Evidence of Generative Syntax in Large Language Models (Appendix)

Anonymous ACL submission

References

Mark Davies. 2008–. The corpus of contemporary american english (coca).

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Idan Landau. 2024. *Elements in Generative Syntax*, chapter Control. Cambridge University.

Maria Polinsky. 2013. *The Cambridge Handbook of Generative SYntax: Grammar and Syntax*, chapter Raising and Control. Cambridge University Press.

Paul M. Postal. 1974. *On raising: One rule of English and its theoretical implications*. MIT Press.

M.E. Sánchez, Y. Sevilla, and A. Bachrach. 2016. Agreement processing in control and raising structures. evidence from sentence production in spanish. *Lingua*, 177:60–77.

A Limitations

Our work still faces limitations in that it does not enable a full reconstruction of the hierarchical syntactic tree. This is a limitation currently inherent to the data and format of LLMs. As can be seen in Figure 3 in the main text, the generative syntax trees consist of branches and nodes that do not overtly appear in the final derivation. That is to say, trace nodes/moved elements are not surfaced, nor are all syntactic elements (such as tense) realized by a separate word. Because of this, an LLM's contextualized word embeddings cannot currently be used to directly derive the sub-surface syntactic trees. The methodology that we've deployed allows us to probe for behaviors that would indicate that LLMs have captured more complex, hierarchically-rich structural information within their embeddings, but this cannot be directly shown the way Hewitt and Manning (2019) did with the one-to-one mappings of dependency parses. Thus, our work is still largely in the tradition of much of linguistics. We cannot directly observe people's mental grammars, but we probe for their knowledge and structures using measurements that indicate how people process and produce language. Similarly, our use of Hewitt and Manning (2019)'s probe also provides an apparatus to measure behaviors that we can use to reverse-engineer the possible behaviors and mechanisms that would derive such results. The interpretability question of LLMs is not far at all from the research questions of linguistics.

B Data Generation

Our data was generated through combinatorics of sets of words for each grammatical role. In short, our sentences followed the base structure of:

(1) [Subject] [past-tense matrix verb] [to] [embedded verb] [direct object].

In order to easily control linear distance, subject verbs were limited to pronominal subjects. Because control verbs are typically volitional, all subjects were prototypically [+HUMAN], but varied in Case, Gender, and Number (see List (2)).

We additionally selected 61 transitive verbs for our embedded verb (see List (3)). Of these verbs, 30 verbs implied human direct objects while 31 implied non-human direct objects. That is to say, a person can *flatter* the king, but it's nonsensical for them to *drink* the king. Conversely, they can drink sodas, but it would be hard to flatter an inanimate soda. This dichotomy was taken into account when selecting direct objects. Thus, when the direct object was a single-word pronominal, inanimate-coded verbs permutated through *it, that, this, stuff,* and *things* while animate-coded verbs permutated through *me, you, him, her, us, them, everyone,* and *someone*. The animate list is longer;

however, the animates were truncated as we omitted direct objects that were the correspondent of the subject. That is to say, if the subject was "she", the direct object would *not* be "her." Additionally, to avoid scope ambiguities, we excluded instances where the subject was "someone" and the direct object was "everyone".[1] Nominal direct objects ("the" + the noun) were more limited as we selected only one plausible noun to pair with the embedding verb.

(2) **Subjects**: You, He, She, We, They, Everyone, Someone

(3) **Embedded Verbs**

a. *Inanimate-coded Verbs*: say, yell, whisper, shout, think, write, read, cook, eat, drink, buy, sell, rent, provide, offer, collect, grab, steal, bump, move, kick, break, destroy, build, wash, wear, sew, mend, fix, enjoy

b. *Animate-coded Verbs*: kiss, hug, slap, wrestle, fight, bully, harass, intimidate, insult, slander, annoy, tease, seduce, flatter, comfort, compliment, question, interrogate, interview, meet, fire, hire, pay, reward, punish, scold, teach, train, serve, admire

(4) **Pronominal Direct Objects**

a. *Inanimates*: it, that, this, stuff, things

b. *Animates*: me, you, him, her, us, them, everyone, someone

(5) **Nominal Direct Objects and Their Corresponding Embedded Verb**: say the words, yell the answer, whisper the clues, shout the lyrics, think the worst, write the essay, read the book, cook the meal, eat the food, drink the sodas, buy the clothes, sell the toy, rent the apartment, provide the supplies, offer the bribes, collect the rocks, grab the keys, steal the gold, bump the table, move the chairs, kick the ball, break the glass, destroy the house, build the tower, wash the socks, wear the uniform, sew a shirt, mend the tears, fix the issue, enjoy the dessert, kiss the puppy, hug the baby, slap the clown, wrestle the children, fight the administration, bully the student, harass the reporter, intimidate the intern, insult the actress, slander the politician, annoy the teenagers, tease the toddlers, seduce the actor, flatter the king, comfort the victims, compliment the model, question the judge, interrogate the witness, interview the suspect, meet the manager, fire the employee, hire the applicant, pay the consultant, reward the winner, punish the cheaters, scold the liars, teach the trainees, train the recruits, serve the queen, admire the hero

We utilized the following suite of diagnostics to select our condition matrix verbs:

1. SR predicates can be replaced by an expletive *it*; SCs cannot. (Polinsky, 2013; Landau, 2024)
   - Base: John seems/wants to annoy his brother.
   - SR: It seems John annoys his brother.
   - SC: *It wants John annoys his brother.

2. SR predicates can be replaced by an expletive *there*; SCs cannot. (Polinsky, 2013; Landau, 2024)
   - Base: A mouse seemed/wanted to be stuck in the house.
   - SR: There seemed to be a mouse stuck in the house.
   - SC: *There wanted to be a mouse stuck in the house.

3. SR predicates allow for idioms to retain their idiomatic meanings; SCs can only retrieve the literal meaning. (Polinsky, 2013; Landau, 2024)
   - Idiom: Every time my friend pet-sits, my fish *go belly up*. *(meaning: my fish die)*
   - SR: My fish seem to go belly up every time my friend pet-sits. *(Die meaning: still easily accessible)*
   - SC: My fish want to go belly up every time my friend pet-sits. *(Die meaning: less accessible if at all)*

4. When SR sentences are passivized, the meaning is equivalent. Passivization of the SC

[1] We did, however, include the distributive scopal alternative in which "someone" is the subject of an "everyone" object. The two readings of this can either be there is some person *X* who [verbs] everyone, or it can be the distributive reading where for every person *X*, they are [verbed] by someone (not necessarily the same someone). The inclusion of a scopal ambiguity was due to an oversight on our part; however, because there were proportionally fewer of these pairings and because these pairings occurred in both conditions, the possible scopal ambiguity should not have an impact on our results.

yields asymmetric meanings. (Sánchez et al., 2016)

- SR: The teachers seemed to select the volunteers. = The volunteers seemed to be selected by the teachers.
- SC: The teachers wanted to select the volunteers. ≠ The volunteers wanted to be selected by the teachers.

5. SRs allow for scope ambiguity, but SCs do not. (Polinsky, 2013; Landau, 2024)

   - SC: Someone from HR seems to win the office raffle every year.
     - *De re* reading: There is someone specific in HR who seems to win the raffle each year.
     - *De dicto* reading: It seems that the winner of the office raffle each year is someone from HR.
   - SR: Someone from HR wants to win the office raffle every year.
     - *De re* reading: There is someone specific in HR who wants to win the raffle each year.
     - *De dicto* reading: inaccessible.

6. Singular subjects of SC predicates can participate in plural-coded verbs,[2] but SRs cannot. (Landau, 2024). By plural-coded verbs, we mean those that necessitate multiple participants (e.g., it's ungrammatical to say "I met at midnight" as "meeting" requires two or more participants).

   - SR: *The student seemed to meet in the library.
   - SC: The student wanted to meet in the library.

From this, we selected 6 SC verbs—all of which met Landau (2024)'s criteria for logophoric control predicates—and 6 SR verbs, listed in List (6).[3]

(6) **Matrix Verbs**

a. *Subject Control Verbs*: wanted, expected, wished, liked, hated, promised

b. *Subject Raising Verbs*: appeared, seemed, happened, began, continued, tended

C Figures

[2] This is known as "partial control," and is a diagnostic for (Landau, 2024)'s logophoric control predicates.

[3] We acknowledge that three of our raising verbs are contentious: *begin* and *continue*, though they do appear as raising verbs in Postal (1974). There are instances of both appearing in the expletive construction (e.g., "It **continued** that the reserve would be 'a back-up solution only'" and "There **began** to be fewer men who paid taxes," both taken from Davies (2008–)).

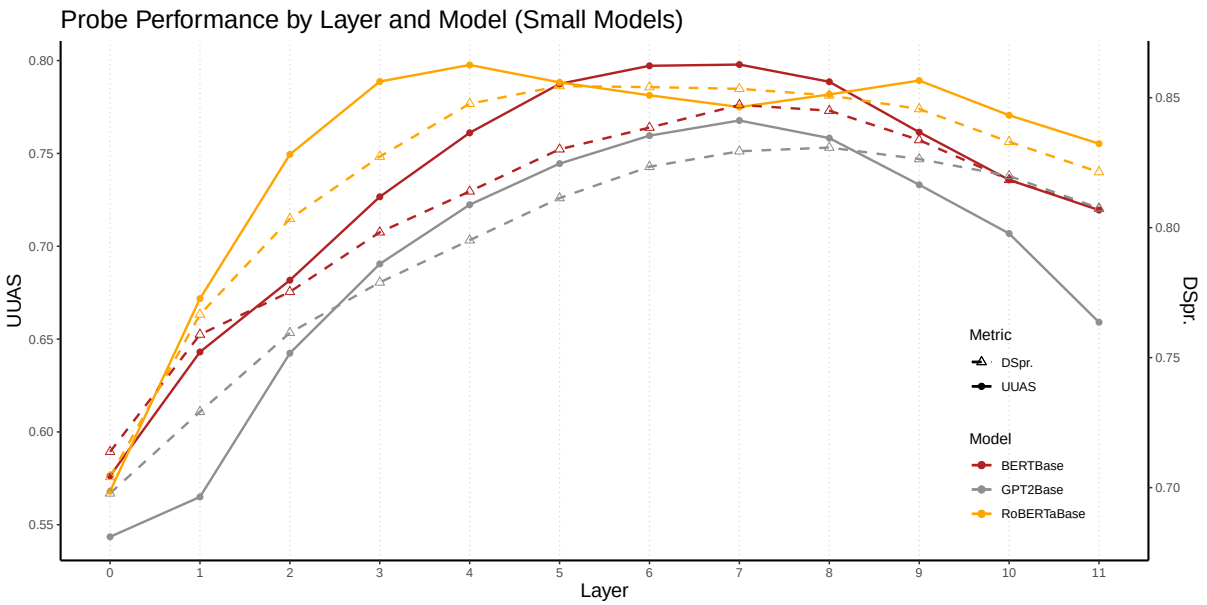

Figure 5: Probe performance for all small models. The solid lines are plotted against the left-hand y-axis and display the performance by Unlabeled Unattached Accuracy Score (UUAS) while the dotted lines plot the average Spearman correlation between the predicted and gold distances (DSpr.) along the right-hand y-axis. Highest-performing probes were `BERT-base-layer7`, `RoBERTa-base-layer4`, and `GPT2-layer7`.

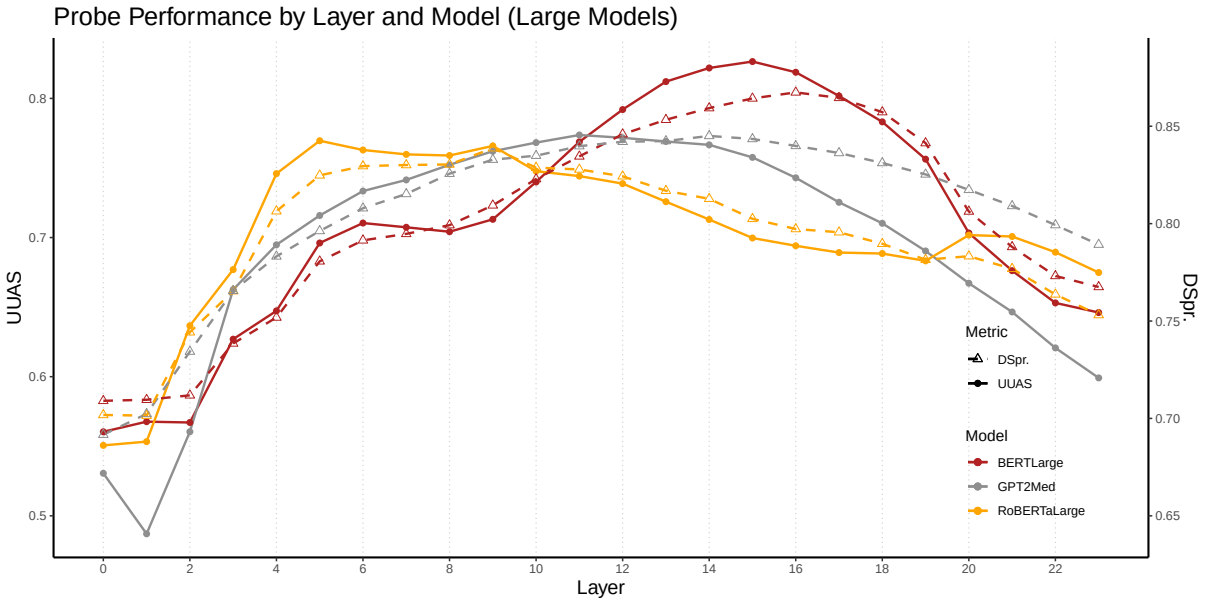

Figure 6: Probe performance for all of the larger models. The solid lines are plotted against the left-hand y-axis and display the performance by Unlabeled Unattached Accuracy Score (UUAS) while the dotted lines plot the average Spearman correlation between the predicted and gold distances (DSpr.) along the right-hand y-axis. Highest-performing probes were `BERT-large-layer15`, `RoBERTa-large-layer5`, and `GPT2-med-layer11`.

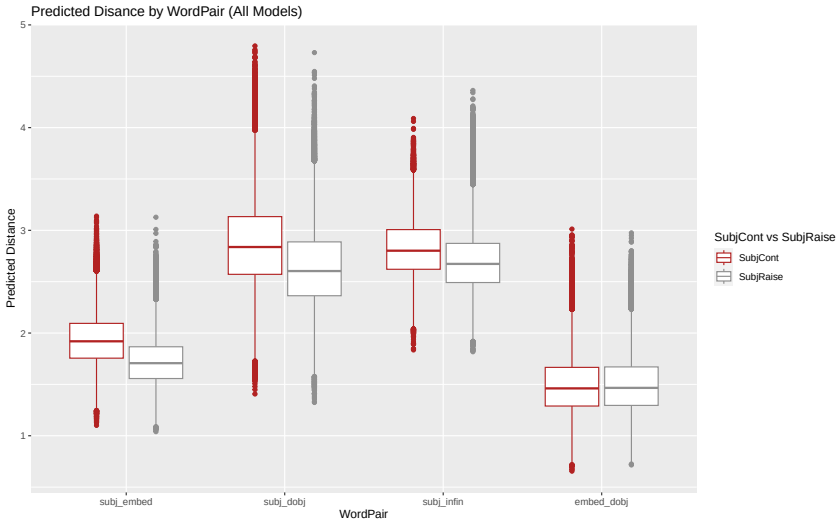

Figure 7: Predicted distances by WordPair for all LLMs. While the SC condition yields longer predicted distances than the SR condition, the baseline of `embed-dobj` shows no difference in the probes' predicted distance for the two conditions.

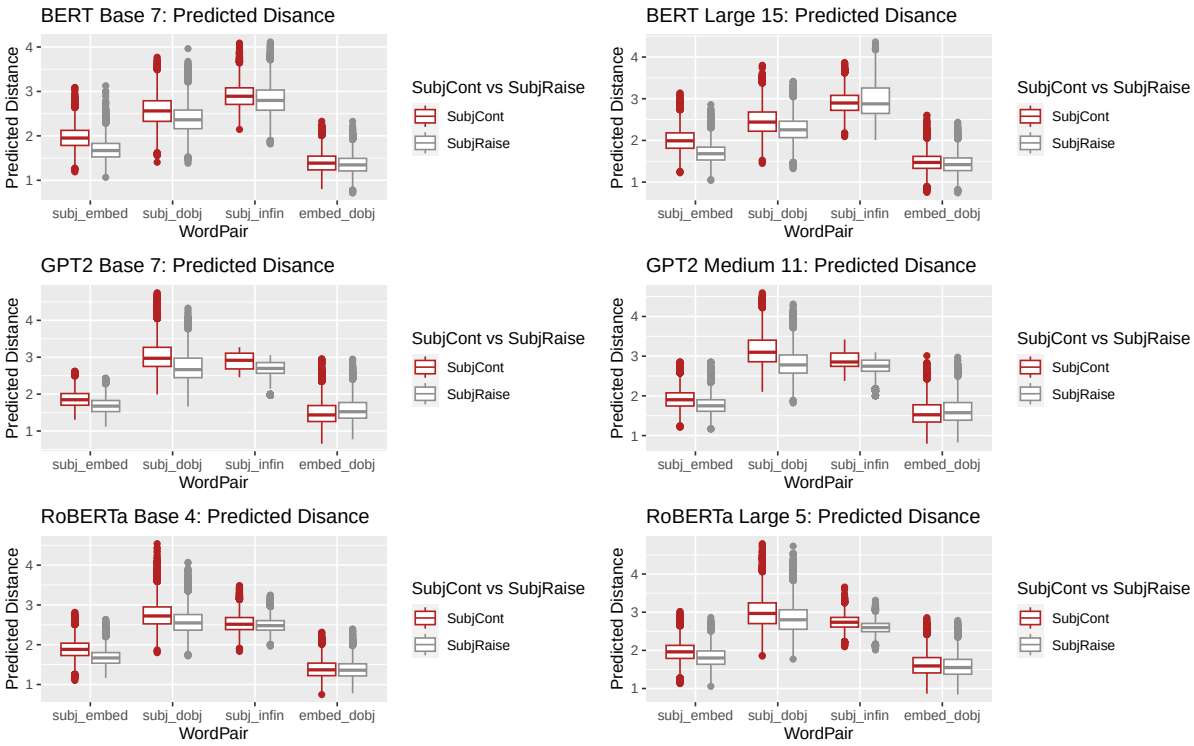

Figure 8: Predicted distances by WordPair and by LLM.