Large language models and linguistic recursion

Maksymilian Dąbkowski and Gašper Beguš

University of California, Berkeley / Project CETI

OVERVIEW

- we investigate the theoretical linguistic abilities of four LLMs:
  - OpenAI's GPT-3.5 Turbo (Brown et al., 2020),
  - OpenAI's GPT-4 (OpenAI, 2023),
  - Meta's Llama 3.1 (Dubey et al., 2024), and
  - OpenAI's o1 (OpenAI, 2024)
- OpenAI o1 outperforms other LLMs on some linguistic analysis tasks
  - able to generate coherent syntactic and phonological analyses (e. g. Chomsky, 2014; Chomsky and Halle, 1968)
- o1 may be the LLM with most advanced metalinguistic abilities—i. e. with abilities not only to use language, but also to reason about it

BACKGROUND

- previous work (e. g. Gulordava et al., 2018; Linzen et al., 2016; Matusevych et al., 2022; Wilcox et al., 2018; Yedetore et al., 2023) has tested the linguistic abilities of neural networks trained on text
- yet, most previous studies only test LLMs' correct language use;
  - not the models' ability to generate analyses of language data
- we test complex metalinguistic abilities of LLMs
  - results can provide insights into their metacognitive abilities

RESEARCH PROGRAM

- *behavioral interpretability* of deep learning — models' performance is evaluated through explicit metacognitive prompts
- transformers (seem to) represent language hierarchically in structures that resemble syntactic trees (Murty et al., 2022)
  - previously, claims evaluated implicitly by looking into the transformers' internal representations
  - applying a linguistic formalism to an LLM's own language ability — testing grounds for accessing its metacognitive abilities?
  - do solutions draw only on distributional knowledge, or also on an understanding of constituency, hierarchical structure, etc.?
- human linguists have arrived at a range of analytical frameworks by reasoning from their mental grammar
  - will LLMs be able to come up with innovative theoretical solutions that were not hypothesized by humans thus far?
  - suggestive: deep neural networks used in protein design (Jumper et al., 2021), geometry (Davies et al., 2021), and cracking unknown communication systems (Beguš, Leban, et al., 2023)

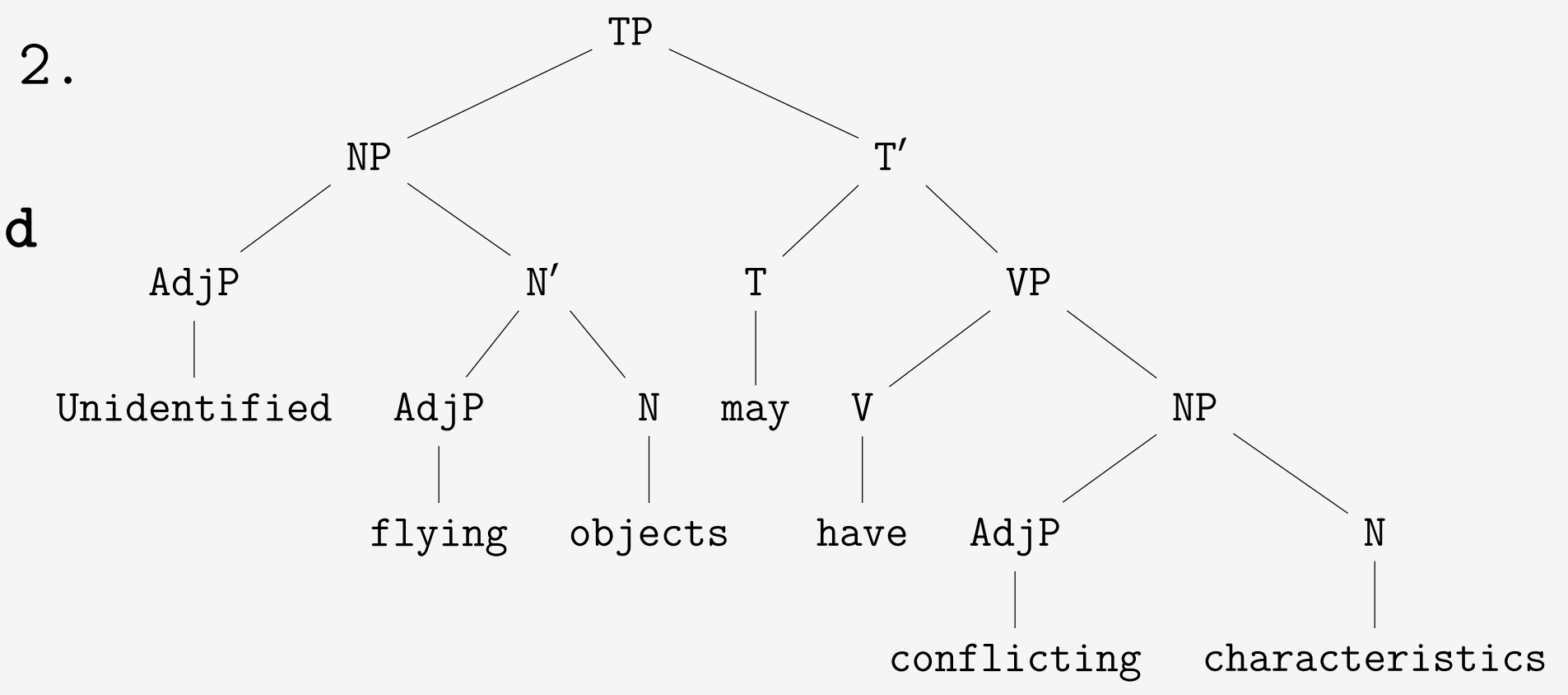
maksymilian gašper project CETI data

recursion task ex

```
Consider the sentence below and complete the following three tasks:
1. Does the sentence in question contain an instance of recursion? If so, identify
the recursive part and say what kind of recursion it is. Note that there are
different types of recursion, e.g. adjectival recursion (an adjective modifying
an adjective-modified noun), Saxon Genitive a.k.a. possessive 's recursion (an 's-
possessed 's-possessor), prepositional phrase recursion (a prepositional phrase
with another prepositional phrase inside), clausal recursion (a clause within
another clause), and so on.
2. Provide code for a syntactic tree representing the structure of the sentence
that can be rendered with LaTeX's forest package. Assume X-bar theory.
3. If you identified that the sentence contains a recursive structure, expand it by
adding another layer of recursion (of the same type) to the identified structure.
"Unidentified flying objects may have conflicting characteristics."
```

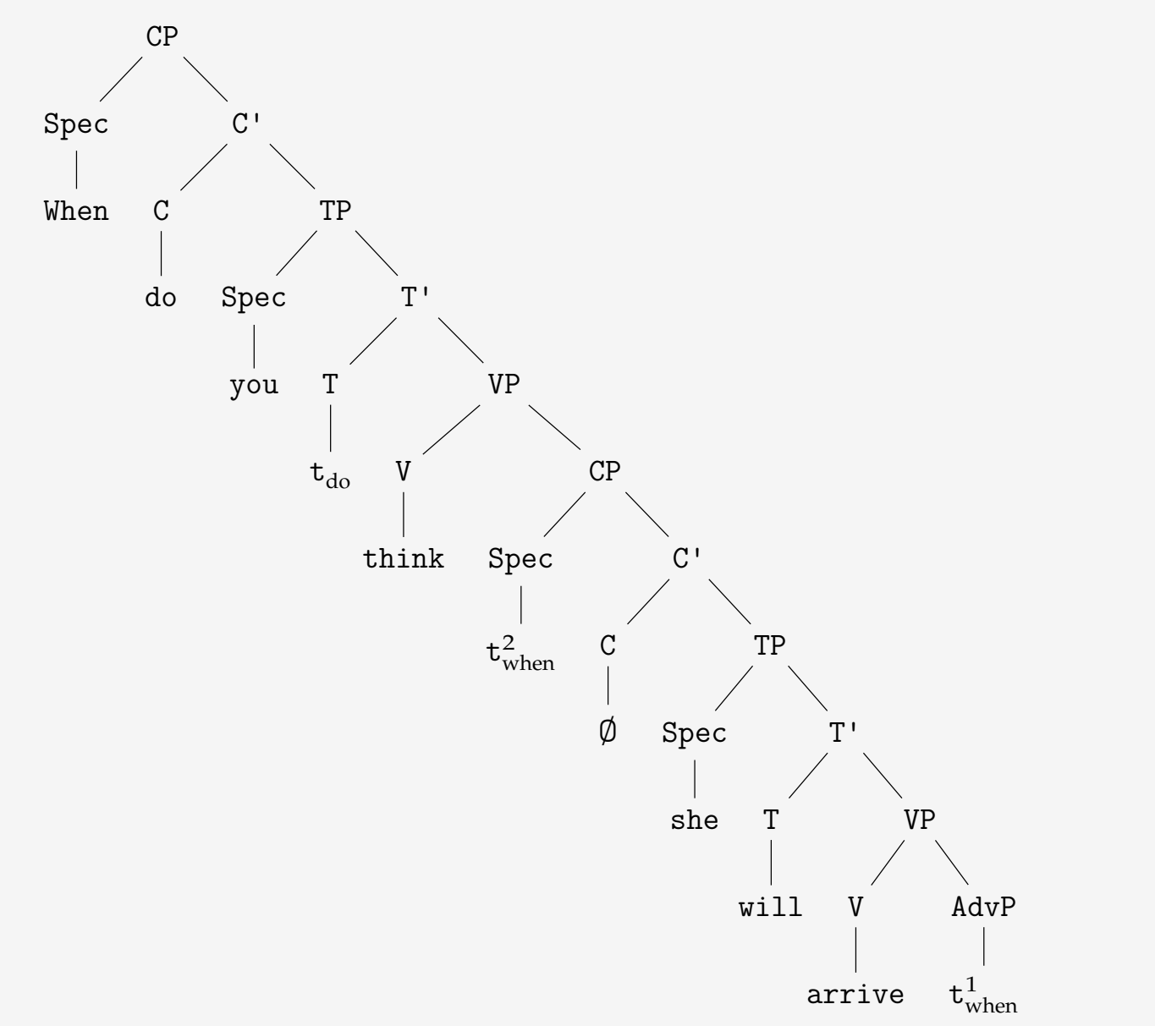
```
1. Yes, the sentence contains an
instance of recursion. The recursive
part is the noun phrase "Unidentified
flying objects", which exhibits
adjectival recursion.
3. "Unidentified recently sighted
flying objects may have conflicting
characteristics."
```

2.



movement task ex

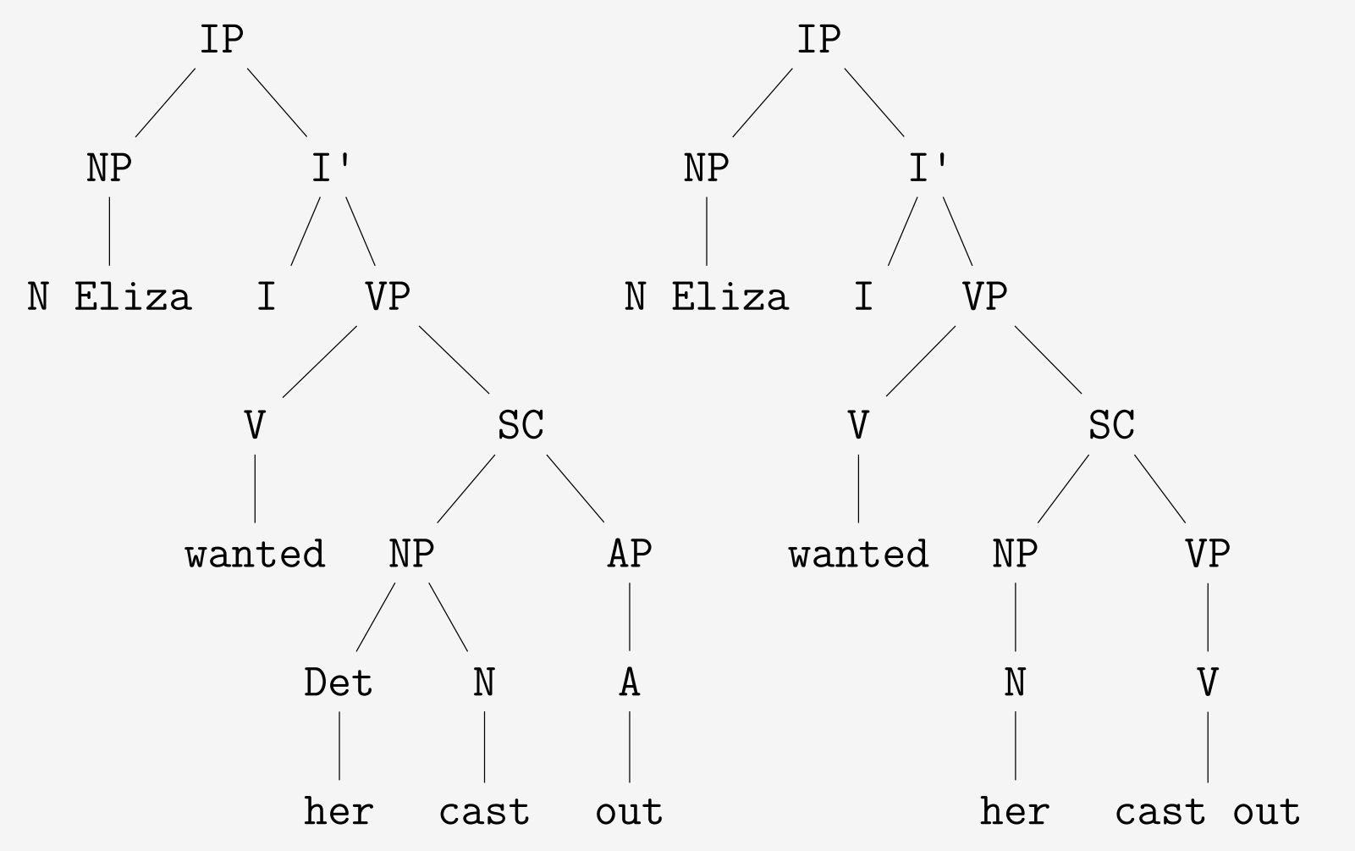
```
Consider the sentence below. Provide
code for a syntactic tree representing
its structure that can be rendered
with LaTeX's forest package. Assume X-
bar theory. Pay special attention to
syntactic movement and represent all
instances of movement with (indexed)
traces. ...
"When do you think she will arrive?"
```



ambiguity task ex

```
Consider the sentence below. Is it
syntactically ambiguous? If so, provide
code for syntactic trees representing
each possible parse of the sentence
that can be rendered with LaTeX's
forest package. Assume X-bar theory.
"Eliza wanted her cast out."
```

```
Yes, the sentence "Eliza wanted her
cast out." is syntactically ambiguous.
```



phonology task ex

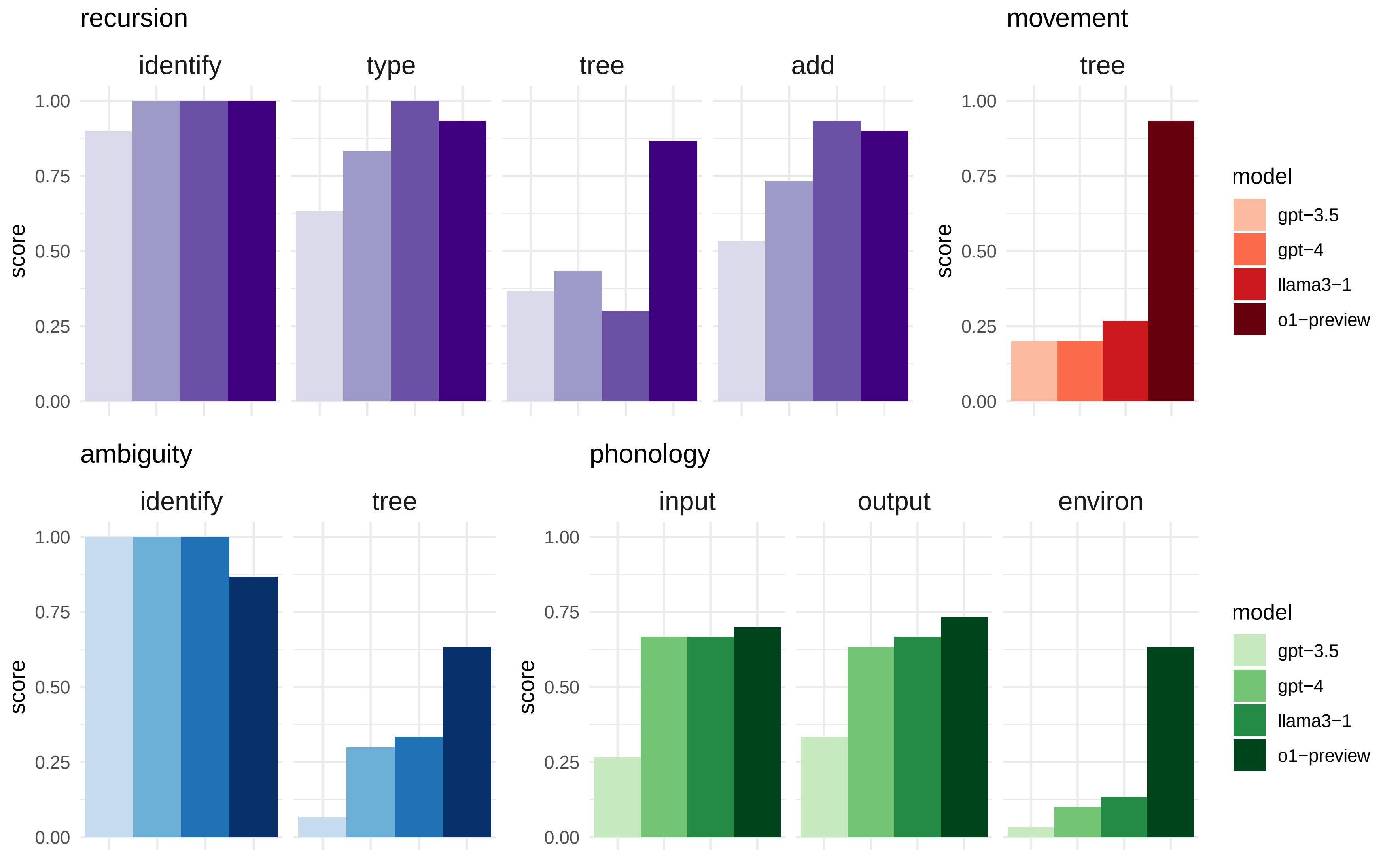
```
Below, you are given a list of 40 words, which are the surface forms in a language
you have not previously encountered. Each surface form is given as a string of
phones separated by spaces. What phonological process operating in this language
can you observe? Focus on the following phones of interest: t d n s z r l t̪ d̪ n̪
θ ð r̪ l̪. State the phonological process as a rule such as e.g. A → B / _ C, A →
B / C _, or A → B / C _ D, where A stands for the underlying phoneme, B for its
surface realization, and C (and D) specify the environment where A is realized
as B. State your rule in the most general possible terms, i.e. refer to natural
classes whenever possible (instead of simply listing the affected phonemes).
h æ θ i l̪          h e n̪ w æ ð          s o t ɑ n          l u l æ          ...
```

```
... alveolar consonants becoming dental after front vowels.
Rule: Alveolar consonants → Dental / Front vowel ___
```

METHODS

- we designed four tasks aimed at testing the models' ability to:
  - identify recursive structures, determine their type, represent them with tree diagrams, and add other layers of recursion,
  - represent syntactic movement with traces,
  - identify ambiguity and represent it with syntactic trees, and
  - write phonological rules specifying the input, output, and environment of their application
- within each task, each model was evaluated on 30 test items / each test item contained an original English test sentence or a constructed phonological dataset
- subtasks of each task was independently graded by three linguistics graduate students as either correct or incorrect; the majority grade was counted

RESULTS



- on many tasks, the performance of all of four models was comparable
- on the more difficult tasks, o1 vastly outperforms other models

TREE DRAWING: o1 scores 0.63–0.93; GPT-4 and Llama 3.1 score ~0.3

RULE ENVIRONMENT: o1 scores 0.63; GPT-4 and Llama 3.1 score below 0.14

DISCUSSION

- our line of work (e. g. Beguš, Dąbkowski, et al., 2023) — the first to show that LLMs can analyze sentence structure in a metalinguistic way, and explicitly solve complex tasks, such as representing recursive structure with syntactic trees
- we speculate that o1's unique advantage in solving linguistic puzzles may result from the model's *chain-of-thought* mechanism, which mimics the structure of human reasoning used in complex cognitive tasks

Beguš, G., M. Dąbkowski, et al. (2023). *Large linguistic models: Analyzing theoretical linguistic abilities of LLMs*. arXiv: 2305.00948 [cs.CL]. Beguš, G., A. Leban, et al. (2023). "Approaching an unknown communication system by latent space exploration and causal inference". In: *Arxiv* 2303.10931. eprint: 2303.10931 (stat.ML). Brown, T. B. et al. (2020). *Language models are few-shot learners*. arXiv: 2005.14165 [cs.CL]. Chomsky, N. (2014). *The minimalist program*. MIT press. Chomsky, N. and M. Halle (1968). *The sound pattern of English*. New York: Harper & Row. Davies, A. et al. (2021). "Advancing mathematics by guiding human intuition with ai". In: *Nature* 600.7887, pp. 70–74. DOI: 10.1038/s41586-021-04086-x. URL: https://doi.org/10.1038/s41586-021-04086-x. Dubey, A. et al. (2024). *The llama 3 herd of models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783. Gulordava, K. et al. (June 2018). "Colorless green recurrent networks dream hierarchically". In: *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1195–1205. DOI: 10.18653/v1/N18-1108. URL: https://aclanthology.org/N18-1108. Jumper, J. et al. (2021). "Highly accurate protein structure prediction with alphafold". In: *Nature* 596.7873, pp. 583–589. DOI: 10.1038/s41586-021-03819-2. URL: https://doi.org/10.1038/s41586-021-03819-2. Linzen, T. et al. (2016). "Assessing the ability of LSTMs to learn syntax-sensitive dependencies". In: *Transactions of the association for computational linguistics* 4, pp. 521–535. DOI: 10.1162/tacl_a_00115. URL: https://aclanthology.org/Q16-1037. Matusevych, Y. et al. (2022). "*Trees neural those*: rnns can learn the hierarchical structure of noun phrases". In: *Proceedings of the annual meeting of the cognitive science society* 44.44, pp. 1848–1854. Murty, S. et al. (2022). *Characterizing intrinsic compositionality in transformers with tree projections*. arXiv: 2211.01288 [cs.CL]. OpenAI (2023). *Gpt-4 technical report*. arXiv: 2303.08774 [cs.CL]. OpenAI (2024). *Openai o1 system card*. URL: https://cdn.openai.com/o1-system-card-20240917.pdf. Wilcox, E. et al. (Nov. 2018). "What do RNN language models learn about filler–gap dependencies?" In: *Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 211–221. DOI: 10.18653/v1/W18-5423. URL: https://aclanthology.org/W18-5423. Yedetore, A. et al. (2023). "How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech". In: *Arxiv preprint arxiv:2301.11462*.

LaTeX TikZposter