explained ratio for the node-category rule is correlated between humans and ChatGPT. Each dot represents a sentence (*N* = 30). **b.** The explained ratio for the node-category rule is correlated with the strength of 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.

A more principled approach to characterizing human behavior would involve developing a set of theories based on constituency and other linguistic or null theories that aim to capture human behavior across all trials and show that overall constituency explained behavioral patterns better.

**Response 2.4**

**Thank you for the insightful comments and valuable suggestions. Following the suggestions, we derived a number of metrics based on the tree-structured (i.e., constituency or dependency representation) and linear sentence representations, and tested which metric could best explain the word deletion behavior.**

**Specifically, we tested three metrics derived from a tree-structured representation, i.e.,**
**(1) constituency;**
**(2) whether the word string formed a connected graph in the dependency representation;**
**(3) mean dependency distance (MDD);**
**and two metrics derived from a linear representation, i.e.,**
**(4) string length, number of words in the string;**
**(5) onset position, number of words prior to the string.**
**Furthermore, each metric was calculated based on both the deleted word string and the remaining word string, resulting in 10 metrics per test.**

**We employed these metrics to discriminate word strings deleted by humans/ChatGPT and randomly selected word strings, based on an SVM classifier. The results (Fig. S4) showed that constituency of the deleted word string was the metric achieving highest discrimination rate, except for Chinese ChatGPT experiment (in which the three metrics derived from tree-structure representations, i.e., constituency, dependency connectivity, and MDD, reached similar performance). The discrimination analysis has been added to Supplementary Results due to the word limit.**

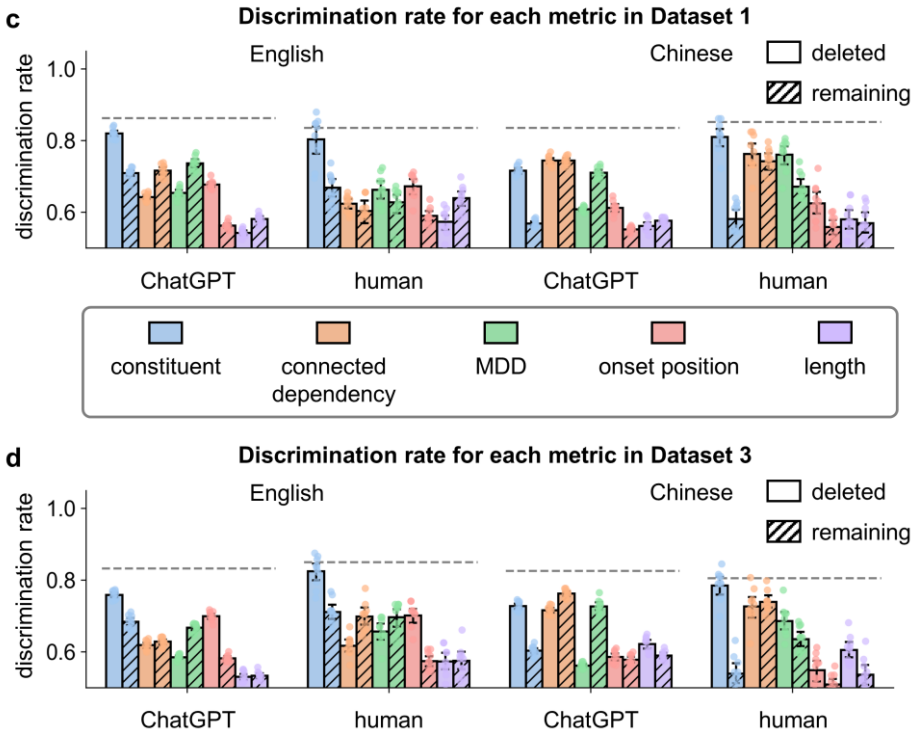


Figure S4. ... **cd.** Discrimination of real human/ChatGPT responses and random word strings based on different word string metrics. Three metrics are

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

2. This points to a more general issue with the paper's presentation and how it synthesizes existing results in the field to motivate the research. Work in the neuroscience of language, specifically (Fedorenko et al., 2016, 2024), has challenged the notion of syntax as a core driver or separate module in language processing as evidenced by activity in the the brain regions responsible for language processing, (see also Goodman, 1997; Linebarger et al., 1983).

**Response 2.5**
**Thank you for the insightful comment. In fact, we did not claim that syntax was a separate module and the word "module" never appeared in the manuscript. Since structural analysis of a sentence may integrate syntactic, semantic, and prosodic information, we analyzed how lexical properties (see Response 2.3) and semantic plausibility (Dataset 6) influence the word deletion behavior.**

**To avoid confusion, we have rewritten many parts of the paper. For example, when discussing the lexical influence on behavior, we mentioned that "*detailed word properties, not just the part of speech, could influence the participants' word deletion behavior*". In Introduction, we mentioned:**

"***sentences across languages show tendency to reduce the mean dependency distance[8,9]. Both behavioral[10,11] and brain responses[12-16] are sensitive to the onset and offset of linguistic constituents. Details of the structured representations, however, remain debated, e.g., the specific representation format and granularity of constituents[17-19].***"

3. Additionally, the authors need to clarify why they consider their task a form of active language use. Why should human behavior in a task that is meta-linguistic and not representative of everyday language use be taken as evidence that they actively use constituency to represent language?

**Response 2.6**

**Thank you for the important question. We claimed that the participants actively used a tree-structured sentence representation when solving a rule inference task. The actively used representation is a representation that can directly influence behavior and can be inferred from behavior. It is in contrast to a representation that can be decoded from either brain activity or model hidden layers (Brennan et al. 2012; Hewitt and Manning 2019) but is in fact not actively accessed by the brain or LLM when solving a task. For example, previous studies have found that the dependency tree of a sentence could be decoded based on the word embeddings in a LLM, but it was concerned that such a representation was artificially decoded and were not actively read out by deeper-layer neurons in the model (Belinkov 2022; van Dijk et al. 2023). Here, we demonstrated that the tree-structured representation could be directly inferred from behavior.**

**To avoid confusion, we have (1) changed "*active language use*" to "*actively contribute to behavior*"; (2) changed "*analyzing language-use behavior*" to "*analyzing rule inferencing behavior*"; (3) explicitly discussed that our deletion task demonstrated "active use" of a tree-structured representation in an inference task while the probe studies demonstrated the tree-structured representation in "everyday" language use. Therefore, these two approaches are complementary: "*Many previous studies have also demonstrated that an internal constituency representation emerges in language models or in the human brain by directly analyzing the internal representation of models, i.e., activation of hidden-layer neurons, or human neural recordings…. the analysis of internal representation is highly complementary to the analysis of word deletion behavior, since the former can be achieved during arbitrary language processing task while the latter can probe whether a constituency representation can actively contribute to behavior.*".**

Other comments

The LSTM baseline is not particularly informative, as it primarily controls for results based purely on learning. The fact that it only requires about 50 demonstrations to learn the task raises questions about the level of linguistic knowledge required to perform the task. For example, if the subjects identify that the task is about grammar and assume they have received some grammar training during their education, they could solve the task by simply referring to their acquired knowledge of

grammar as taught in school. This approach contrasts with how a preschool child learns and uses language actively.

**Response 2.7**

**Thank you for the insightful comment. The concern was that the participants might start to use school grammar to solve the task after a few trials, just like how the baseline LSTM model learned to perform the task. To address this concern, we first analyzed whether the constituent rate significantly increased after a few trials, but this phenomenon was not observed (Fig. S1). Why did the LSTM model, but not the human participants and ChatGPT, show a strong learning effect? The difference lay in how the models were trained – The former LSTM model and the current GPT-2 model were trained in a supervised manner using the expected answer to each question, while the expected answer was not provided during human/ChatGPT experiments. We have now emphasized the difference in training: "*we also trained a baseline model, i.e., an initialized GPT-2 model, to learn the word deletion task based on the ordinal position and lexical properties of each word. When the expected answer was provided to allow supervised learning, such a model still needed hundreds of samples to reach the constituent rate of humans or ChatGPT (see Supplementary Results and Fig. S5).*".**

In the last paragraph before the section "Sensitivity to Syntactic Category," the authors argue that the phenomena cannot be explained by statistical cues in word properties. However, I am uncertain how this conclusion is drawn from the LSTM results.

**Response 2.8**

**We apologize for the unclear statement. The LSTM had access to word properties through pretrained word vectors, but failed to explain the word deletion behavior with just a few demonstrations. Therefore, we meant to conclude that the properties of individual words were not sufficient to explain word deletion behavior – For example, if the participants intended to delete words that were semantically related to the words deleted in the demonstration, such a trend would have been easily captured by the LSTM. The LSTM analysis had been replaced with the GPT-2 analysis and the sentence has been removed.**

If the demonstration did not follow a constituency rule, the authors' notion suggests that humans would still be biased toward using constituency-based parsing to modify the test. I believe that demonstrating this bias would enhance the results by providing additional evidence of

humans' reliance on constituency structures, even in non-standard scenarios.

**Response 2.9**

**Thank you for the insightful question. However, we do not assume that humans or ChatGPT will delete a constituent if what is deleted in the demonstration is not a constituent. In fact, we assume the opposite – As is discussed in the paper, we assume the default strategy of both humans and ChatGPT is to output a grammatical and meaningful word string, which is generally a constituent. Therefore, participants who do not carefully examine the demonstration may tend to type in a constituent as a default tendency for language production. To avoid that this tendency is taken as evidence for a tree-structured sentence representation, we choose to delete a constitute and leave a word string that is generally not a constituent – We hypothesize that if the participants cannot discover a clear pattern in the remaining word string, they start to analyze the deleted word string. We have made these points clear in the Discussion. For example, "*In the demonstration, a grammatical and meaningful sentence was transformed into a word string that was generally not grammatical or meaningful. Therefore, the participants had to avoid their natural tendency to produce grammatical and meaningful expressions and infer the uncommon rule underlying the transformation.*".**

**To directly answer the question here, if demonstration keeps a constituent and removes a nonconstituent, we hypothesize that the participants will learn to keep a constituent, since it is consistent with their default strategy.**

Another point of concern is the sensitivity of large language models (LLMs) to prompts. If this is a one-shot learning task that does not depend on context beyond the demonstration sentence, why does Figure S2 show that the prompt can substantially affect the model's behavior (as seen with prompt 2, prompt 3, and prompt 4)? This observation echoes the earlier comment regarding the non-constituent and other responses that the models provide. This is in contrast to humans, who show consistency in their response patterns (Figure S3).

**Response 2.10**

**The reviewer correctly pointed out that the LLM results partly depended on the prompt – Different prompts do not affect the general conclusion but have quantitative influences on the results. Reliance on the prompt,**

**however, is commonly observed in LLMs (Sclar et al. 2024). For example, by just adding "Let's think step by step" to a prompt, the LLM performance on an inference task can boost by about 40% (Kojima et al. 2022) and different LLMs prefer different prompts in different tasks (Chang et al. 2024). In this experiment, the difference between prompts was quantitative instead of qualitative and all prompts supported the same conclusion – The constituent rate was higher than the nonconstituent rate and the preference for node-/parent-category rule differs between languages.**

The authors need to do more work in both the introduction and discussion sections to highlight theories and experimental results that challenge the original constituency parsing hypothesis. The first paragraph of the introduction lacks a comprehensive review of the various approaches that exist in the field to address the main question. While the authors refer to the constituency as an influential hypothesis supported by research in psychology and neuroscience, they fail to cover alternative accounts that challenge this hypothesis. Empirical work by researchers such as Fedorenko, among others, highlights significant observations that contradict the dominance of unique syntactic parsing for explaining activity in regions in the human brain that process language, for a review see (Fedorenko et al., 2024).

**Response 2.11**

**We fully agree that we should not only discuss the constituency parsing hypothesis, and we have thoroughly changed the title, Introduction, and Discussion. We actually agree with Fedorenko et al. that there may not be a specific syntactic region in the brain, but we did not include detailed discussions on this topic since the paper does not attempt to address the neural substrate for language processing. To address this concern, we have (1) rewritten most parts of the Introduction; (2) changed the discussion section title from "Syntactic capacity of LLMs" to "Sentence representation in LLMs"; (3) briefly mentioned that "*In the human brain, a distributed language network contributes to sentence representation*[73-75]*".***

In figure 5C, it is unclear to me why the points depicting the results are in bands. Can authors show the results per-subject basis, to see whether the effect is present in every subject.

**Response 2.12**

**We apologize for the confusion. In fact, data points in Figs. 5cd (Figs. S13cd in the revised manuscript) were the results from individual participants. We have now clarified this in the figure caption: "*… Each data point represents the result from a participant or a run of ChatGPT.*". Since each participant (or run of ChatGPT) only received 6 tests in each condition, the proportion of target string could only take 7 possible values, which was why the data points concentrate on 7 values.**

In figure 5C the baseline of chatGPT proportions is much lower compared to other figures, why is this the case?

**Response 2.13**

**Fig. 5c (Fig. S13c in the revised version) shows the proportion of the target strings, instead of the total constituent rate. ChatGPT often deleted constituents that were not the target strings. In this experiment, the constituent rate was 0.66 ± 0.013 (95% CI) for ChatGPT (see Supplementary Results).**

**Reviewer #3 (Remarks to the Author):**

- The tree reconstruction using CKY parsing is very interesting but the current setting sounds not apt to me. Given Figure 3, the probabilities associated with all strings in a derivation are treated as scores and they are simply summed to compute the score of a derivation. Since each probability could be regarded as a probability of a constituent given a sentence, I think they should be multiplied, or summed in log-domain, to represent the probability of deletion in a recursive manner. If summing is appropriate, I'd expect a justification for the approach.

**We thank the reviewer for his/her very constructive feedback.**

**Response 3.1**

**We agree with the reviewer that it is more reasonable to sum log probability and we have now analyzed how different operations, i.e., linear summation, linear multiplication, log summation, and log multiplication, influenced tree reconstruction. All operations, except for log multiplication, yielded similar results (Fig. S10). In the main text, we chose to report the result for linear summation since the linearly summed probability can be interpreted as the ratio of tests explained by a tree. However, we mentioned that "*In the example here, the explained ratio of a tree was the sum of explained ratio of its nodes. Since the sum of likelihood was often calculated on the log scale, we also tried log-scale integration of explained ratios and obtained similar results (see Fig. S10).*".**

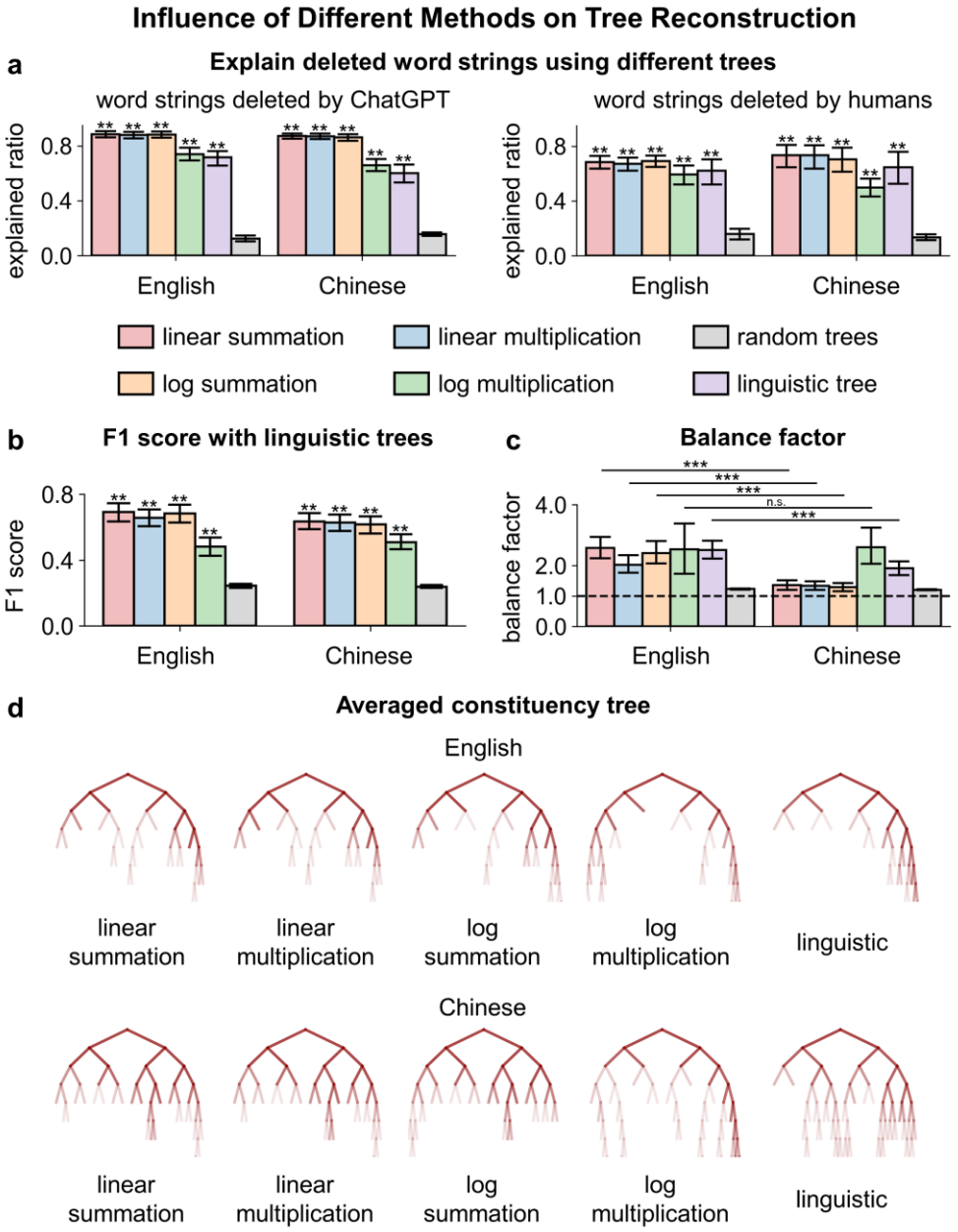


Figure S10. Constituency tree reconstructed using 4 different methods differing in how the explained ratio is integrated, i.e., linear summation (reported in the main text), linear multiplication, log summation and log multiplication. The results are generally consistent except for log multiplication. **a**. Explained ratios for different trees. **b**. F1 score with respect to the linguistic tree. **c**. Balance factor of each constituency tree. **d**. Visualization of the constituency trees averaged over all test sentences. $^{**}p < 0.01$, $^{***}p < 0.001$.

- Experiment 2 carries out experiments on arbitrary constituent categories, but I'd expect more fine-grained analysis to see if there exist any different tendencies by the categories, e.g., NP, VP and PP. In particular, it is of interest in measure the impact of PP attachment.

**Response 3.2**

**Thank you for the highly constructive suggestion. We have now separately analyzed the constituent rate and explained ratio for different constituent categories in Experiment 2 (Dataset 3 in the revised version). The results were reported in Table S3 for constituent categories with more than 10 samples in the human experiment. In general, the results are consistent across constituent categories – Participants preferred to delete constituents, and the responses to English and Chinese sentences are separately better explained by parent- and node-category rules. There are, however, a few exceptions – For ChatGPT, the constituent rate was below 0.3 for the NP[PP] structure in English and ChatGPT deleted multiple strings in 41.1% trials (Table S3). An example of the NP[PP] category demonstration was "*Some found it on the screen ~~of a personal computer~~*". ChatGPT might infer that the task was to shorten the sentence. An example ChatGPT response was "*Competition in ~~the~~ sale ~~of complete bikes is~~ heating up too*", which captured the gist of the sentence but did not delete a constituent.**

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 |

- Penn-Treebank and Chinese Treebank are very popular data sources and have been used in many prior studies. I'm wondering if the data leakage might have an impact to this studies since the dataset, at least non-bracket texts, are found elsewhere mainly intended for language modeling. Also note that ChatGPT is capable to show Penn-

Treebank like bracketing without any few-shot examples. It is better to measure how accurately GhatGPT can generate the bracket structure that is consistent with the gold annotation, and how that will be related to the accuracies of the deletion task.

**Response 3.3**

**Following the constructive suggestion, we have asked ChatGPT to explicitly parse the sentence structure, using the prompt described in a recent study (Tian et al. 2024). The explicitly parsed constituency tree reached F1 scores comparable to that of the deletion-based constituency tree (Fig. S12a). However, the explicitly parsed constituency tree differed from the deletion-based constituency tree, and performed significantly worse at explaining the word deletion behavior (Fig. S12b). Furthermore, for individual sentences, the F1 score of explicitly parsed trees was not strongly correlated with the F1 score of deletion-based trees (Fig. S12c), suggesting that the capacity to explicitly parse a sentence dissociated from the capacity to implicitly extract linguistic constituents in the word deletion task.**

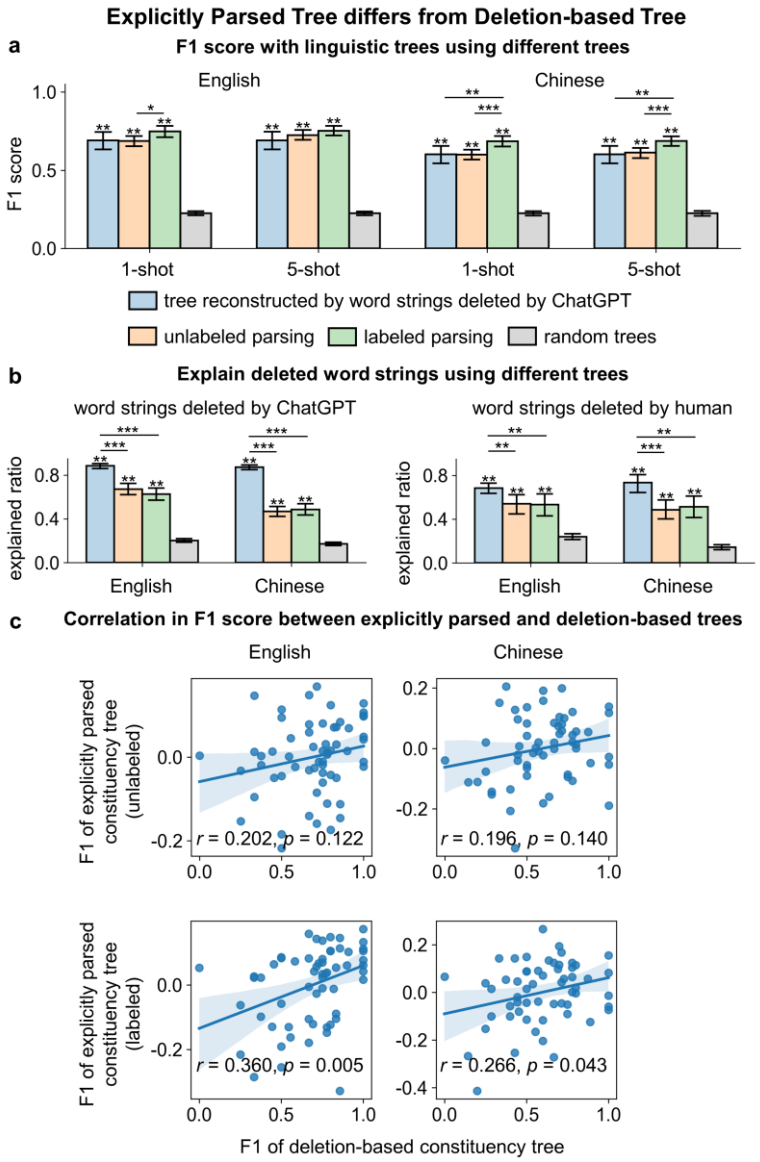


Figure S12. Comparison between syntactic tree explicitly parsed by ChatGPT and deletion-based tree. Explicit parsing is achieved using prompts[2] using either 1-shot or 5-shots learning. **a**. F1 score with respect to the linguistic tree. One-shot and 5-shot parsing obtain similar performance and we 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. **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. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

**Please also see Response 1.1 for other analyses to address the data leakage problem. The comparison between the explicitly parsed trees and deleted-based trees have been added to Fig. S12.**

Minor comment and suggestions:

- The first experiments comprises two sub-experiments, one for measuring the accuracies of a noun phrase deletion and the other for investigating the impact of parent constituency. It is not clear why these experiments are treated as a single experiment given that they have different theme.

**Response 3.4**
**Experiments in the previous manuscript referred to behavioral testing on humans/ChatGPT, but we agree it could be misleading. Therefore, the data collected during Experiment 1 is now referred to Dataset 1 and there are two separate analyses on the same dataset.**

- It is not clear whether an LSTM is a good baseline, and I think GPT-2 will be a better baseline since that follows Transformer architecture which is also used in ChatGPT. If possible, it is better to run an experiment on GPT-2 as an additional baseline to test the capacity.

**Response 3.5**
**Following the valuable suggestion, we have now used GPT-2 to construct baseline models. With a single demonstration, the constituent rate of GPT-2 was not significantly above chance. In Dataset 1 (Figs. S5a & S5b), after trained with about 100-500 demonstration-test pairs, the GPT-2 model reached the performance of ChatGPT and its performance further increased with more training samples. However, GPT-2 consistently preferred the node-category rule for both Chinese and English, not showing the language-dependent preference as humans and ChatGPT. A similar pattern is observed for Dataset 3 (Figs. S5c & S5d). The results of GPT-2 have been added to Fig. S5.**

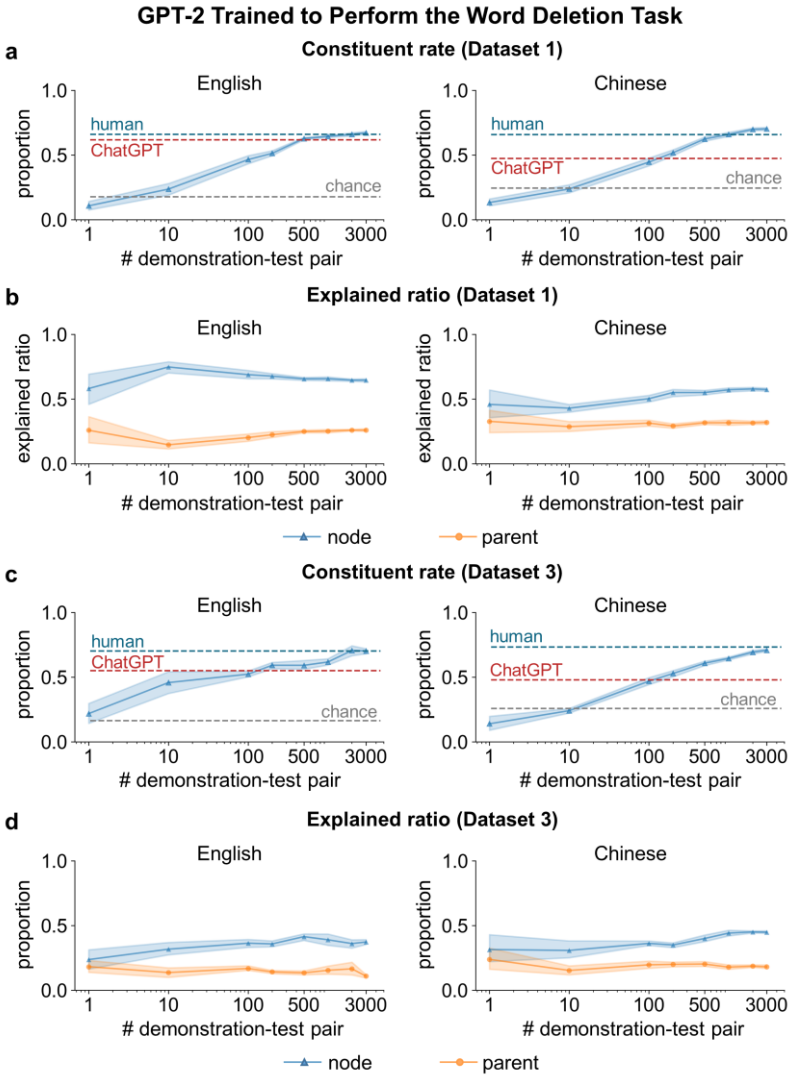


Figure S5. Performance of a baseline model, i.e., a naive GPT-2 model, trained to perform the word deletion task. The naive GPT-2 model is not pretrained, and its input includes word position embeddings and pretrained word embeddings. **ac.** Constituent rate of the GPT-2 model as a function of the number 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.

- If possible, it is interesting to measure the impact of L2 for Experiment 3 to see if any biases in the tree structures.

**Response 3.6**

**Following the valuable suggestion, we have now tested L2 speakers on Experiment 3 (Dataset 4 in the revised version), following the same**

**procedure for native speakers. The results are generally consistent between L2 and native speakers. The deletion-based constituency tree reconstructed by L2 speakers exhibited a slightly higher F1 score than ChatGPT and native speakers (Fig. S11b), and explained about 80% of the strings deleted in the L2 experiment (Fig. S11a).**

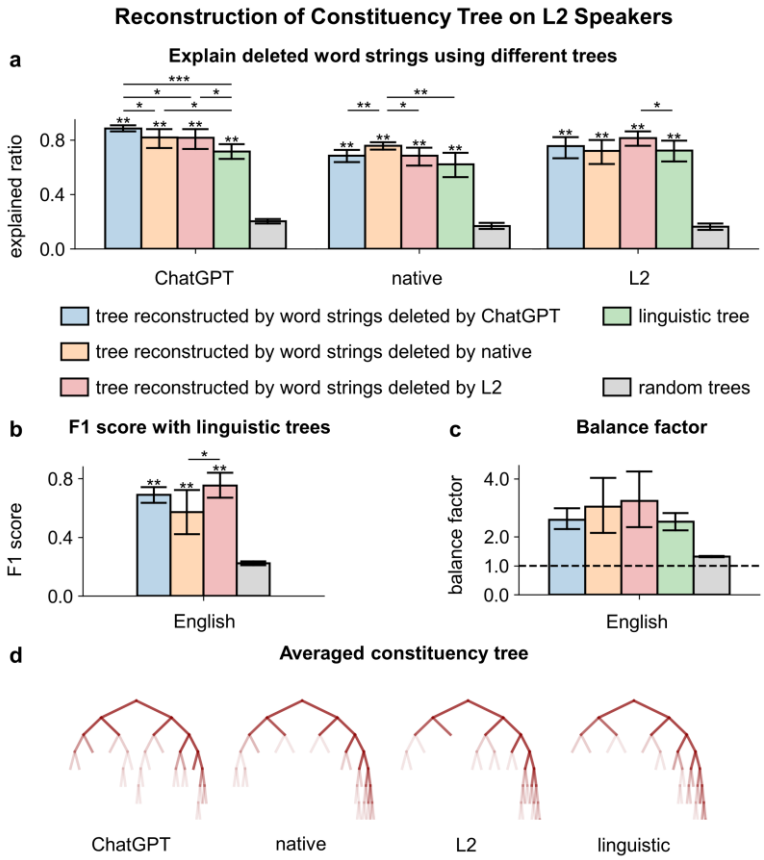


Figure S11. Reconstruction of constituency tree based on word deletion behavior for L2 English speakers ($N$ = 34, 19-29 years old, mean 23 years old; 22 females), following the same procedure for native speakers in Dataset 4. The results are generally consistent to ChatGPT and native speakers. **a**. 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. S11), 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. **d**. Visualization of the constituency trees averaged over all test sentences. *$p$ < 0.05, **$p$ < 0.01, ***$p$ < 0.001.

- The current experiments are based on one-shot example, i.e., a single demonstration, for each prompt to an LLM. I think it is of interest to see if more shots, e.g., 3-shots, would significant impact the performance or not, in the same manner as observed in LSTM experiments.

**Response 3.7**

**Following the valuable suggestion, we have now tested N-shot learning on ChatGPT (N ≤ 5). The constituent rate slightly improved with more demonstrations (Dataset 1: from 0.473 to 0.515 for Chinese and 0.618 to 0.780 for English. Dataset 3: from 0.479 to 0.513 for Chinese and 0.550 to 0.702 for English), while the explained ratio of the node- and parent-category rules was not clearly influenced by the number of shots. The results of N-shot learning have been added to Figs. S8 and S9.**

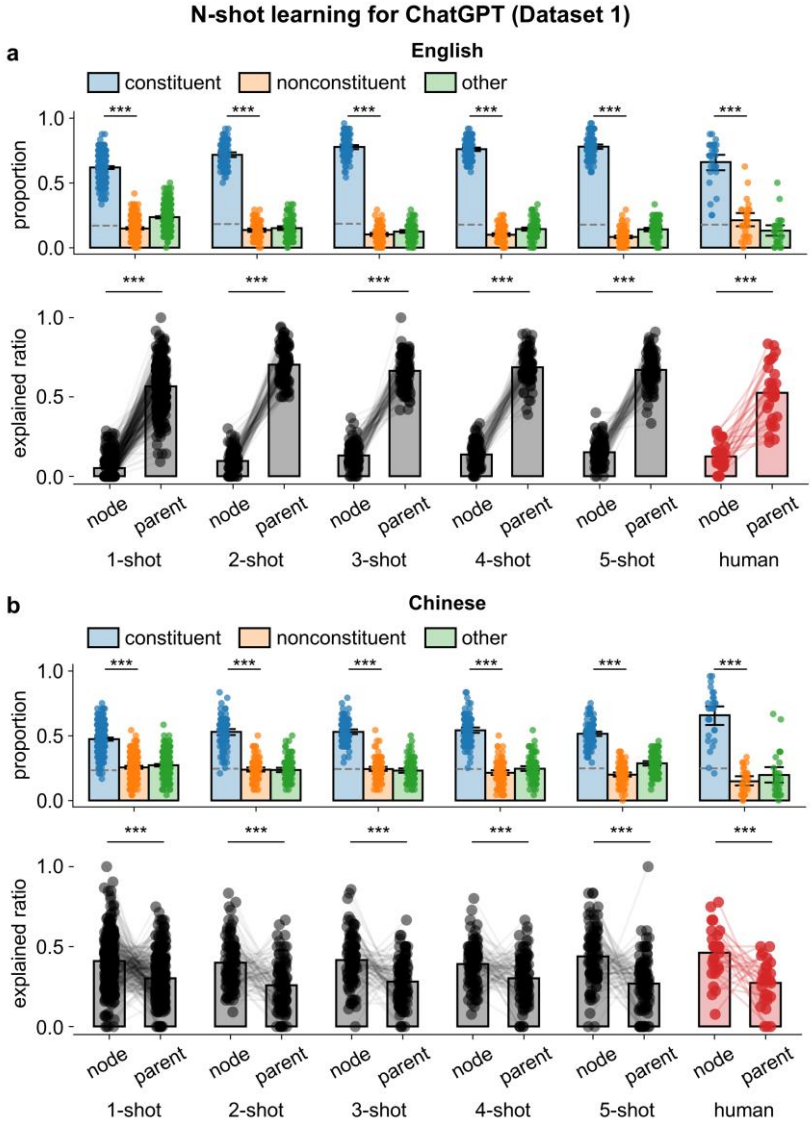


Figure S8. ChatGPT N-shot learning for Dataset 1. **a.** Constitute rate slightly improves with more demonstrations. **b.** The explained ratio for the node- and parent-category rules do not clearly change with more demonstrations. ***$p$ < 0.001.

- Constituency test has been investigated for unsupervised parsing and it seems to be marginally related: https://aclanthology.org/2020.emnlp-main.389/

**Response 3.8**

**Thank you for pointing out this highly relevant paper and we have now cited it – "*Based on linguistic constituency tests, algorithms have been designed to parse the constituency structure of a sentence in an unsupervised manner*[53]."**.

**Reference**


Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and language*, *120*(2), 163–173.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., et al. (2024). A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, *15*(3). https://doi.org/10.1145/3641289

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=e2TBb5y0yFf

Oxford University Press. (1991). *Oxford Collocation Dictionary (Bilingual version)*. Foreign Language Teaching and Research Press Pub.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=RIu5lyNXjT

**We would like to thank the editors and reviewers again for their highly constructive feedback. We are pleased that the reviewers have found the revised manuscript satisfying.**

REVIEWER COMMENTS:
Reviewer #2 (Remarks to the Author):
2.1. Interpretation of Results in Figure 5
Lexical vs. syntactic information: The results in Figure 5 suggest that the absence of lexical information dramatically affects the "constituent rate" in models across both languages. This pattern, in my view, does not strongly support a primarily syntactic constituency-based deletion process. Indeed, the authors also note a similar pattern in Supplementary Figure S6 for Dataset 2. I would recommend adjusting the claim about the separability between lexical information and syntax; or if appropriate moving it to the supplementary section. The data as it stands suggest that lexical cues are intertwined with whatever syntactic pattern might be at play.

**Response 2.1**
**We agree with the reviewer that lexical cues are intertwined with syntactic cues during the deletion task. The purpose of the experiment in Fig. 5 is to show that syntactic cues can influence the word deletion task, instead of suggesting that lexical cues and syntactic cues are independent.**

**For the particular experiment reported in Fig. 5, the constituent rate was low because ChatGPT often refused to answer the questions, e.g., by responding "*It is impossible to guess what John would say for the sentence as his speaking style does not seem to follow any logical pattern or rule.*" The ratio of "refuse to reply" is higher than 0.33 in English (see the original Table S2). The questions in the experiment might be too challenging to answer since (1) the test sentence was not semantically coherent and (2) the syntactic cues were weak – The demonstration sentence was randomly selected from treebanks (following Dataset 1) and its structure was not required to match with the test sentence.**

**We have now provided a new version of the experiment, in which the task difficulty was reduced by requiring the demonstration sentence to have a similar (but not the same) structure with the test sentence. Following Dataset 2, the demonstration sentence had a VP[NP] structure while the test sentence had a VP[PP[NP]] structure. See Figure R1 for examples. In the new experiment, ChatGPT answered the questions most of the times and the constituent rate was higher than 0.7, comparable to Dataset 2 (see Fig. S8 and Table S2).**

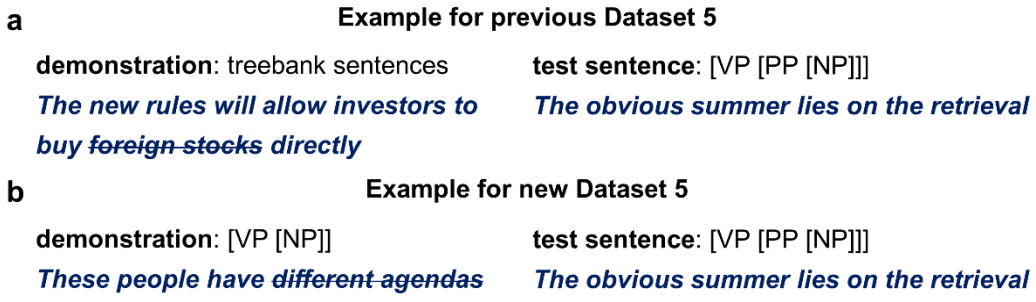


Figure R1. Examples for the previous Dataset 5 and new Dataset 5 (Dataset S1 in the revised manuscript).

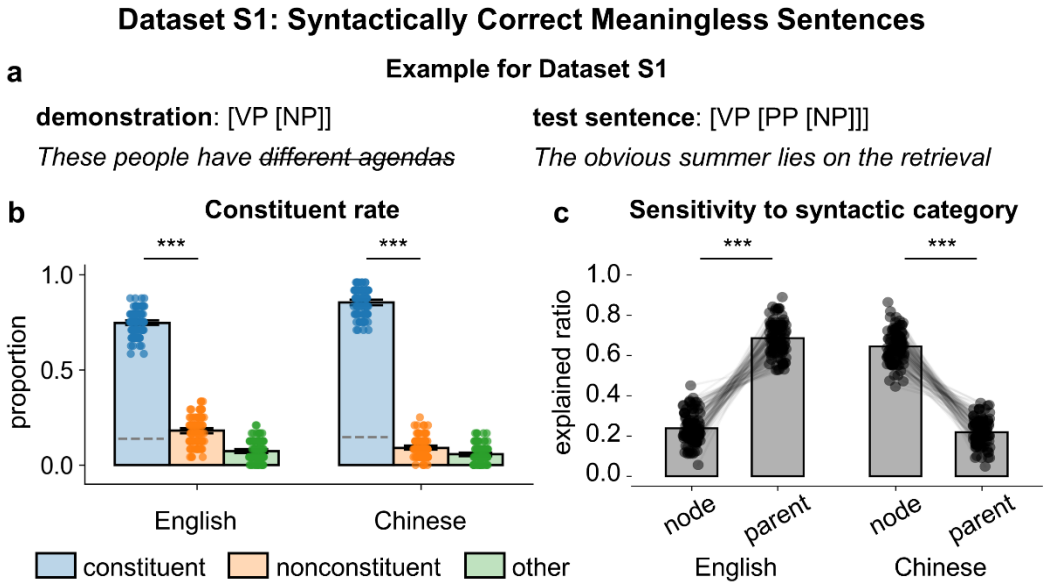


Figure S8. Results of Dataset S1. The test sentences were syntactically correct but meaningless sentences that are constructed based on the test sentences of Dataset 2. The demonstration sentence was the same as Dataset 2. **a**. Example sentences. **b**. Constituent rate of word strings deleted by ChatGPT. **c.** Explained ratio for the node- and parent-category rules. Results are consistent with the results of Dataset 2. ***$p < 0.001$.

**Following the constructive suggestion of the reviewer, we have moved Fig. 5 to the Supplementary Material and adjusted claims that might have suggested separability between lexical and syntactic information. For example, we removed "*confirming primary reliance on syntactic structure*" when describing the results of Dataset S1.**

**We changed "*These results indicated that humans and ChatGPT mainly relied on syntactic structure instead of semantic plausibility…*" to "*These results indicated that native speakers and ChatGPT were not strongly influenced by sentence-level semantic plausibility …*"**

Data leakage discussion (Lines 533–535, 536–541, and 541–543).
The manuscript takes the results from Figure 5 as evidence against data leakage, uses Figure S6 to argue that LLMs are not relying on "only" syntactic information, and then concludes that these outcomes generalize to sentences not used for LLM training (and are not simply based on possible phrase annotations). However, it is unclear if these observations alone can justify the conclusion that there is no data leakage or that the behavior generalizes beyond the training set. The current evidence may not be sufficient to rule out other explanations (e.g., partial memorization of corpus, or alternative strategies models employ).

**Response 2.2**
**Thank you for the insightful comment. The purpose of the data leakage analysis was not to demonstrate that data leakage could not happen. Instead, it was to demonstrate whether the current conclusions could generalize to sentences that were not used to train LLMs. In the new Dataset 5 (see Response 2.1), the constitute rate was comparable to other datasets that used semantically coherent sentences. We believed that this result demonstrated that our results could generalize to the semantically implausible sentences that we constructed in the study and were not possibly included in LLM training. We agree, however, with the reviewer that we should further explore alternative strategies that humans and LLMs might employ and please see Response 2.3 for details.**

2.2. Differences Between Humans and LLMs
There are places where model behaviors clearly diverge from humans. For instance, in Supplementary Figure S2, LLM responses are sensitive to different prompts, whereas human behavior is relatively robust.

In addition, the text sometimes implies that if LLMs do not delete the word, they do not understand the task. However, an alternative view is that the LLM's behavior reflects its inference over the instruction plus sentence modification example. This is not necessarily a failure to understand, but rather a different type of reasoning or strategy.

**Response 2.3a**
**Thank you for the insightful comment. We have added discussion on why LLMs are sensitive to prompts:**

***"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 ChatGPT's performance was more variable across different prompts than that of human's. 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"".***

I think it would strengthen the paper if the authors could speculate further on why LLM behavior might differ from human behavior. For instance, do LLMs rely more on surface-level or distributional cues? Are they more literal or less flexible with instructions?

**Response 2.3b**
**Following the constructive suggestion, we included two new metrics to quantify the difference between human and LLM behaviors. Response 2.3a suggested that the LLM was more engaged in role playing. Therefore, it was possible that LLM might be more biased to infer a plausible speaking style while the human participants might be more**

**biased to infer a more abstract rule. Therefore, we hypothesized that the LLM responses were (1) semantically more consistent with the test sentence, and (2) more grammatically acceptable. Semantic similarity was characterized using SimCSE (Gao et al. 2021) and grammatical acceptability was evaluated using GPT-4o, based on prompts described in a previous study (Zhang et al. 2024).**

**Unlike the analysis of constituency/dependency metrics related to the deleted word string, the semantic similarity and grammatical acceptability analyses did not require the response to delete a word string and therefore could be applied to the "other" responses. Therefore, the two analyses were separately applied to responses that deleted word strings and the "other" responses, excluding the conditions in which LLMs refused to answer the question.**

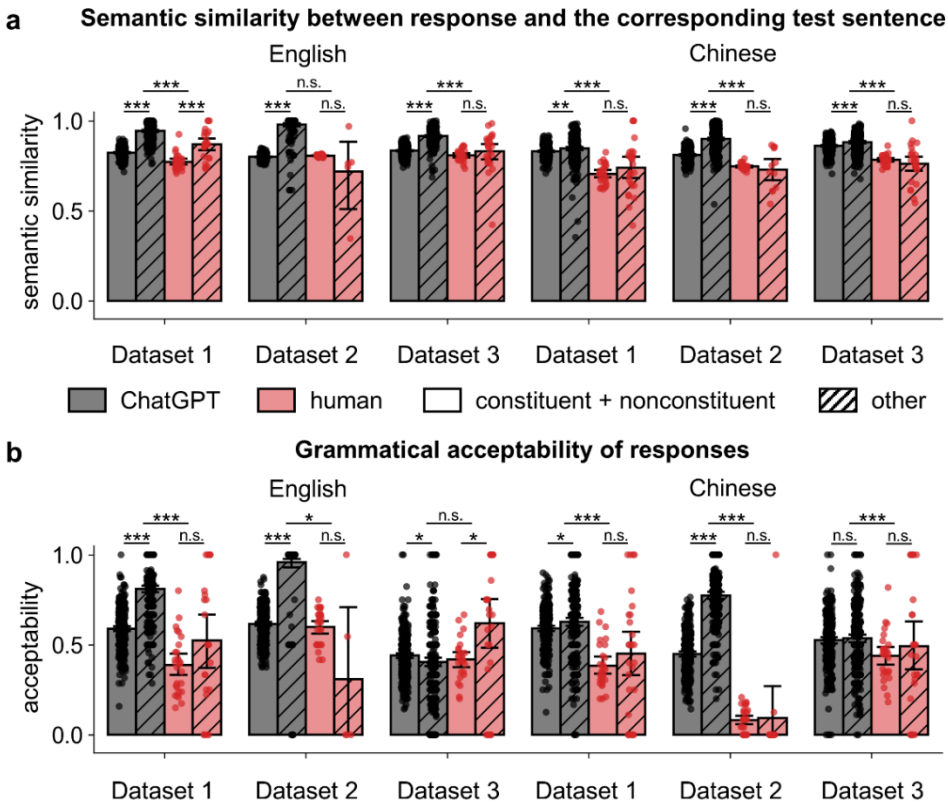


Figure S4. Semantic similarity and grammatical acceptability in Datasets 1-3. **a**. Semantic similarity between the response and the corresponding test sentence. **b**. Grammatical acceptability of responses. ChatGPT tends to preserve the semantics of the test sentence and generate more grammatically acceptable responses than human. Each dot on the bar graph represents data from a human participant or a single run of ChatGPT. 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.

**The results showed that the ChatGPT responses in general were semantically more similar to the test sentence and was more grammatically acceptable than humans (Fig. S4), and was particularly higher in the "other" responses. These results potentially reflected a competition between the instruction of the word deletion task and the natural tendency of LLMs to generate a coherent sentence based on the context.**

3. Minor Comments and Clarifications
1.Figure S4 (SVM Results for Remaining Strings). For Dataset 1 (English/ChatGPT), constituency, dependency connectivity, and MDD are equally good at discriminating the remaining string. For Chinese/ChatGPT and Chinese/humans, dependency connectivity seems to outperform constituencies. This is mostly true for Dataset 3 as well.

Please clarify in the text what you conclude from these differences. Why might dependency connectivity perform better in Chinese? How should we interpret "remaining string" results in the broader context of language processing?

**Response 2.4**
**Thank you for the insightful question.**

***1. Regarding better performance of dependency grammar on Chinese***

**There could be more uncertainty or ambiguity associated with identifying constituents in Chinese compared to English, since Chinese lacks case markers (Li and Thompson 1989; Wu and He 2015) and has a relatively flexible word order (Huang 1998; Sybesma 2021). Constituency grammar is also understudied in Chinese compared with English. Therefore, one possibility is that dependency grammar is more suitable to a language with flexible word order such as Chinese (Kübler et al. 2009). Another possibility is that the treebank constituency annotations used in the study is suboptimal. We have now discussed this issue when describing the results:**

***"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]."***
**Please see Response 2.7 for additional discussions.**

***2. Regarding the remaining string***

**Since the instruction only asked the human participants and LLMs to infer a speaking style and did not explicitly ask them to delete a word string. The participants could in principle infer the rules by analyzing either the remaining word string or the deleted word string. An analysis of the remaining string is more straightforward since the string is directly available in the text. Nevertheless, in most cases, the participants would fail to find a rule in the remaining string and therefore would turn to the deleted string.**

**In the results, it was clear that the constituency of the deleted string more strongly influenced behavior than the constituency of the remaining string. In contrast, dependency-based metrics of the remaining string, however, more strongly influenced behavior than dependency-based metrics of the deleted string. A possible explanation was that if the participants attempted to infer a rule just based on the remaining string, i.e., what is available in the text input, they were likely to infer a rule based on dependency-based metrics, since the remaining string was in general not a constituent. In contrast, if the participants turned to analyze the deleted string, they were likely to infer a rule based on constituency since the deleted string was always a constituent. We have added the following discussions to the manuscript:**

***"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."***

2.Lines 62–78: Clarity of Argument. The paragraph is somewhat confusing regarding the argument about LLMs not explicitly encoding constituents or dependencies, and how that "raises the possibility" that these structures are not necessary for language understanding. Recommendation: Please clarify how one leads to the other. The absence of explicit encoding does not necessarily imply these constructs are unnecessary—it might only imply they are not represented in the same way as symbolic syntactic structures.

**Response 2.5**
**We thank the reviewer for pointing this out. We have incorporated what the reviewer's comment into the Introduction:**

**"*There is a growing body of work showing successful performance of large language models (LLMs) in many language processing tasks[20-22] but how LLMs understand language remains difficult to explain[23-25]. LLMs do not explicitly encode constituents or dependencies, but the absence of explicit encoding does not necessarily imply these constructs are entirely unnecessary for LLMs. One possibility is that a latent constituency/dependency representation can emerge and functionally explain the linguistic performance of LLMs.*"**

3.Line 367: CKY Reference. The statement about CKY is described as "not informative." Please elaborate what CKY is.

**Response 2.6**
**We have now briefly introduced the CKY algorithm in the *Results*:**

**"*…we used the CKY algorithm to find the binary constituency tree that had the highest explained ratio among all possible binary trees that could generate the test sentence (see Methods).*"**

**We have also added a more detailed description of the CKY algorithm in the *Methods*:**

**"*The CKY algorithm was a dynamic programming algorithm used to find the most probable parse tree for a sentence under a Probabilistic Context-Free Grammar (PCFG). A parse tree was a binary tree in which the leaf nodes were the words in the sentence and the edges was not allowed to cross. Here, the probability of a parse tree was defined as the explained ratio."***

4.Dependency Grammar vs. Constituency Grammar. The authors argue that the dependency grammar approach is still inferior to constituency-based analysis (Supplementary figure S4). Follow-up Question: If demonstration sentences were explicitly designed to illustrate dependency-based deletions, would the same advantages of constituent-based analysis hold?

**Response 2.7**

**Thank you for the question. We want to clarify that we did not argue that the dependency approach is inferior to constituency approach. Instead, we emphasized the commonality between the two approaches, as the basic forms of these two types of representations are isomorphic to each other (see Introduction, page 1). Depending on the specific implementations, it is conceivable that one can find an optimal dependency-based analysis that outperforms the constituency-based analysis for current datasets, and vice versa. To answer the follow up question, we think it is an interesting empirical question whether the advantages of constituent-based analysis still hold if demonstration sentences are explicitly designed to illustrate dependency-based deletions. In future work it is worth exploring what specific implementations and optimizations of the structural analysis can best explain human or LLM behavior, using different types of stimuli. We have added the following paragraph to the Discussion section.**

**"*Furthermore, the current study investigated whether humans and LLMs build a hierarchically-structured or linear representation of sentences. Different types of structural analysis may focus on different structural notions (e.g. constituency structure, dependency relations, etc.). Future work should further explore the similarities and differences when different implementations of structural representations are applied to explain human and LLM behavior, especially when demonstrations are***

***constructed to highlight different variations of structural representations.*"**

Overall, the revised manuscript is much clearer and stronger than before. Nevertheless, additional clarifications on the points above would help solidify the central claims—particularly around the interplay of lexical vs. syntactic cues, the possibility of data leakage, and the differences between LLM and human deletion patterns. Moreover, addressing ambiguities in certain datasets (especially Dataset 5) would ensure that readers clearly understand how these results factor into the main conclusions.

I hope these comments and suggestions are helpful in refining the final version of the paper. Thank you again for the careful revisions and for engaging with the feedback so thoroughly.

**Thank you again for the very constructive feedback. The comments have been very helpful in strengthening the paper.**

**Reference**


Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6894–6910). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.552

Huang, C.-T. J. (1998). *Logical relations in Chinese and the theory of grammar*. Taylor & Francis.

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Sybesma, R. (2021). Voice and Little v and VO–OV Word-Order Variation in Chinese Languages. *Syntax*, *24*(1), 44–77. https://doi.org/10.1111/synt.12211

Wu, F., & He, Y. (2015). Some Typological Characteristics of Mandarin Chinese Syntax. In *The Oxford Handbook of Chinese Linguistics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199856336.013.0020

Zhang, Z., Liu, Y., Huang, W., Mao, J., Wang, R., & Hu, H. (2024). MELA: Multilingual Evaluation of Linguistic Acceptability. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2658–2674). Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.146