Limitations

First, our experiments were conducted on only two neural language models (BERT-base and BERT-large). It remains unclear whether similar results would be obtained for larger models or other architectural variants. However, our method is applicable to any open neural model, making it feasible to extend this analysis to a broader range of models in future research.

Second, this study focused solely on English data. It is uncertain whether similar layer-wise syntactic structure construction patterns would be observed when applying our method to other languages. Nevertheless, our approach is language-agnostic, making cross-linguistic analysis an important direction for future work.

Furthermore, semantic cues may influence the results of syntactic probes. Our study does not fully account for these potential semantic confounds. Future research should consider methods to more rigorously isolate syntactic information, such as using Jabberwocky sentences as demonstrated by Maudslay and Cotterell (2021).

Lastly, our method relies on dependency parsing, primarily due to the use of the structural probe from Hewitt and Manning (2019), which analyzes distances between tokens in the embedding space. This approach is inherently tied to formalisms like dependency grammar that focus on relationships between terminal symbols (tokens). As a result, our method may not be directly applicable to other grammatical theories or parsing approaches that involve non-terminal symbols, such as constituency grammar. This limitation arises because analyzing distances between tokens does not capture the hierarchical structures represented by non-terminals. Future work could explore adapting our method or developing new probing techniques that can handle non-terminal representations to verify the generalizability of our findings.

Ethical considerations

The training corpus is extracted from public web pages and thus could be socially biased, despite its popular use in the NLP community.

A Example Sentences for Structure Sets

Below are the example sentences corresponding to the four primary structure sets described in §5.1:

`Macro` with micro relations `nsubj` and `dobj`:
e.g.) The concert caused a major stir.

`Macro` with micro relations `nsubj` and `prep`:
e.g.) The match ended in a goalless draw.

`Macro` with micro relations `nsubj` and `attr`:
e.g.) Her parents were music professors.

`Macro` with micro relations `nsubj`, `prep`, and `dobj`:
e.g.) The film received positive reviews from critics.

B The Experimental Results for GPT-2 Models

Figure 8 and Figure 9 show the experimental results with the same experimental setup as §5, but conducted with GPT-2.

C Hyperparameters

Hyperparameters for our experiments are shown in Table 1. All models were trained and evaluated on 4× NVIDIA RTX A5000 (24GB). The total computational cost for all experiments in this paper is about 120 GPU hours.

Optimizer	Adam
Learning rate	1e-3
Number of epochs	40
Learning rate scheduler	ReduceLROnPlateau
Batch size	32

Table 1: Hyperparameters for our experiments

D License of the Data and Tools

The licenses of the data and tools used in this paper are summarized in Table 2. We confirmed that all the data and the tools were used under their respective license terms.

Data/tool	License
`spacy` (Honnibal et al., 2020)	MIT
`transformers` (Wolf et al., 2020)	Apache 2.0
`Wikitext-103` (Merity et al., 2016)	CC-BY-SA 3.0

Table 2: License of the data and tools

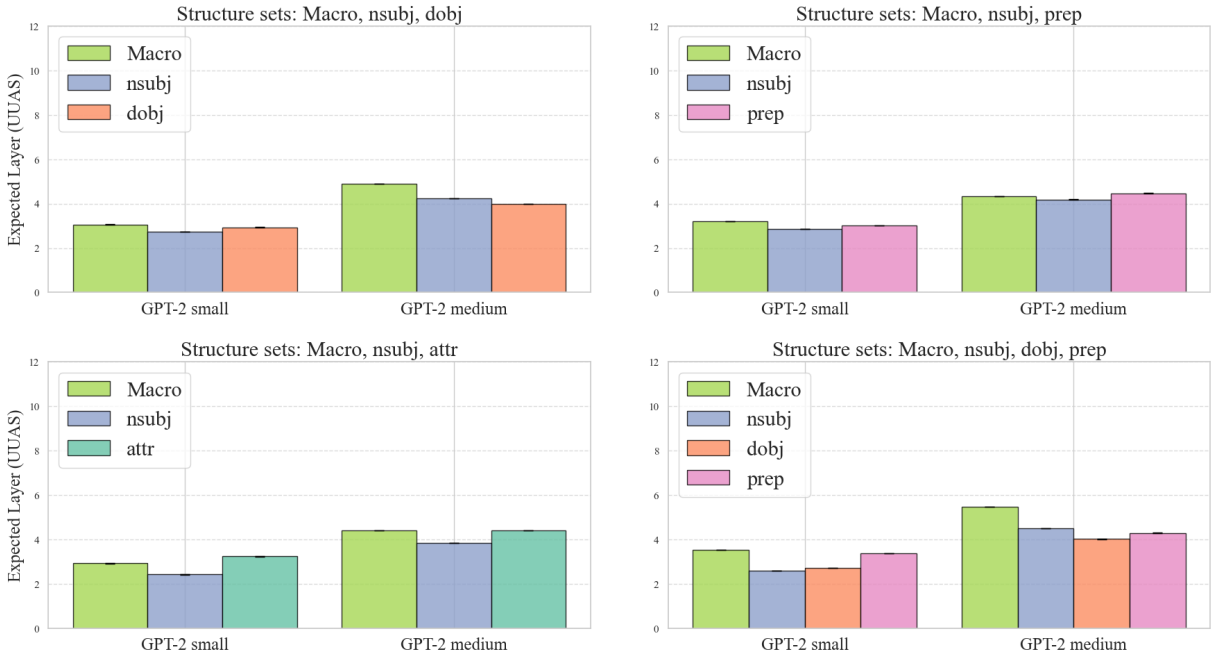

Figure 8: Expected layer for each GPT-2 model across different structure sets. Error bars represent standard deviation across 5 random seeds.

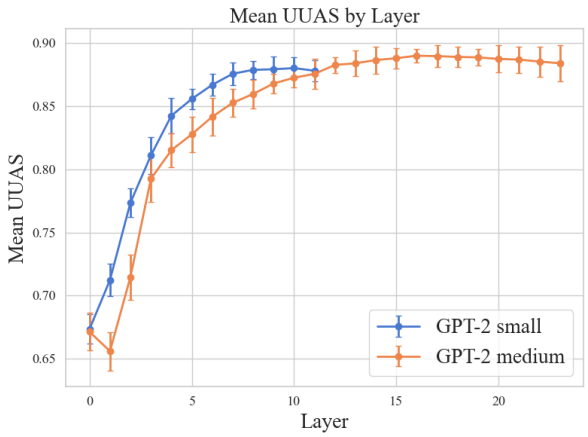

Figure 9: Global UUAS by each layer for each GPT-2 model. Error bars represent standard deviations across 5 random seeds.