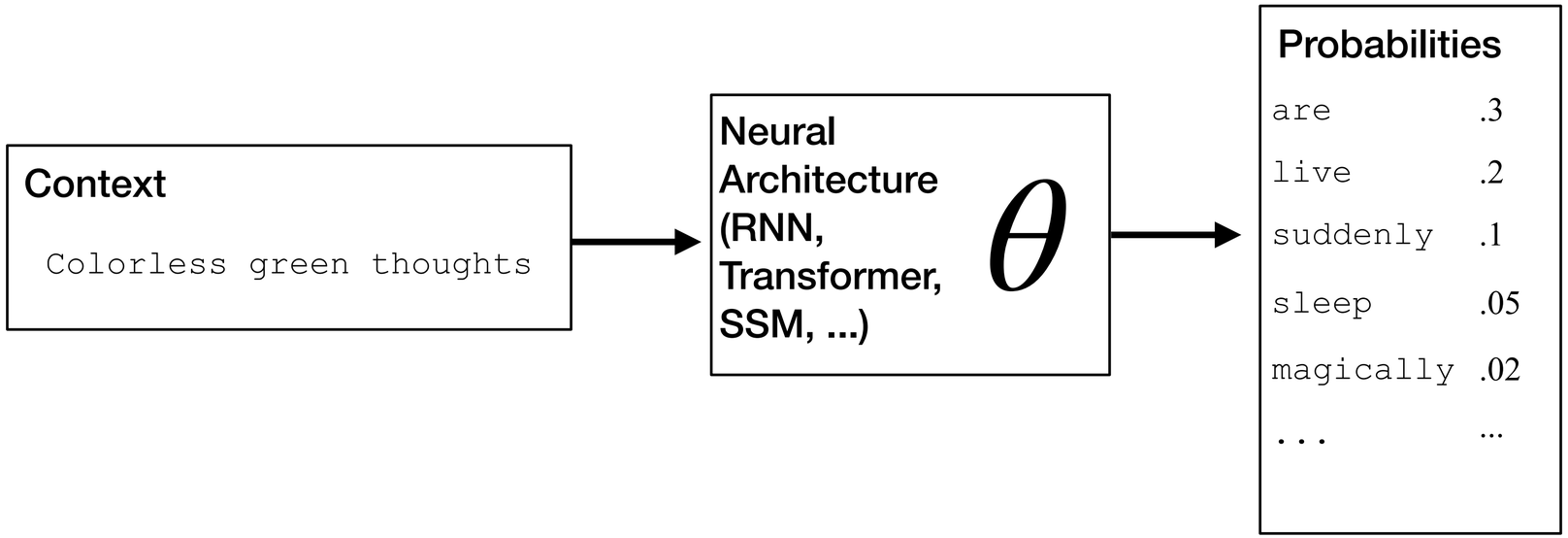
What can language models tell us about language?

Richard Futrell
Department of Language Science
University of California, Irvine
@rljfutrell

GLOWing Lecture
2026-01-23

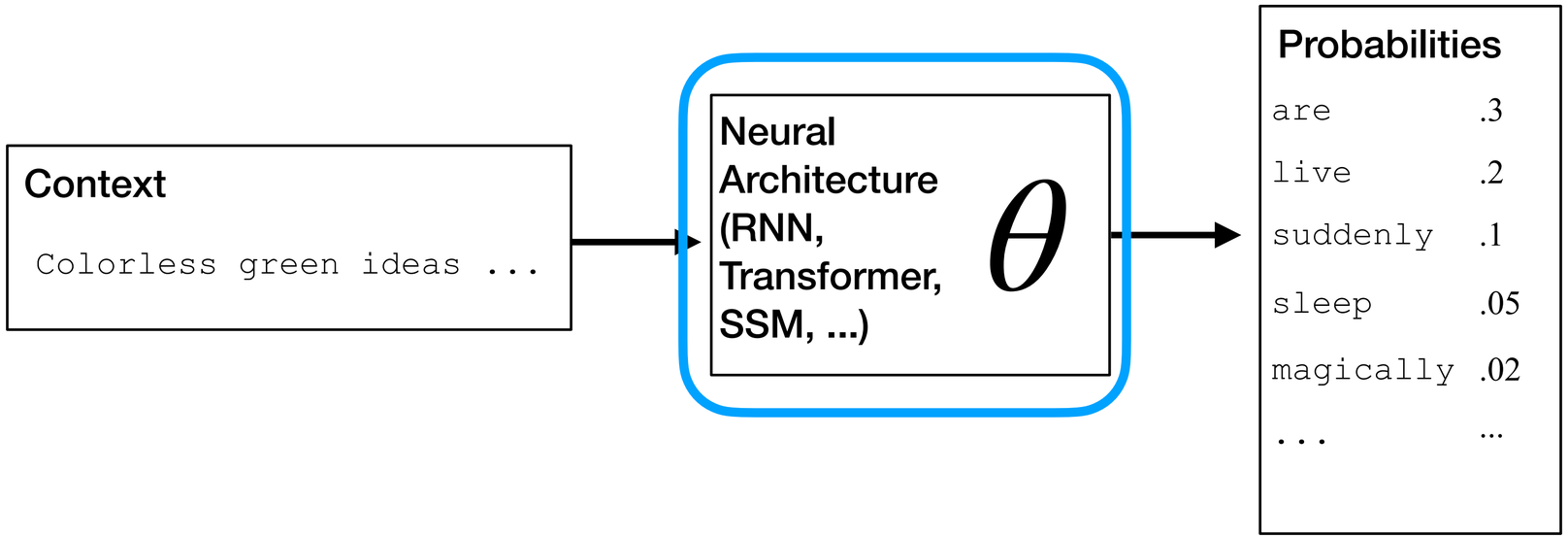
What do I mean by "language model"?

- Probabilistic models trained to have high predictive accuracy when predicting text data.



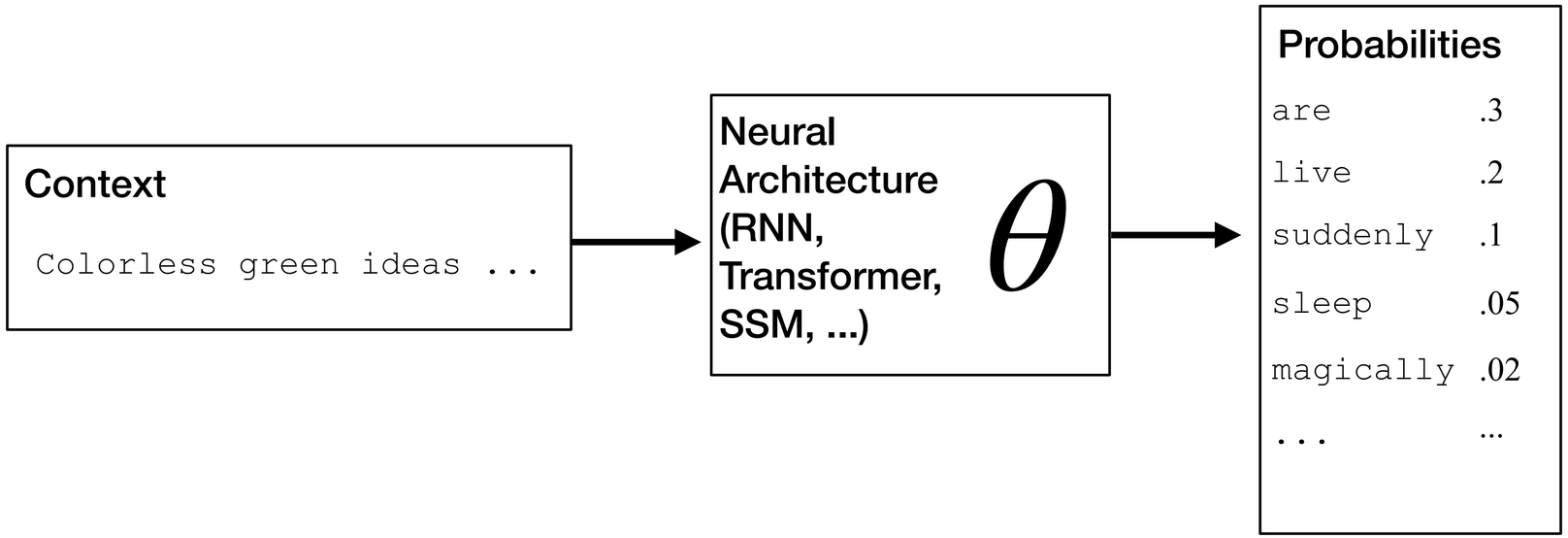
- I will discuss language models that are *unlike commercial systems like ChatGPT* ("frontier models"):
  - "Base models" (no additional training using reinforcement learning or fine-tuning)
  - When possible, open-weight, open-source, open-training-data
  - Substantially smaller than commercial systems (small enough to train/run on a laptop)
- Analysis of frontier models is interesting (e.g. Beguš et al. 2025 have shown that they can draw accurate syntactic trees for some challenging sentences), but not my focus here.

What do I mean by "language model"?



- The neural architecture is *highly flexible* (not just an *n*-gram model!)
  - For example, *in principle*, many architectures can implement a recognizer for context-free languages (Korsky & Berwick, 2019)
  - They *can* perform complex computations based on highly abstract features
- The question is what they *actually do* when trained to predict text.

What do I mean by "language model"?



- These are the *only known artificial systems* that can process natural language to the level of being able to have nontrivial interactions with humans.

Topics

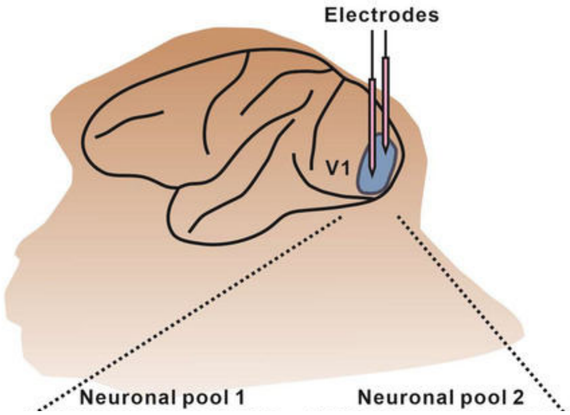
- What can such models tell us in principle?
- Evidence for Linguistic Structure in LMs
- Learning and Representation
- Conclusion

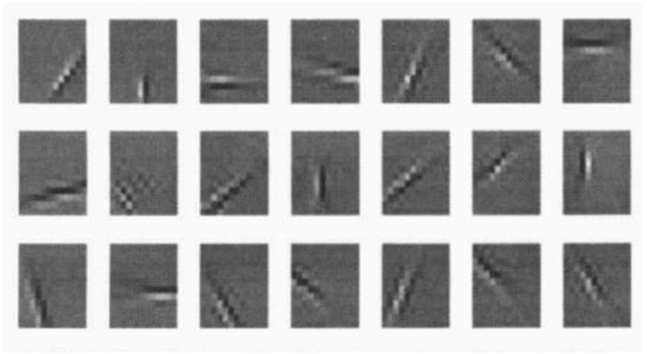
Example: Neural Networks in Vision

- Neural networks are *loosely* inspired by brain architecture.
- *Perhaps*: LMs are just not like the human brain, so they are irrelevant to human linguistic cognition, or no more relevant than airplanes (engineering artifact that flies) are to ornithology (study of biological organisms that fly).
- *Counterexample:* Despite these limitations, neural networks have played a central role in our developing understanding of *visual cognition*.

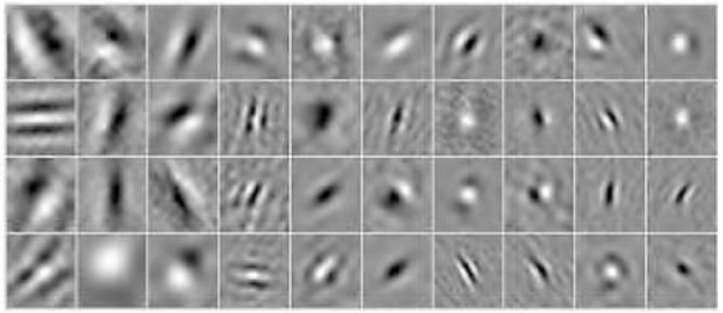
Visual Cognition

Neural Networks in Vision

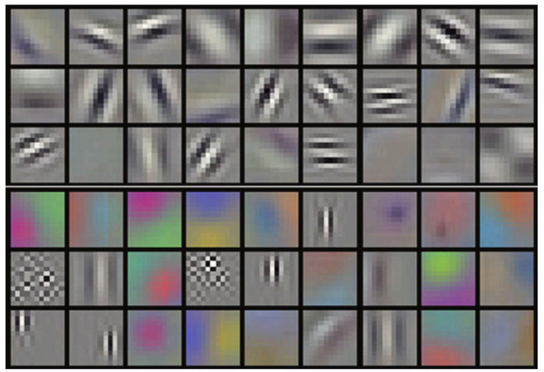




Receptive fields derived by training neural networks to predict natural images (Bell et al., 1997)



Receptive fields in macaque early visual cortex (Zylberberg et al., 2011)



First-layer "receptive fields" in AlexNet, trained on image classification (Krizhevsky et al., 2012)

Cao & Yamins (2024)

Vision Example

- Neural models contributed to an *explanatory* account of edge detection in vision
  - Edge detection is the result of (1) the function of vision, (2) the statistics of visual input, (3) general principles of efficient information processing.
  - They did so despite not capturing all of visual cognition (Bowers et al., 2022 BBS).
- We have argued that LMs can contribute to linguistics similarly (Futrell & Mahowald, 2026 [to appear in BBS]).
  - Not as replacements or proxies for models of linguistic cognition
  - Rather as **comparative systems** where we are conscious that they share some properties with humans but not others
  - Here I will be presenting *some pieces* of the larger argument in that paper.

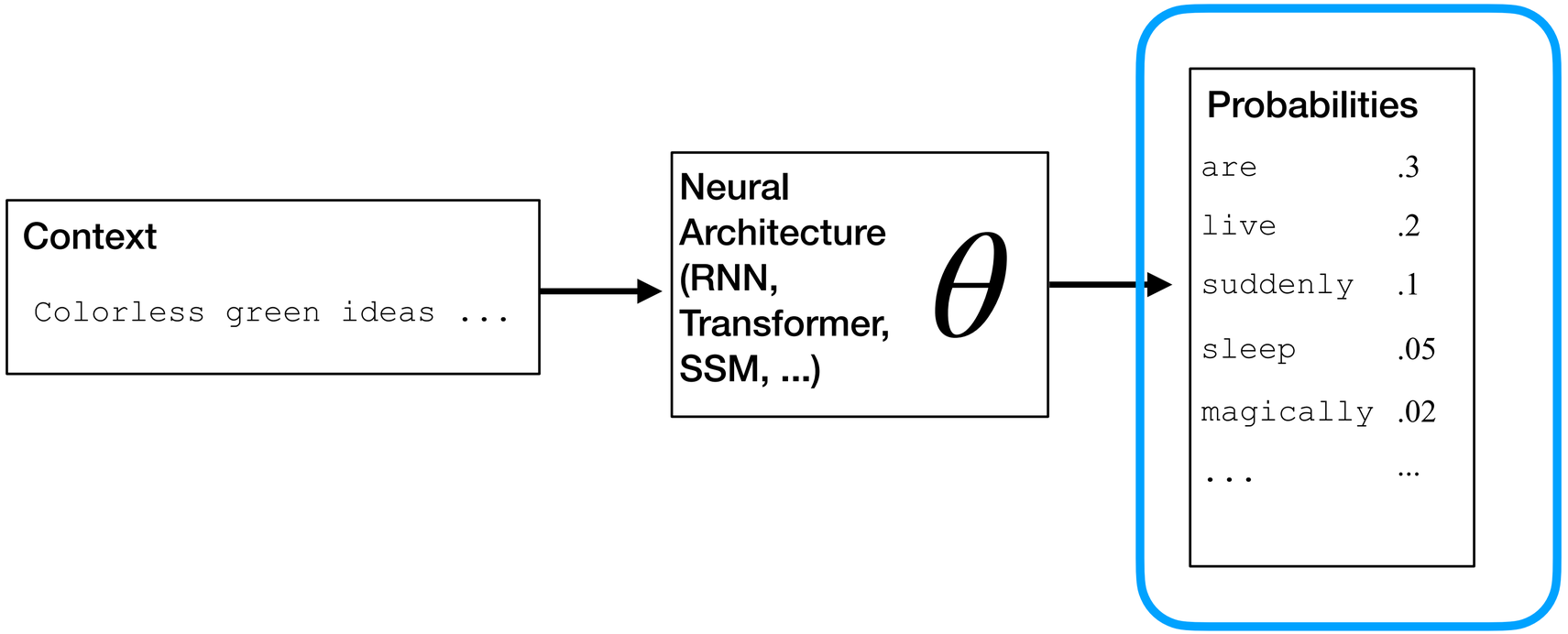
Key Positions

- LMs do not replace or supplant linguistic theories (*contra* Piantadosi, 2023), which provides the best known formal characterization of human linguistic competence.
- But they do inform questions of linguistic interest, by serving as systems that
  - 1. Demonstrate what is possible in a system not limited to that characterization.
  - 2. Generate hypotheses for neural representation of linguistic structures.
  - 3. Demonstate ways of thinking about *learning* and *representation* that might be new to formal linguists.

Topics

- What can such models tell us in principle?
- Evidence for Linguistic Structure in LMs
- Learning and Representation
- Conclusion

Behavioral Assessment of Linguistic Structure



Grammaticality is not Probability

(1) Colorless green ideas sleep furiously. → GPT-2 → $p \sim 2^{-90}$

(2) * Furiously sleep ideas green colorless. → GPT-2 → $p \sim 2^{-93}$

~~P(some grammatical sentence) > P(some ungrammatical string)?~~

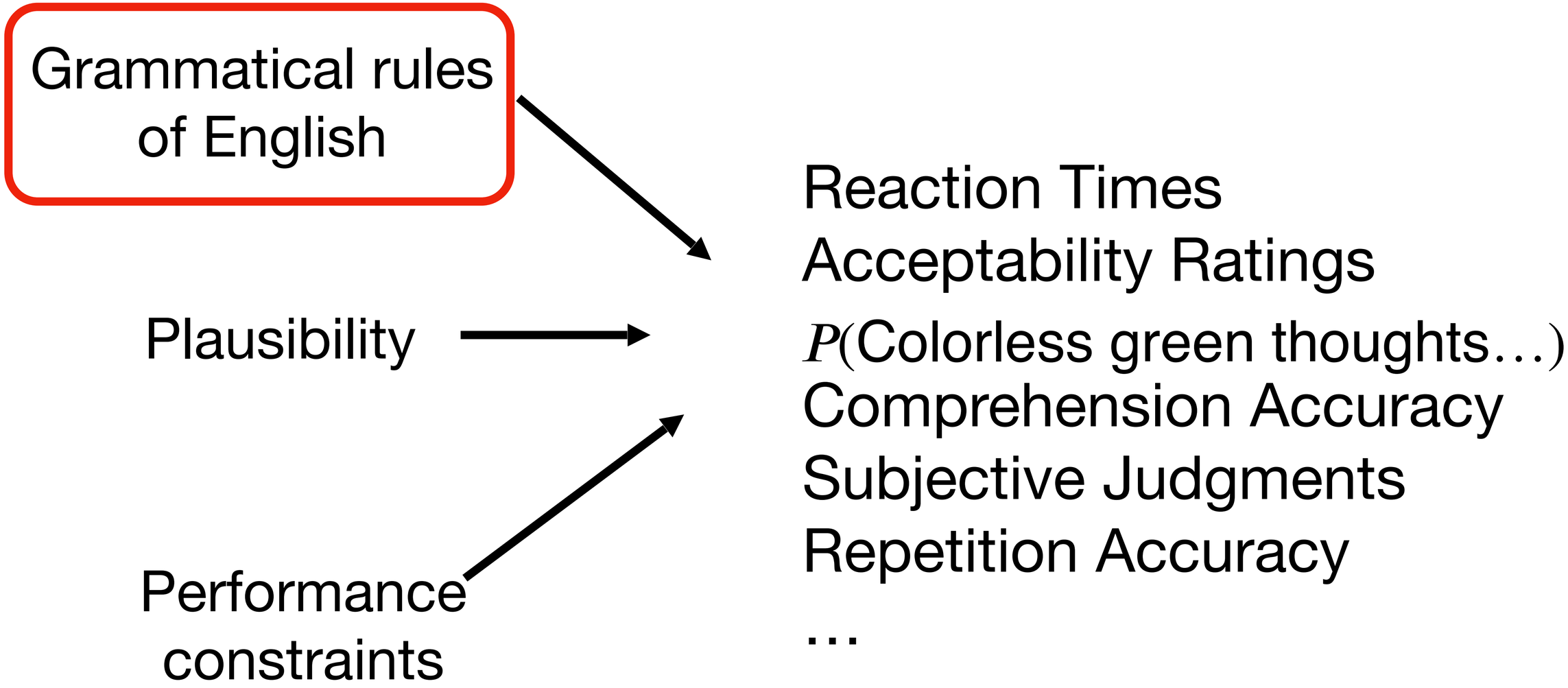
* Snails died the old. → GPT-2 → $p \sim 2^{-49}$

The ancient crustaceans expired. → GPT-2 → $p \sim 2^{-55}$

"Flavorless sour thoughts dream angrily"
vs. *"Angrily dream thoughts sour flavorless"

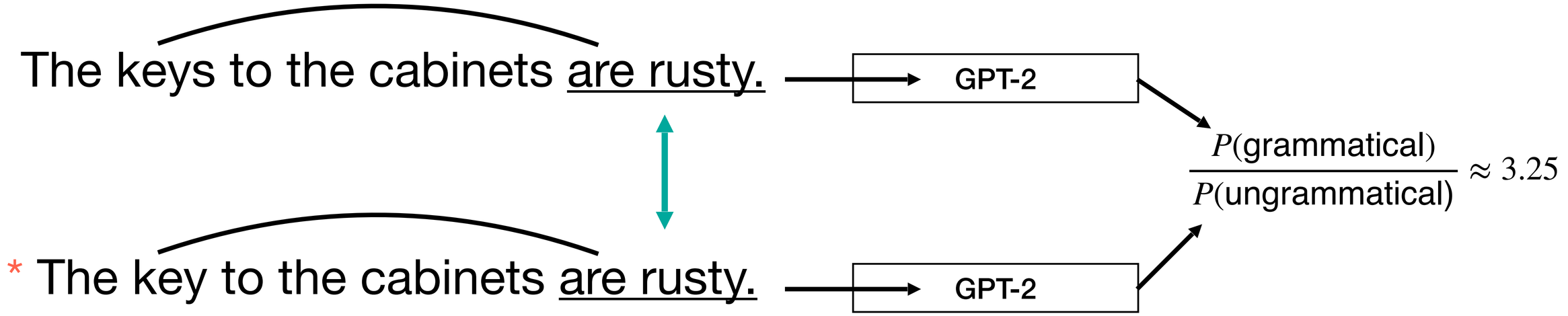
Chomsky (1957); Wilcox, Futrell & Levy (2024)

But Probability Gives Evidence for Grammaticality



Wilcox, Futrell & Levy (2024)

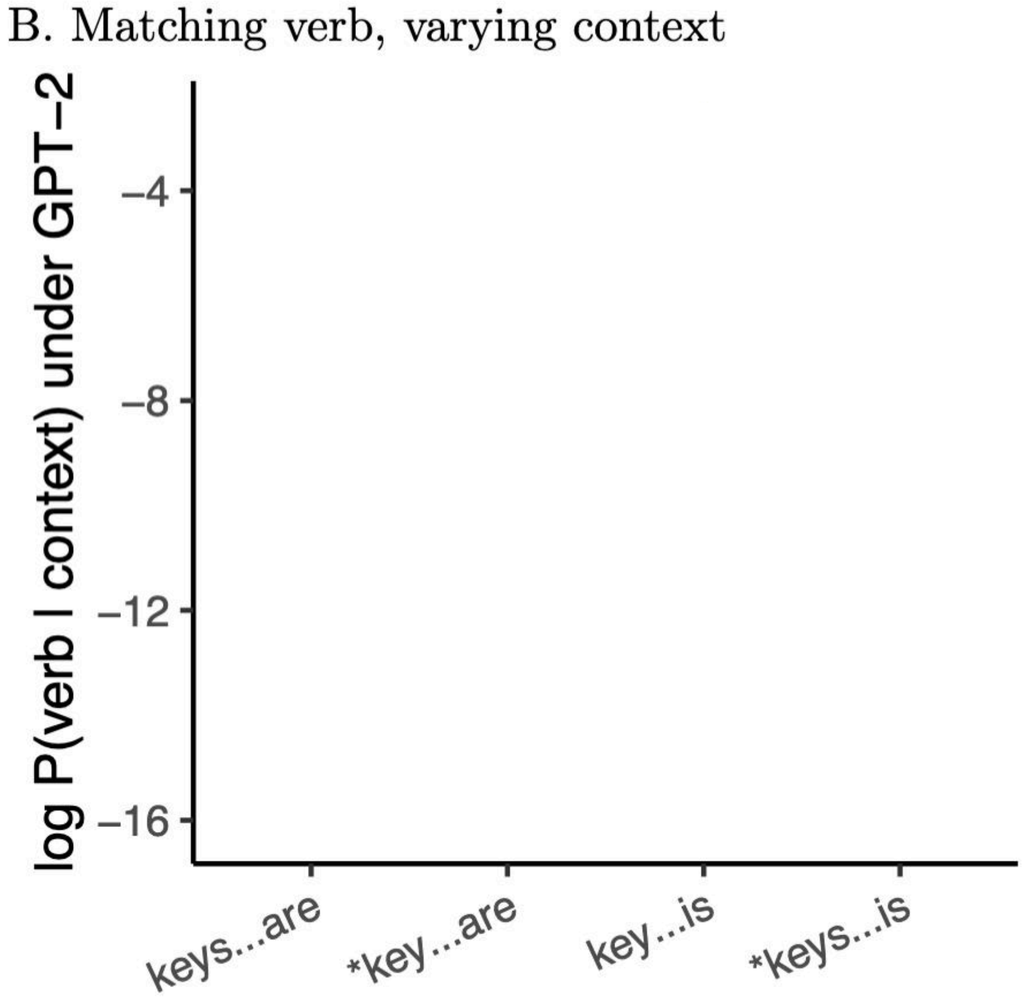
Evaluation Using Minimal Pairs



Linzen et al. (2016), Futrell et al. (2019), Hu et al. (2020); Wilcox, Futrell & Levy (2024)

Example Results

Sentences like: The {key/keys} to the old cabinets {is/are} ...



- There is no global probability threshold for grammaticality, nor would we expect there to be according to probability theory (Hu, Wilcox, Song, Mahowald & Levy, 2026 *TACL*)
- Nevertheless, the ungrammatical paired sentences get systematically lower probability.
- We use a large number of different sentences so that the idiosyncratic properties of individual sentences for the LM wash out.

Marvin & Linzen (2018); Hu et al. (2020); Gauthier et al. (2020); Futrell & Mahowald (2026)

Targeted Syntactic Evaluation: Wh-Dependencies

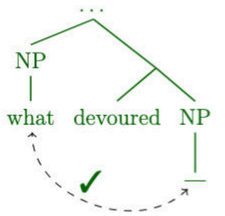
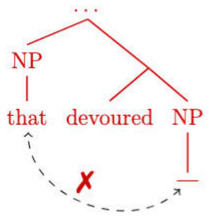
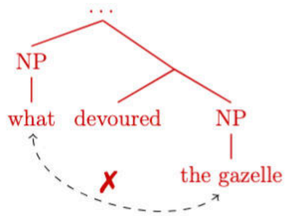
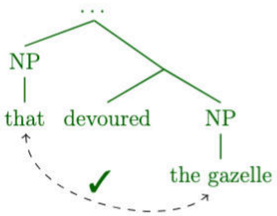
	What (+Filler)	That (-Filler)
+gap	 I know **what** the lion devoured ___ yesterday.	 *I know **that** the lion devoured ___ yesterday.
-gap	 *I know **what** the lion devoured **the gazelle** yesterday.	 I know **that** the lion devoured **the gazelle** yesterday.

Figure 1: Schematic demonstrating our 2 × 2 interaction design for measuring the filler–gap dependency. The portion of the sentence in which we measure surprisal is underlined.

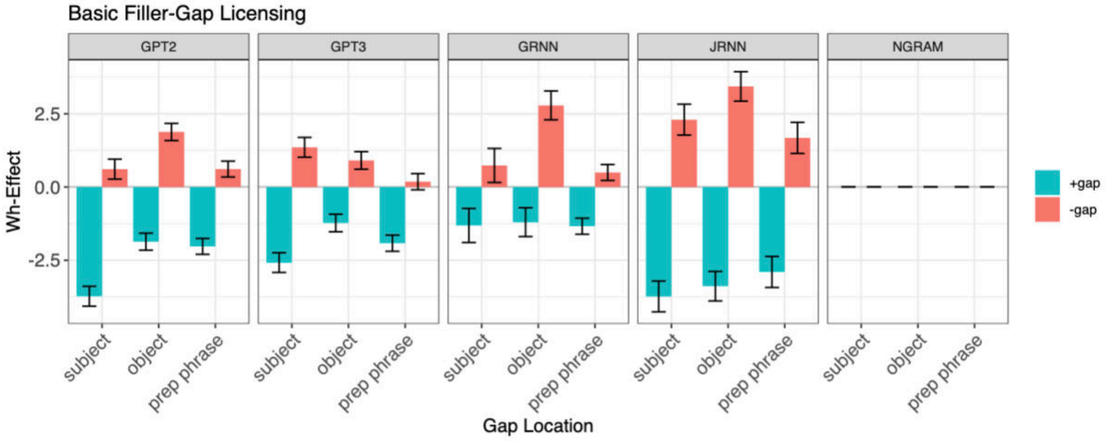


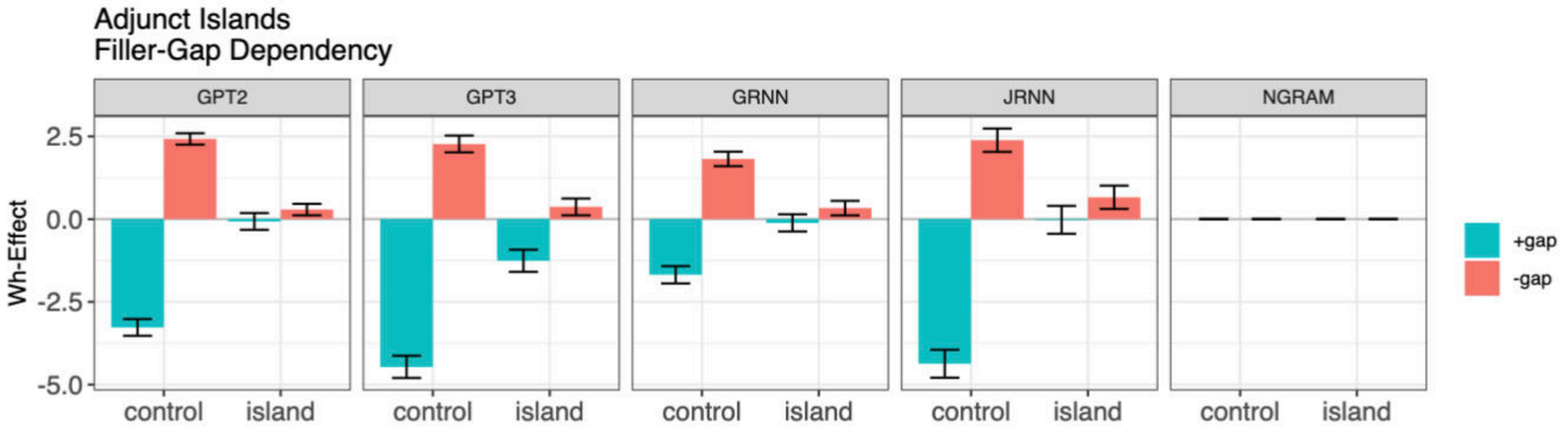
Figure 4: Basic Licensing. If models are learning the filler–gap dependency, we expect negative wh-effect in the *+gap* condition (blue bars) and a positive wh-effect in the *-gap* condition (red bars).

Wilcox, Futrell & Levy (2024)

Targeted Syntactic Evaluation: Island Constraints

control: I know what the librarian placed _ on the wrong shelf.
island: *I know what the patron got mad after the librarian placed _ on the wrong shelf.

Adjunct Islands



Wilcox, Futrell & Levy (2024)

Targeted Syntactic Evaluation: Island Constraints

control: I know what the librarian placed _ on the wrong shelf. island: *I know what the patron got mad after the librarian placed _ on the wrong shelf.	Adjunct Islands
control: I know what the actress bought _ yesterday. island: *I know what the actress bought the painting that depicted _ yesterday.	Complex NP Islands
control: I know what the man bought _ at the antique shop. island: *I know what the man bought _ and the painting at the antique shop.	Coordination Islands
control: I know how expensive a car you bought _ last week. island: *I know how expensive you bought _ car last week	Left Branch Islands
control: I know who the seniors defeated _ last week. island: *I know who for the seniors to defeat _ will be trivial.	Sentential Subject Islands
control: I know what _ fetched a high price. island: *I know who the painting by _ fetched a high price.	Subject Islands

- Not that-trace effects, not parasitic gaps

Wilcox, Futrell & Levy (2024)

Targeted Syntactic Evaluation: Island Constraints

- Our conclusion in Wilcox, Futrell & Levy (2024: 37): "Our tests reveal that these weakly biased models acquire impressively sophisticated generalizations regarding the filler-gap dependency and island constraints from even a childhood's quantity of linguistic input, though in some cases we find acquisition failures."
- Convergent with other modeling approaches showing that island constraints may be acquirable from non-language-specific learning principles (Pearl & Sprouse, 2013; Legate & Yang, 2024; Dickson, 2025)
- Lan, Chemla & Katzir (2024, LI) claim further failures on parasitic gaps and ATB movement.

Further Filler-Gap Complexities

- Lan, Chemla & Katzir (2024, LI) claim further failures on parasitic gaps and ATB movement.
- Our response:
  - 1. If there *were* failure here, even then it would not undermine our claim that the models learned the filler-gap depedency and island constraints.
    - Syntactic frameworks differ in the extent to which ATB and parasitic gaps involve different theoretical machinery beyond the basic *wh*-dependency.
    - Island constraints have famously gone through many different formal explanations (A-over-A constraint; subjacency; barriers; grammaticalized processing constraints). So we should be conservative when we want to claim that failure to learn one aspect of the *wh*-dependency weakens claims of learning another.

Further Filler-Gap Complexities

- Following up, Lan, Chemla, & Katzir (2024, LI) argue for further acquisition failures for ATB movement and parasitic gaps.
- Our response:
  - 2. Even then, the results are not obviously failures.
  - If these were reading time results, following psycholinguistic methods, we would conclude that there is evidence that the dependency.

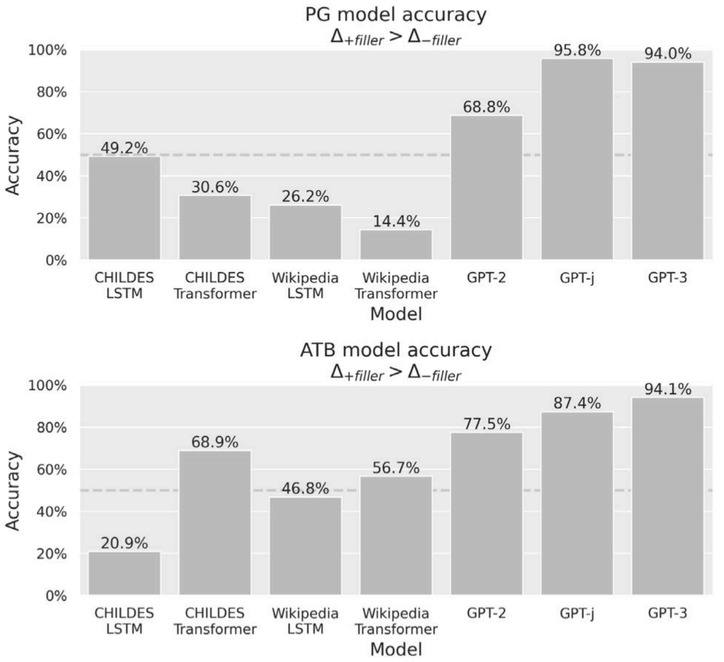
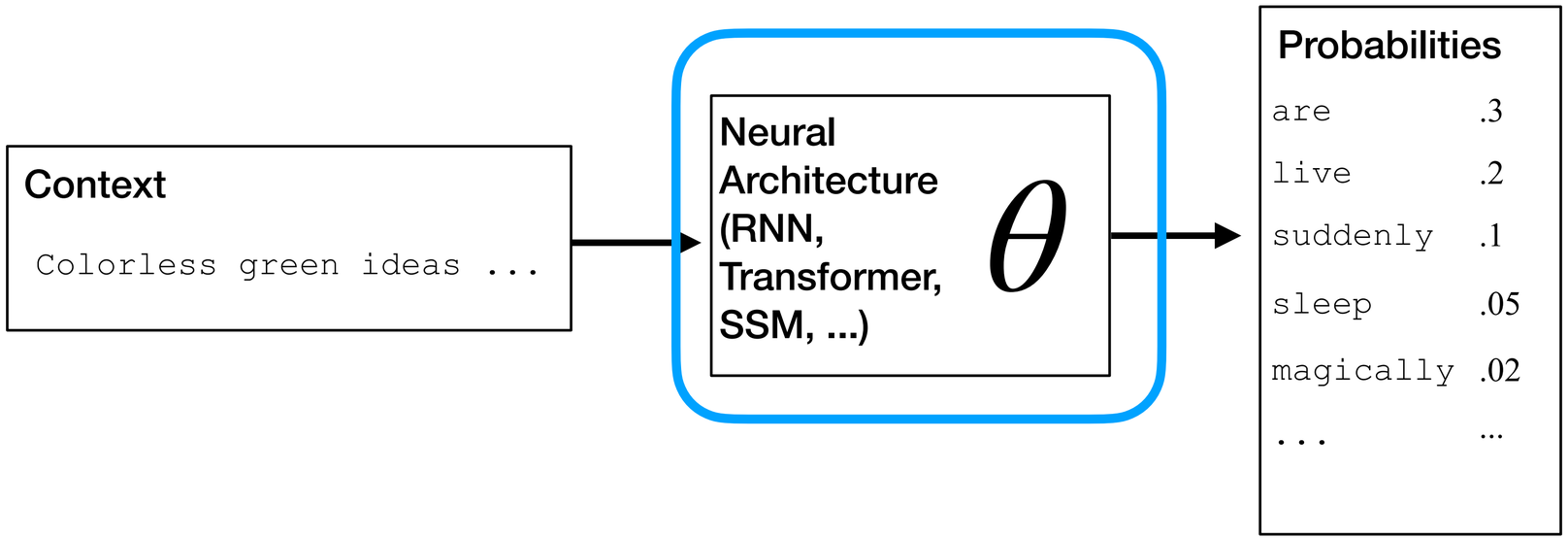


Figure 6
Model accuracy on the difference-in-differences condition for the parasitic gap (PG) and across-the-board (ATB) datasets. Accuracy is measured as the ratio of cases where $\Delta_{+filler} > \Delta_{-filler}$, that is, when the model shows a relative higher preference for a gap when the gap follows a filler than when it does not.

Internal Assessment of Linguistic Structure

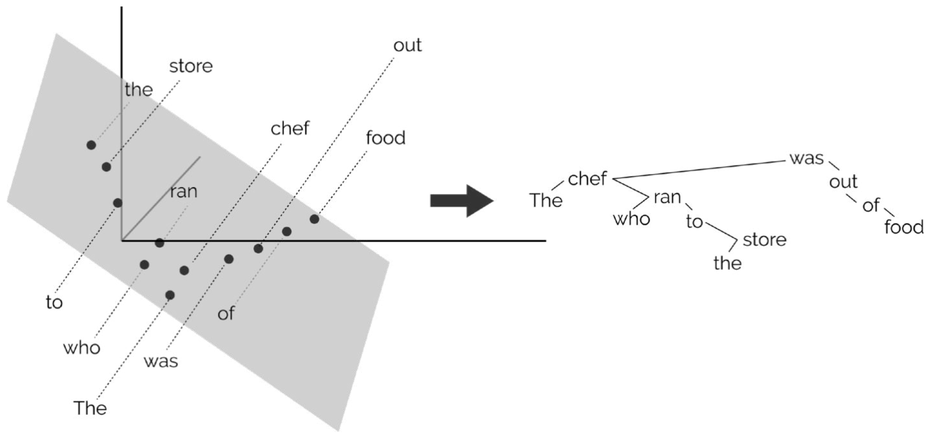


Why evaluate this way?

- Why not just ask the model whether something is grammatical and why?
  (Leivada et al., 2023; Beguš et al., 2025; ...)
  - 1. This probes only meta-linguistic awareness -- interesting but not the key question.
  - 2. LM explanations are not faithful (Madsen, Chandar & Reddy, 2024)
- There is a better way...

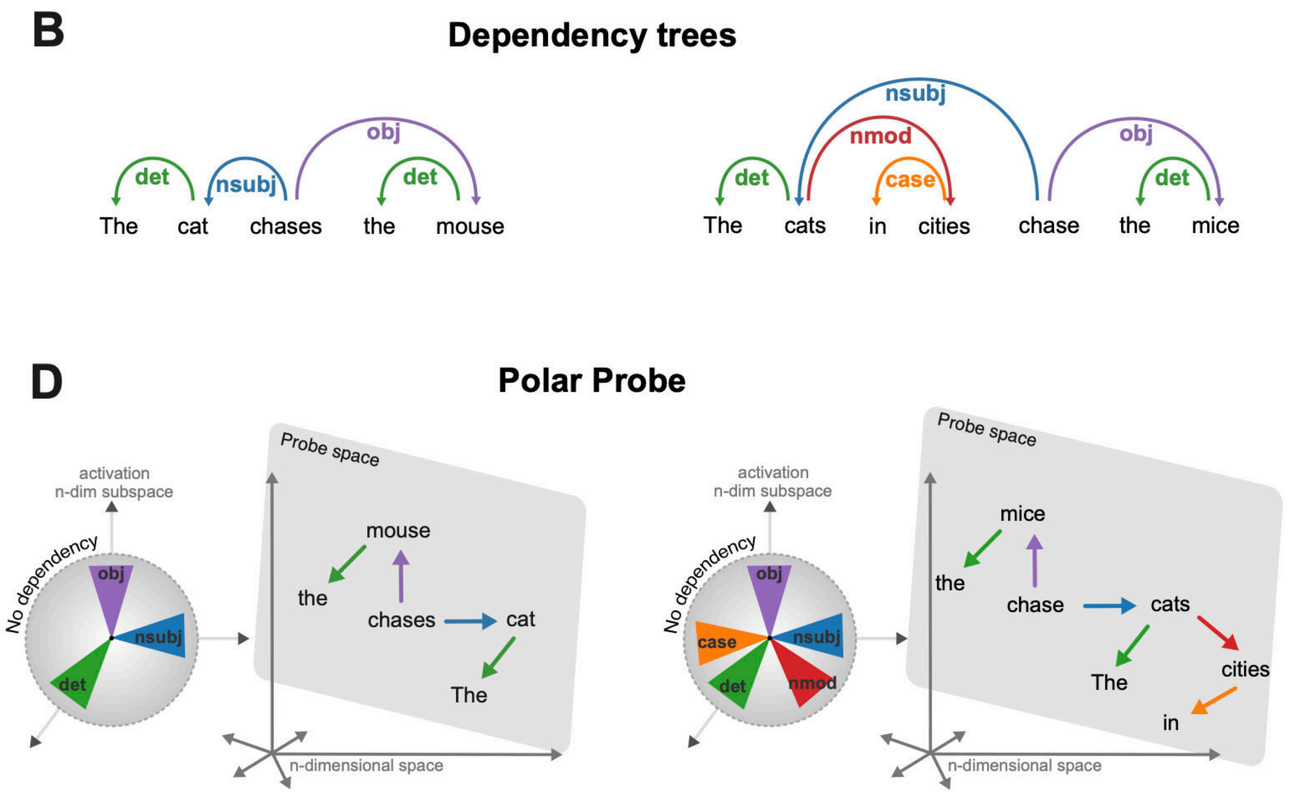
Probing for Syntactic Structure

- (Open) LMs are **glass boxes**: we can see their internal computations.
  - The question is how to interpret what we find.
  - **Mechanistic interpretability** has made large advances in understanding *why* models behave as they do.
- Syntactic relations and features are represented through geometric relations among high-dimensional vectors that represent words in context (Hewitt & Manning, 2020; Eisape et al., 2022; Diego-Simon et al. 2024, 2025; Arora et al., 2024; many others)



Syntactic Structure in LMs

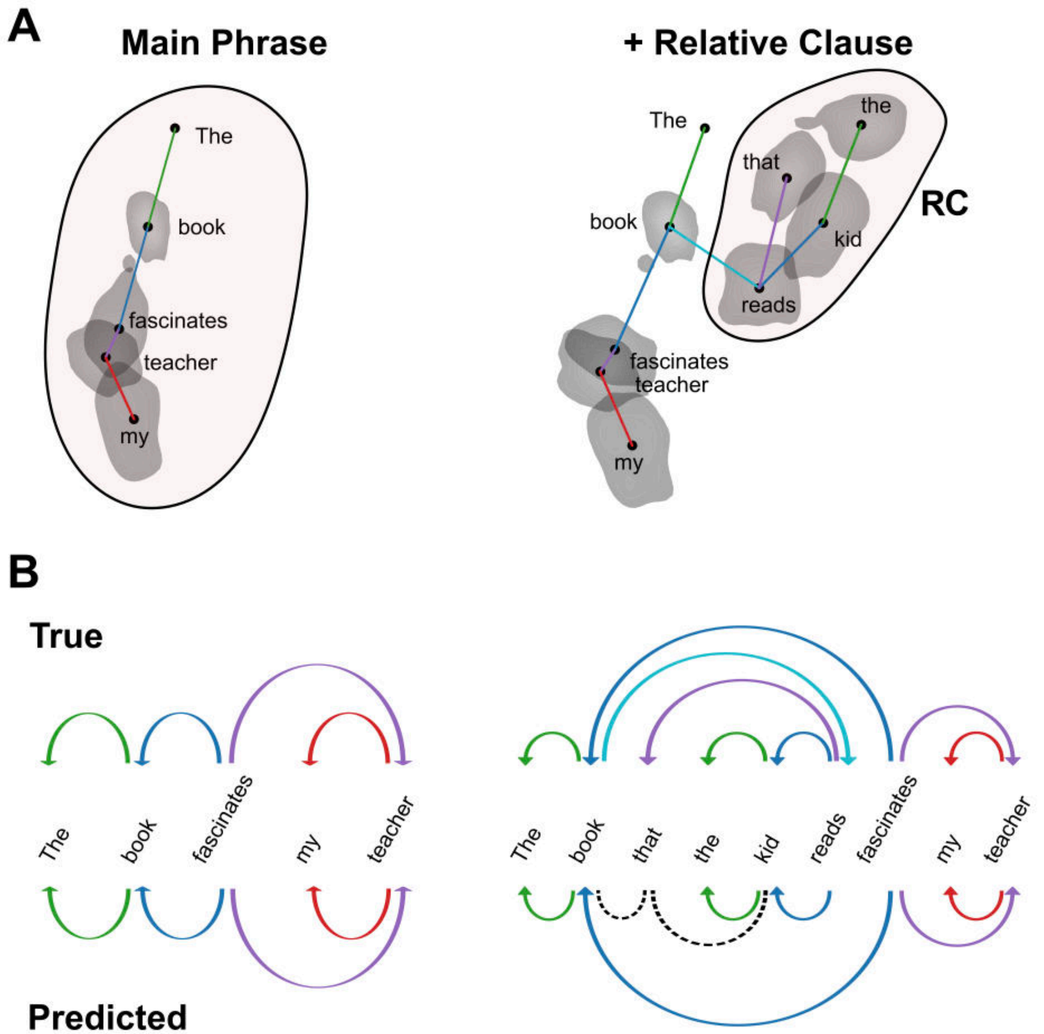
- For example, Diego-Simón et al (2024) look at syntactic dependency relations and find
  - Vectors for tokens linked in syntactic dependencies have certain geometric relationships with each other.
  - The *angle* between them encodes the *syntactic features* involved in the dependency.



Diego-Simón et al. (2024, *NeurIPS*)

Syntactic Structure in LMs

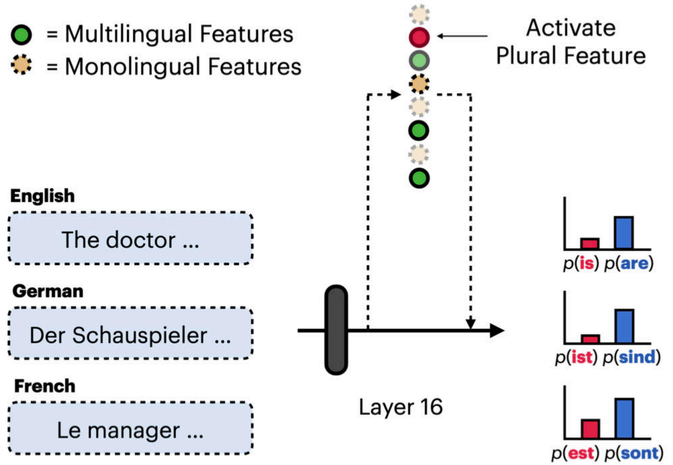
- The revealed parses are not always perfect.
- But they do capture nontrivial hierarchical and recursive structure.



Diego-Simón et al. (2024, *NeurIPS*)

Manipulating Syntactic Structure in LMs

- These representations of syntactic structure and features are **causally active** in determining what the model outputs, and they are **abstract**.
- For example: you can find the high-dimensional vector corresponding to the feature "plural" on a subject noun in English.
  - If you go into the network and add this vector in, you can change the output verb form.
  - *The same vector* also works to change verb forms in other languages!



Brinkmann et al. (2025, *ACL*)

Syntactic Structure in LMs

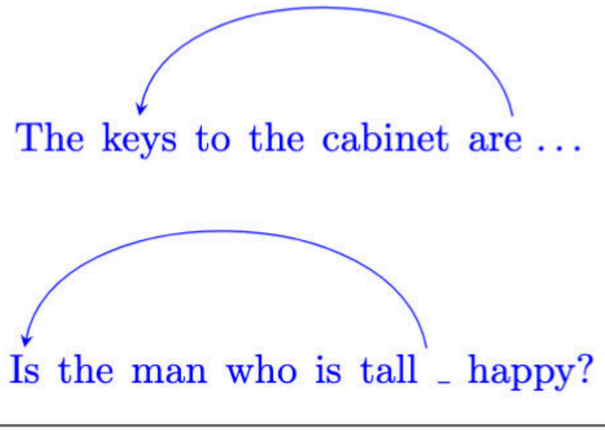
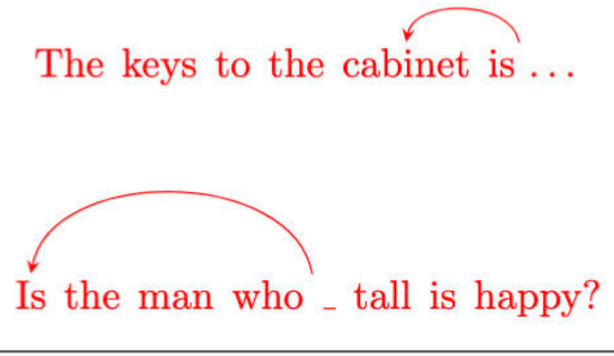
- Neural LMs learn enough abstract, hierarchical linguistic structure to be interesting comparative systems for linguistics.
  - Further experiments and probes are needed, and syntacticians have much to contribute!
- All the mechanisms I described are viable hypotheses for how these structures are implemented in the brain, or any brain-like system.
  - Whether or not the brain learns them, or they are built in!
- They show how syntactic structures do not have to be discrete and symbolic; they can be continuous and vector-valued.

Topics

- What can such models tell us in principle?
- Evidence for Linguistic Structure in LMs
- Learning and Representation
- Conclusion

How to Learn from Data

Data	
The key is pretty.	The keys are pretty.
Are the keys pretty?	I like the keys to the cabinet.
The man is happy.	Is the man happy?
The man who is tall is happy.	I like the man who is tall.

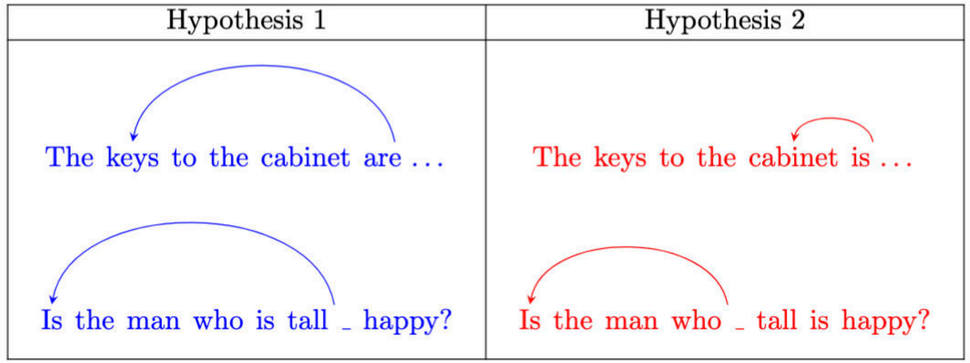
Hypothesis 1	Hypothesis 2
	

- Both hypotheses are equally consistent with the data.
- To arrive at Hypothesis 1, not Hypothesis 2, you need an **inductive bias** (a.k.a. the "evaluation measure" from Chomsky, 1965: 31-37), *which is not present in the data.*
- This logic is not in dispute. The question is: *What is the nature of that bias?*
- *Both for humans, and for LMs!*

Futrell & Mahowald (2026)

Inductive Bias Requires Restriction?

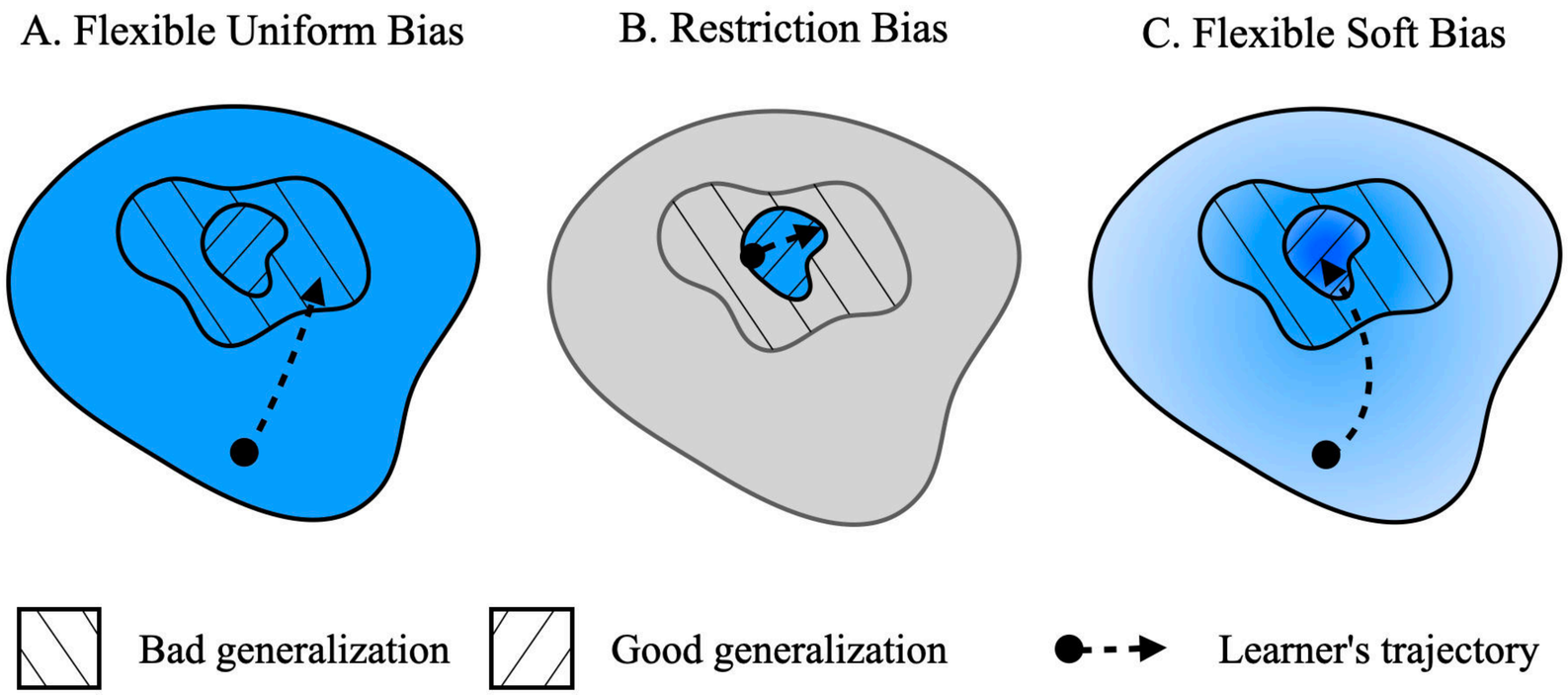
- Usual Solution: Your learner must be *restricted* to only consider certain hypotheses.
  - The learner represents linguistic input using an **innate formal system**, specific to language and based on hierarchical structure, that cannot even form the non-hierarchical representations (Chomsky, 1965, 1971, 1981).
  - A linguistic formalism (like Minimalist Grammars) is meant to model such a system.
- LMs provide an alternative view of how inductive bias can work.



Futrell & Mahowald (2026)

Inductive Bias from a Simplicity Bias

- Large neural networks learn more effectively when their hypothesis space is *less* restricted (Wilson, 2025)
- They have an implicit **simplicity bias** that gives them a ***soft* inductive bias**.
- The inductive bias comes from many sources, not only the architecture.
  - You can think of it as minimizing a description length (eg, Huang et al., 2025), but this is not always helpful.



Wilson (2025)

Learning and Linguistic Theory

- Traditional Approach: Learners need a restrictive theory of mental representations, and linguistic theories are such theories
- Modern Machine Learning Outlook: A restricted hypothesis space is less important than a simplicity metric (Wilson, 2025)
  - For linguistics: "Explanation through constrained description" is not the only valid approach to explanatory adequacy (Haspelmath, 2008)
    - It is not bad if a linguistic formalism "overgenerates" grammars (eg, alleged Turing-completeness of HPSG) as long as you can define an appropriate simplicity metric
    - There are many possible soft biases that can help learning

Inductive Bias in LMs

- To the extent that LMs do form linguistic generalizations, it must be because they have some inductive bias which is aligned in some way with the structure of language
- One way to find out what that bias is is to ask how well they learn artificial languages, including "impossible" languages

Learning Languages with Disrupted Structure

***HOP languages** perturb verb inflection
using counting rules

Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

Learning Languages with Disrupted Structure

***HOP languages** perturb verb inflection
using counting rules

1. **NOHOP**
   He cleans his very messy bookshelf.

Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

Learning Languages with Disrupted Structure

***HOP languages** perturb verb inflection using counting rules

1. **NOHOP**
   He clean~~s~~ his very messy bookshelf.

Learning Languages with Disrupted Structure

***HOP languages** perturb verb inflection using counting rules

1. **NOHOP**

   He cleans his very messy bookshelf.

Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

Learning Languages with Disrupted Structure

***HOP languages** perturb verb inflection using counting rules

1. **NOHOP**

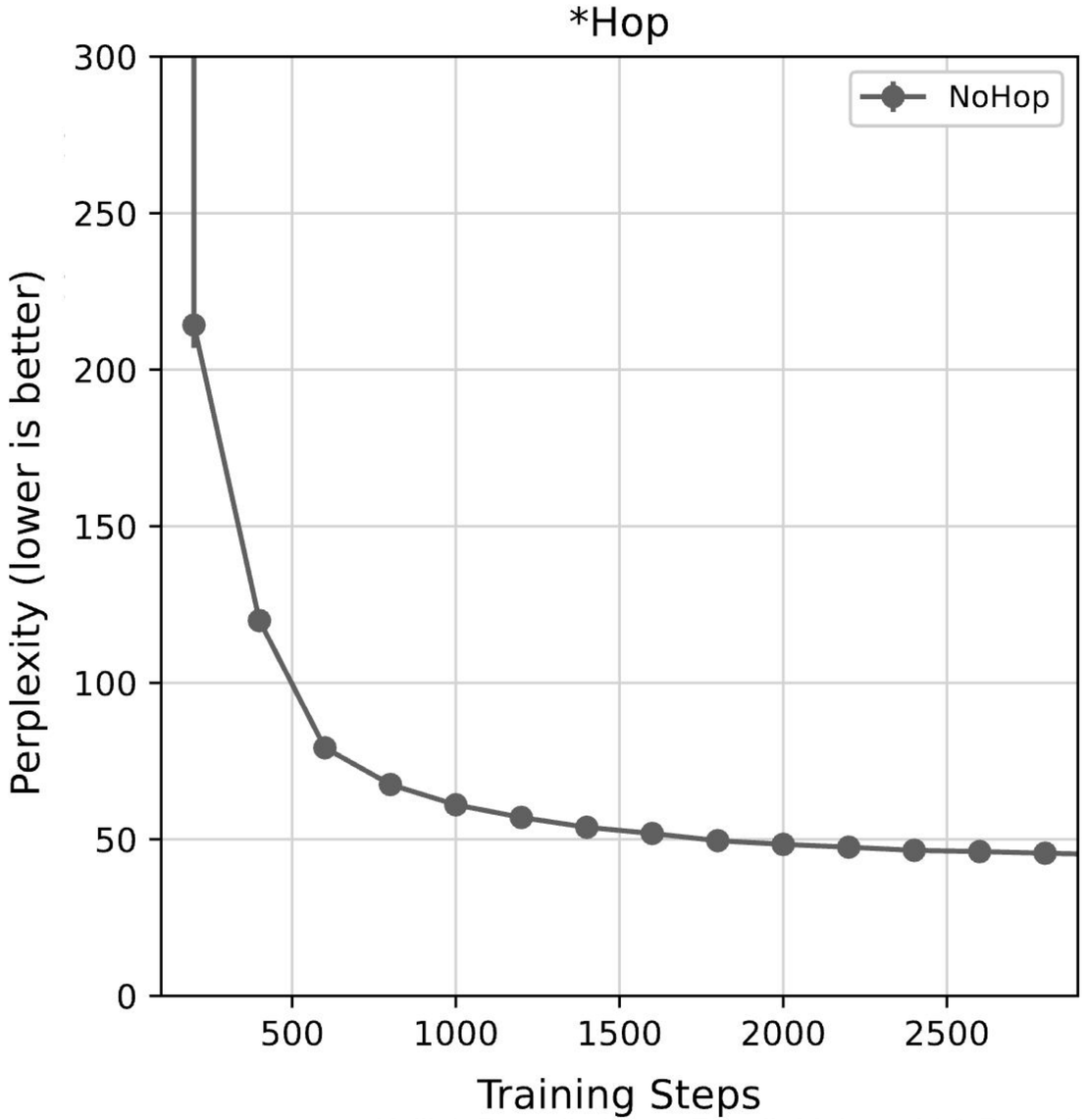
Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

Learning Languages with Disrupted Structure

***Hop languages** perturb verb inflection using counting rules

1. **NoHop**

He clean S his very messy books he lf .



Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

Learning Languages with Disrupted Structure

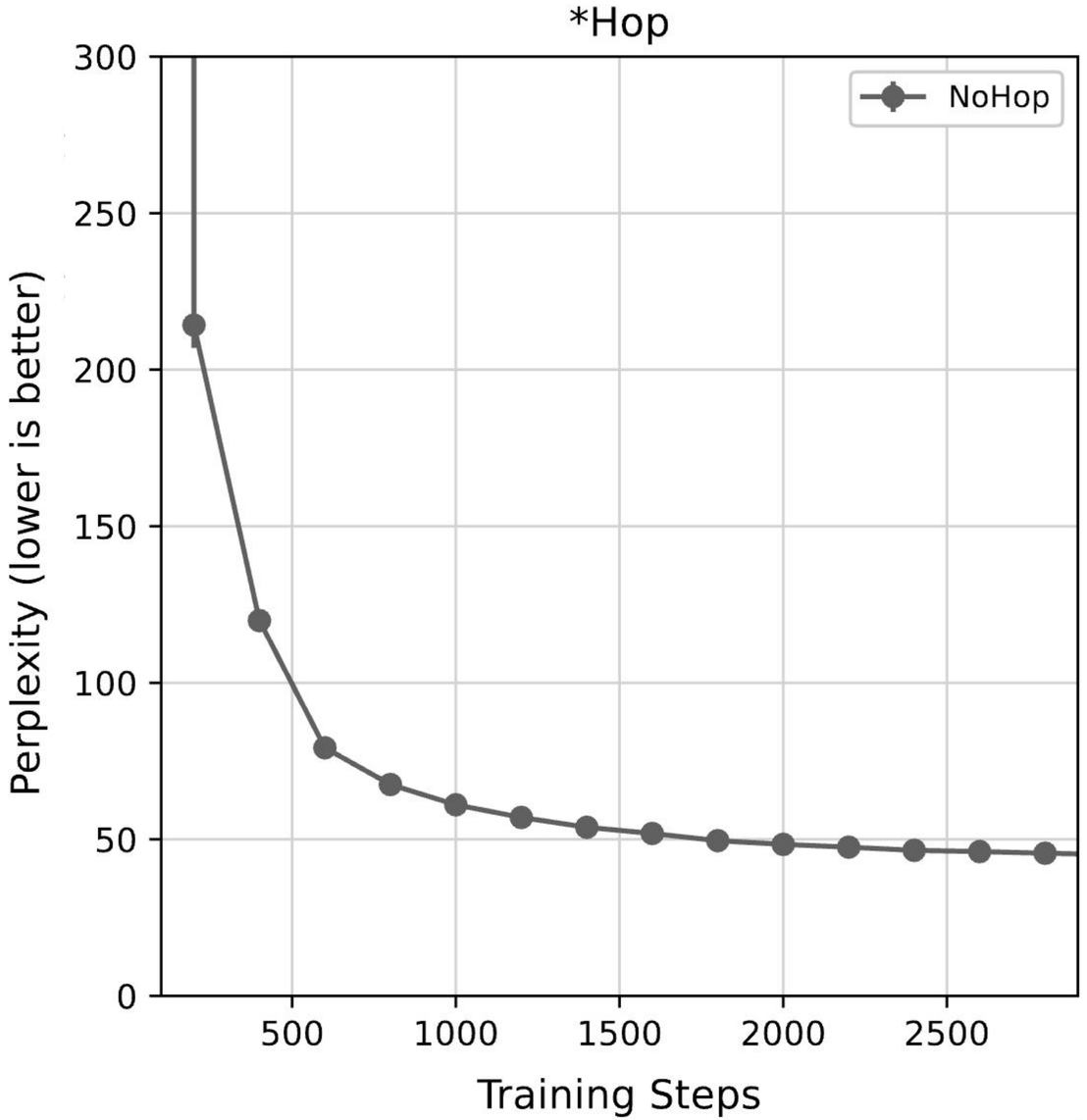
***Hop languages** perturb verb inflection using counting rules

1. **NoHop**

   He clean S his very messy books he lf .

2. **TokenHop**

   He clean S his very messy books he lf .



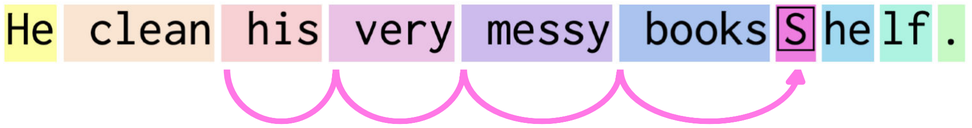
Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

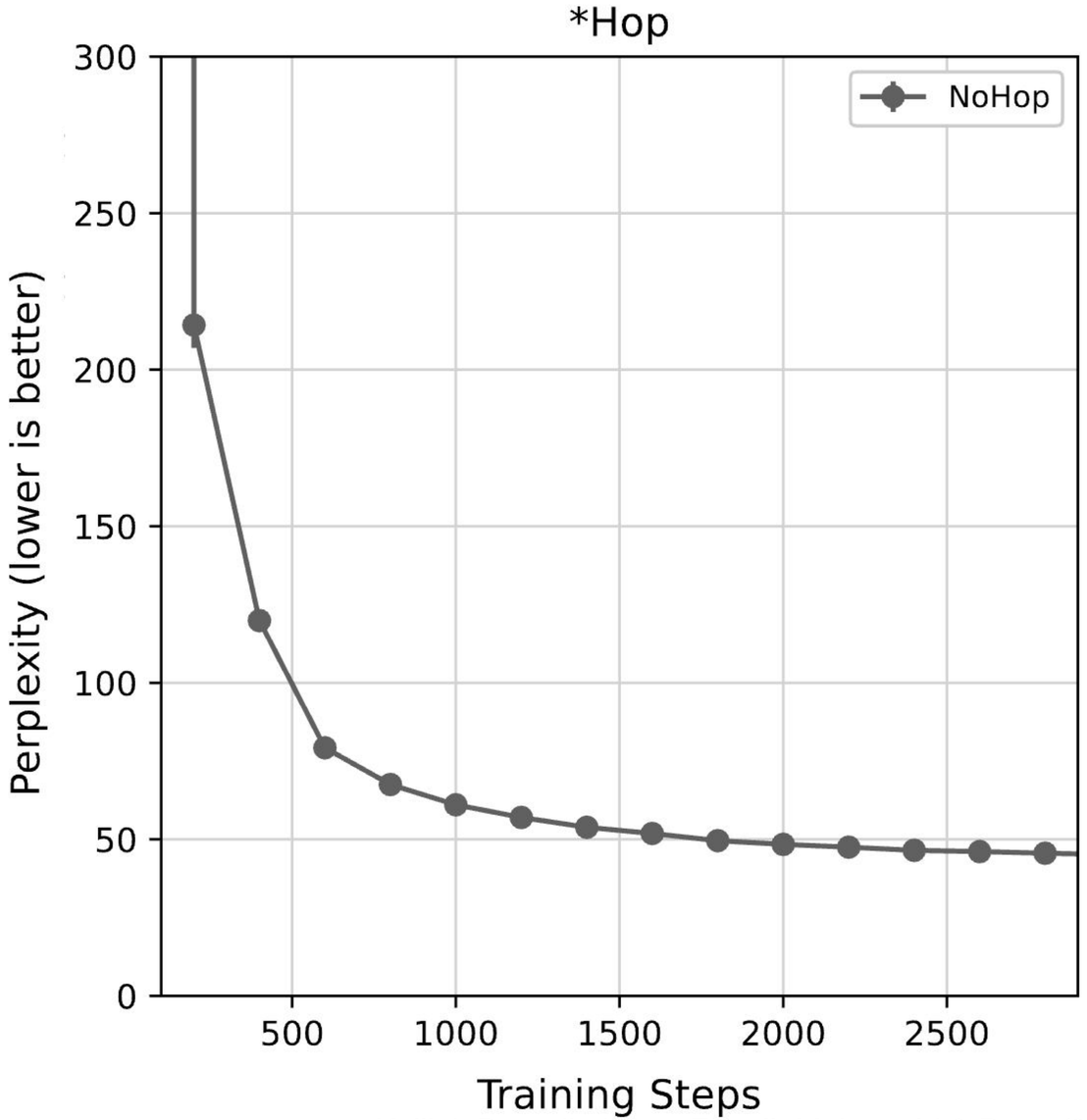
Learning Languages with Disrupted Structure

***Hop languages** perturb verb inflection using counting rules

1. **NoHop**

2. **TokenHop**





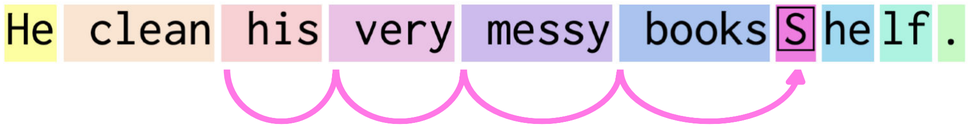
Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

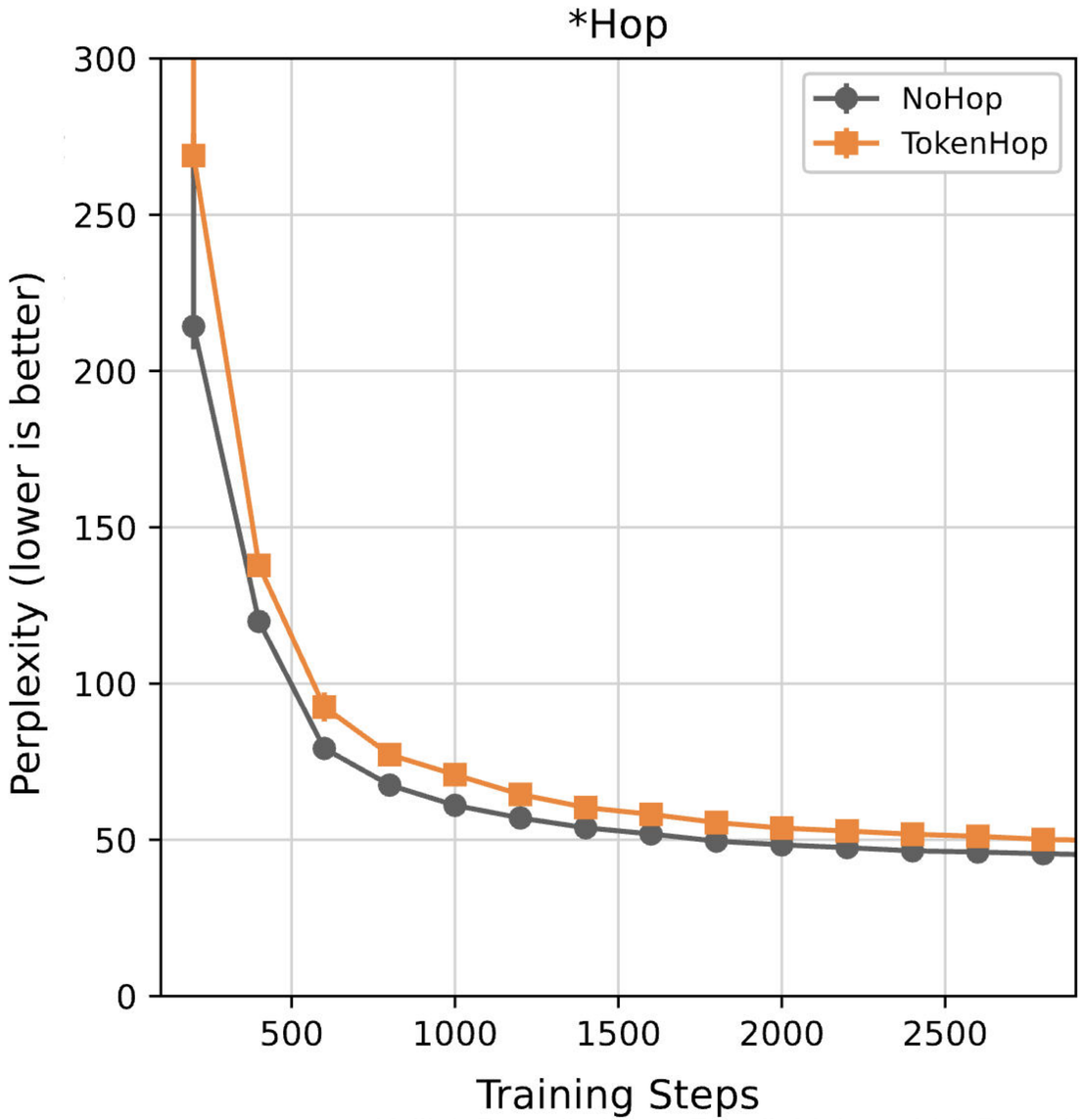
Learning Languages with Disrupted Structure

***Hop languages** perturb verb inflection using counting rules

1. **NoHop**

2. **TokenHop**





Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

Learning Languages with Disrupted Structure

***Hop languages** perturb verb inflection using counting rules

1. **NoHop**

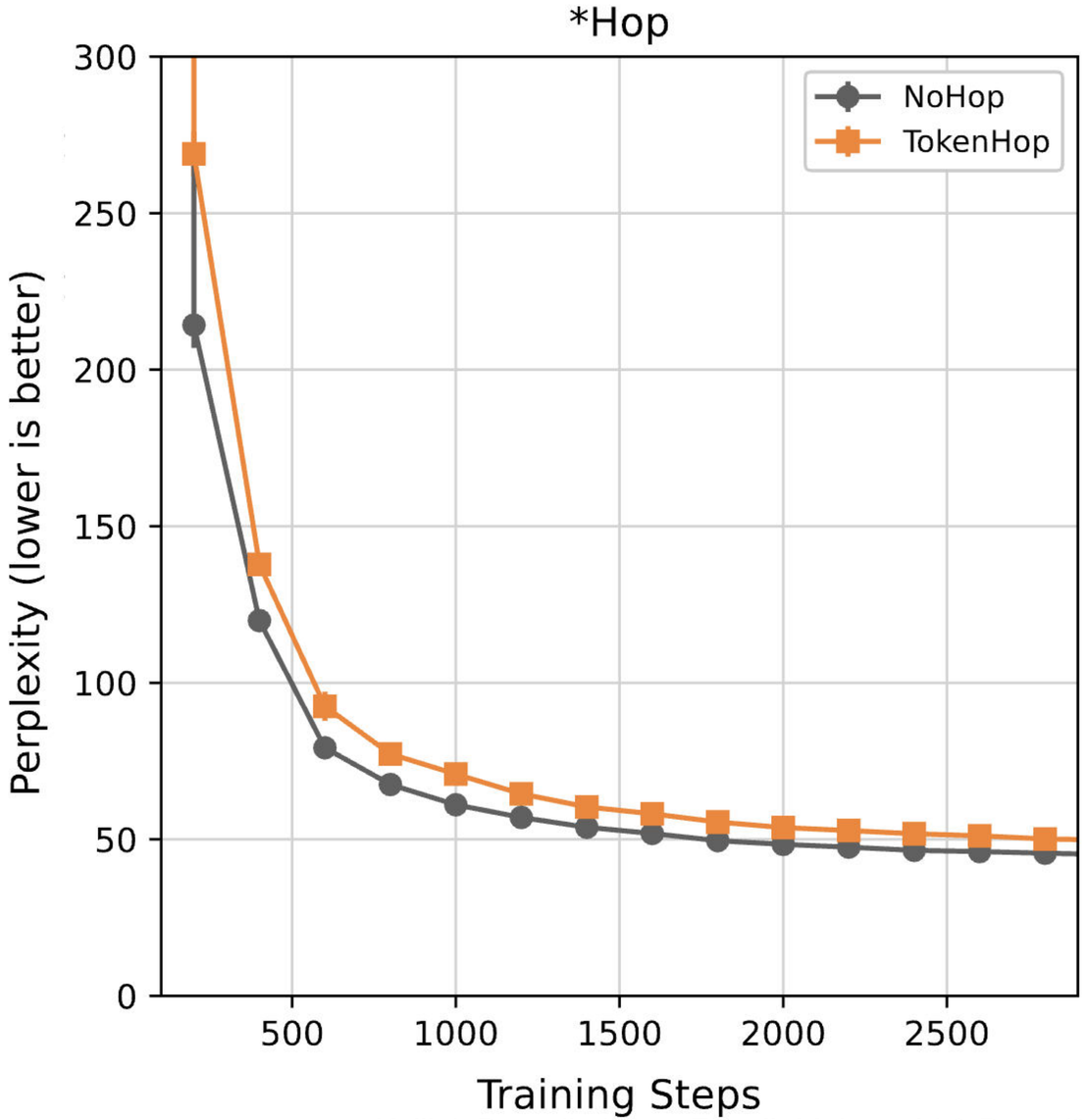
   He clean S his very messy books he lf .

2. **TokenHop**

   He clean his very messy books S he lf .

3. **WordHop**

   He clean S his very messy books he lf .



Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

Learning Languages with Disrupted Structure

***Hop languages** perturb verb inflection using counting rules

1. **NoHop**

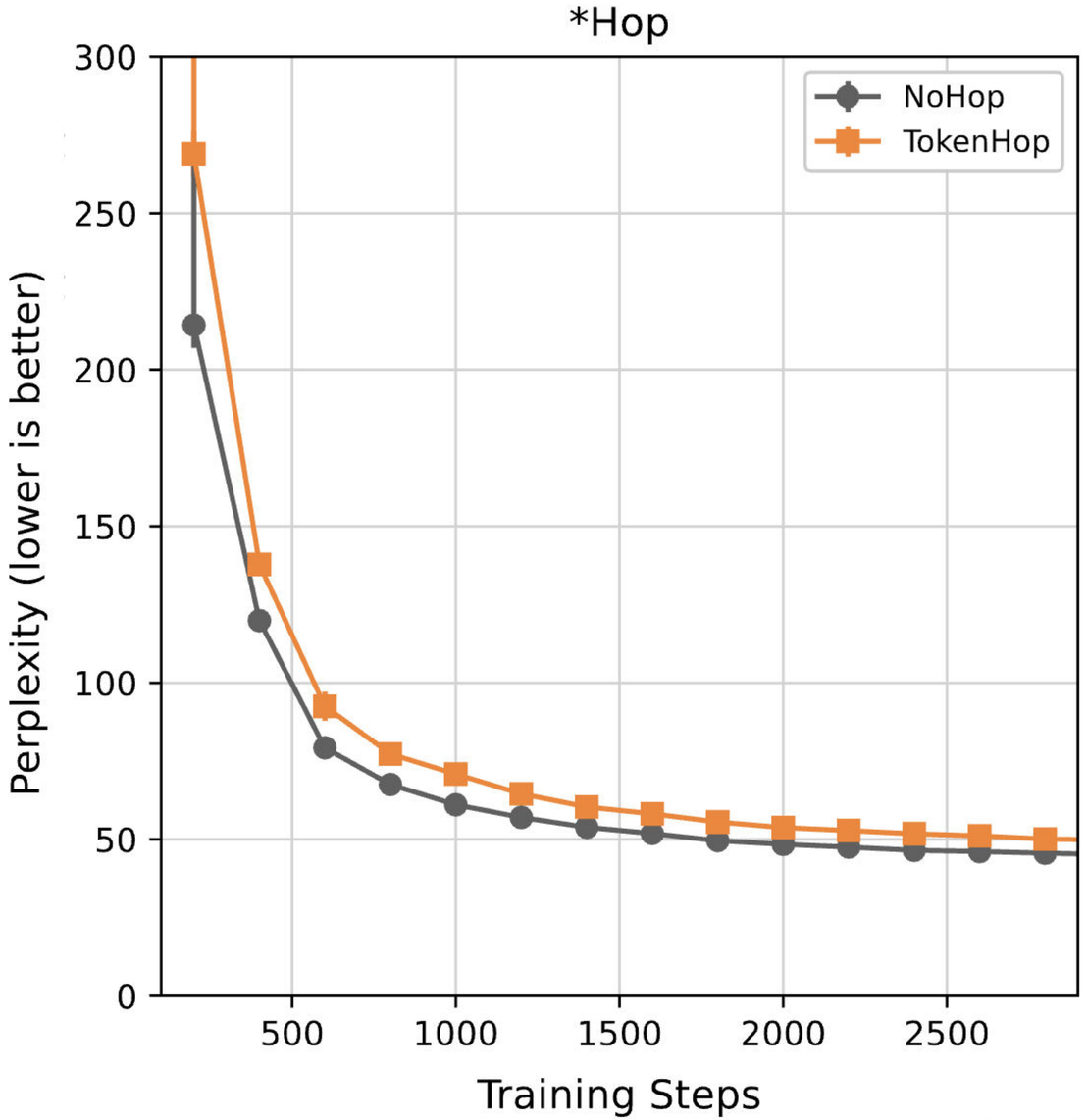
   He clean S his very messy books he lf .

2. **TokenHop**

   He clean his very messy books S he lf .

3. **WordHop**

   He clean his very messy books he lf S .



Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

Learning Languages with Disrupted Structure

***Hop languages** perturb verb inflection using counting rules

1. **NoHop**

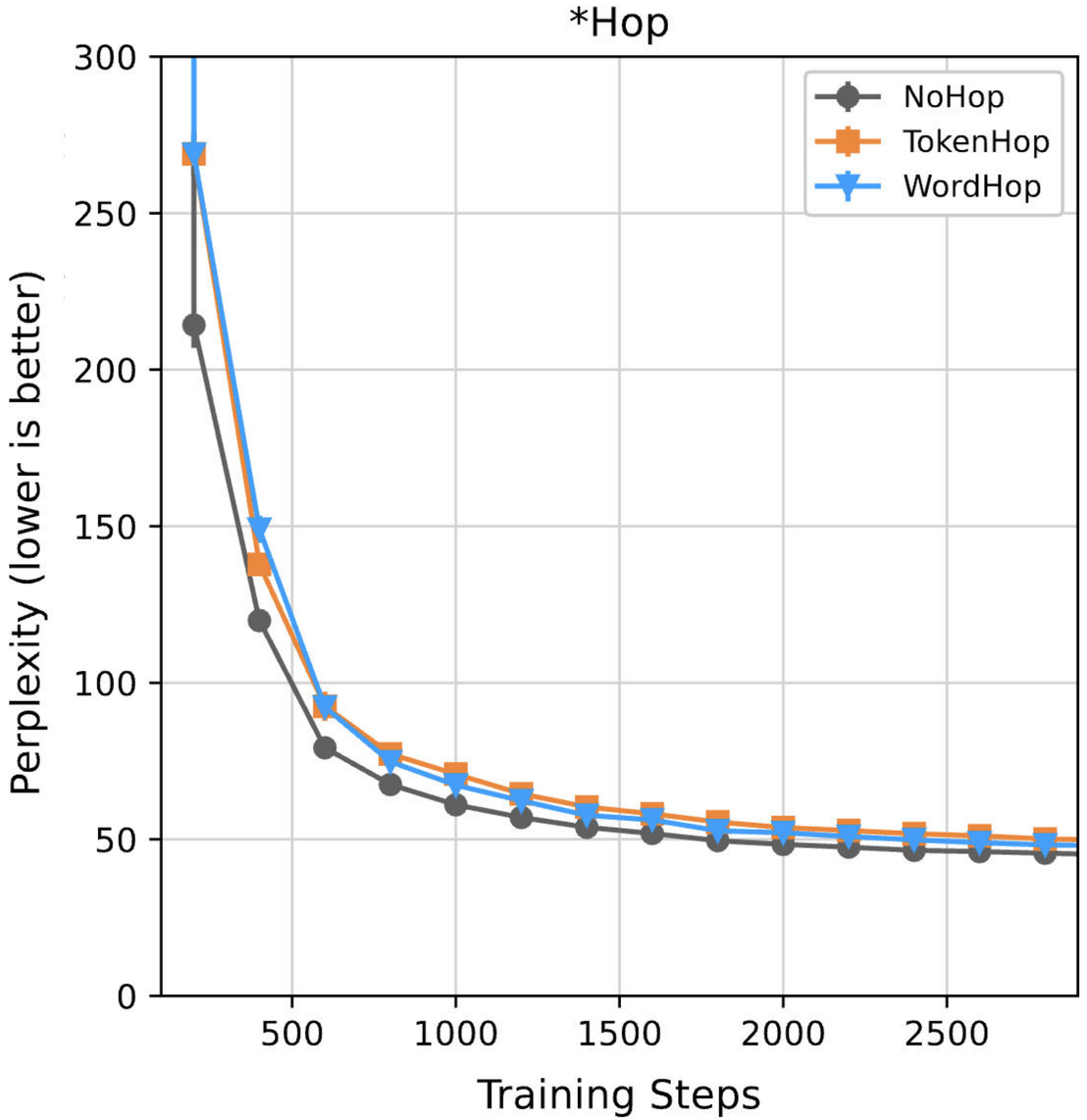
   He clean S his very messy books he lf .

2. **TokenHop**

   He clean his very messy books S he lf .

3. **WordHop**

   He clean his very messy books he lf S .



Kallini, Papadimitriou, Futrell, Mahowald & Potts (2024, ACL)

Impossible Languages: Complications

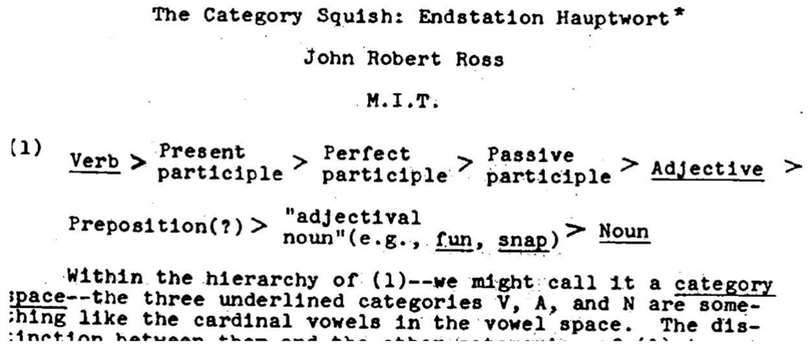
- Subsequently, Ziv et al. (2025) and Yang et al. (2025) report some impossible languages that seem easier to learn than real ones.
- Also, Hunter (2025) claims we do not properly control hierarchical vs. non-hierarchical languages.
  - I think a collaboration to find a pair of "languages", one hierarchical and one not, *controlled for statistical properties*, would be fruitful.
  - A number of simple suggestions (like languages from the fMRI experiments by Musso et al., 2003) have problems.
  - Also, it is not clear how hierarchical structure formally rules out things like *Hop.

Impossible Languages: Upshot

- The experiments show some inductive bias in LMs which is partially aligned with language, although weaker than human learners (Yang et al., 2025)
- We think part of that inductive bias is **information locality**: a tendency for related elements to be close (Futrell et al., 2020; Mansfield & Kemp, 2023; Someya et al., 2025)
  - Matches ideas from the functional typological literature: It is a statistical version of Hawkins' (2004) principle of **Minimize Domains**, used to explain Greenbergian word order universals
  - In LMs, arises from the function (predicting next word), not the architecture (Transformer)
  - In humans, hypothesized to arise from pressures of incremental processing (production, comprehension, parsing) (Gibson, 1998; Futrell et al., 2020; Hahn, Jurafsky & Futrell, 2020)
- More generally, ideas from the functionalist literature seem to match what LMs do...

LMs are a proof of concept for linguistic representations that are less discrete and categorical

- Word meanings: Represented as vectors encoding statistics of usage patterns (Erk, 2012, Potts, 2019) rather than as discrete predicates (Heim & Kratzer, 1998).
- Syntactic categories: Syntactic categories in LMs are fuzzy and exist in a space of functions (Ross, 1972, Comrie, 1989, Croft & Poole, 2008).
  - For example: LMs have a crosslinguistic feature for "grammatical subject", but passive subjects are less "subject-y" than active subjects (Papadimitriou et al., 2021).



The Category Squish: Endstation Hauptwort*

John Robert Ross

M.I.T.

(1) Verb > Present participle > Perfect participle > Passive participle > Adjective >

Preposition(?) > "adjectival noun"(e.g., fun, snap) > Noun

Within the hierarchy of (1)--we might call it a category space--the three underlined categories V, A, and N are something like the cardinal vowels in the vowel space. The distinction between them and the other

- Compositionality: Neural networks naturally capture gradient compositionality.
  - For example "green tea" is *more* compositional than "green thumb" but *less* than "green car" (Baroni et al., 2014). Neural nets represent compositional meanings, but they do not require a discrete compositional vs. non-compositional distinction.
- Linguistic levels: Linguistic levels (phonology, morphology, syntax, semantics) are represented in different layers of neural networks, but *softly and not strictly* (Belinkov, 2018).
  - As we find in psycholinguistics, where information from multiple layers can be combined flexibly in real-time processing.

Topics

- What can such models tell us in principle?
- Evidence for Linguistic Structure in LMs
- Learning and Representation
- Conclusion

Key Positions

- LMs do not replace or supplant linguistic theory.
- But they do inform questions of linguistic interest, by serving as systems that
  - 1. Demonstrate what is possible in a system that is not limited to certain formal structures.
  - 2. Generate hypotheses for neural representation of linguistic structures.
  - 3. Demonstate ways of thinking about *learning* and *representation* that might be new to formal linguists
- They open up the range of ideas and formal devices for linguistic theory.

Conclusions

- Linguistically informed computational work on LMs is already taking place within linguistics departments, where computational researchers are working alongside syntacticians, semanticists, phonologists, language documentation experts, sociocultural linguistics, and experts in a wide variety of languages and language families.
- This is an exciting time for linguistics!

Acknowledgments

- Thanks to Chris Potts, Harvey Lederman, David Beaver, Ted Gibson, Greg Hickok, Laura Kalin, Connor Mayer, Kanishka Misra, Lisa Pearl, Greg Scontras, Steve Wechsler, Xin Xie, and the UCI Quantitative Language Collective for helpful comments on this presentation and on Futrell & Mahowald (2026).
- Thanks for your attention!

References

- Begus, G., Dabkowski, M., & Rhodes, R. (2025). Large linguistic models: Investigating LLMs' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*.
- Korsky, S. A., & Berwick, R. C. (2019). On the computational power of RNNs. *arXiv preprint* arXiv:1906.06349.
- Cao, R., & Yamins, D. (2024). Explanatory models in neuroscience, Part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, 85, 101200.
- Zylberberg, J., Murphy, J. T., & DeWeese, M. R. (2011). A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *PLoS Computational Biology*, 7(10), e1002250.
- Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23), 3327-3338.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... & Blything, R. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46, e385.
- Futrell, R., & Mahowald, K. (2026, to appear). How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*.
- Piantadosi, S. T. (2024). Modern language models refute Chomsky's approach to language. In *From fieldwork to linguistic theory: A tribute to Dan Everett*, 353-414.
- Wilcox, E. G., Futrell, R., & Levy, R. (2024). Using computational models to test syntactic learnability. Linguistic Inquiry, 55(4), 805-848.
- Chomsky, N. (1957). Syntactic structures. Walter de Gruyter.
- Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521-535.
- Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.
- Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020, July). A systematic assessment of syntactic generalization in neural language models. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 1725-1744).
- Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.
- Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020). SyntaxGym: An online platform for targeted evaluation of language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 70-76).
- Hu, J., Wilcox, E. G., Song, S., Mahowald, K., & Levy, R. P. (2026). What Can String Probability Tell Us About Grammaticality?. Transactions of the Association for Computational Linguistics, 14, 124-146.
- Lan, N., Chemla, E., & Katzir, R. (2024). Large language models and the argument from the poverty of the stimulus. Linguistic Inquiry, 1-28.
- Pearl, L., & Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, *20*(1), 23-68.
- Dickson, N. (2025). Acquiring Syntax by Chunking Trees: A computational account of child syntactic learning (Doctoral dissertation, University of California, Irvine).
- Diego Simon, P. J., d'Ascoli, S., Chemla, E., Lakretz, Y., & King, J. R. (2024). A polar coordinate system represents syntax in large language models. Advances in Neural Information Processing Systems, 37, 105375-105396.
- Diego-Simón, P. J., Chemla, E., King, J. R., & Lakretz, Y. (2025). Probing syntax in large language models: Successes and remaining challenges. arXiv preprint arXiv:2508.03211.
- Hewitt, J., & Manning, C. D. (2019, June). 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) (pp. 4129-4138).

References

- Brinkmann, J., Wendler, C., Bartelt, C., & Mueller, A. (2025). Large language models share representations of latent grammatical concepts across typologically diverse languages. arXiv preprint arXiv:2501.06346.
- Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT press.
- Chomsky, N. (1971). Deep Structure, Surface Structure, and. Semantics: An interdisciplinary reader in philosophy, linguistics and psychology, 183.
- Chomsky, N. (1981). Knowledge of language: Its elements and origins. Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 295(1077), 223-234.
- Wilson, A. G. (2025). Deep learning is not so mysterious or different. arXiv preprint arXiv:2503.02113.
- Haspelmath, M. (2008). Parametric versus functional explanations of syntactic universals. In The limits of syntactic variation (pp. 75-107). John Benjamins Publishing Company.
- Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. arXiv preprint arXiv:2401.06416.
- Hunter, T. (2025). Kallini et al.(2024) do not compare impossible languages with constituency-based ones. Computational Linguistics, 51(2), 641-650.
- Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Büchel, C., & Weiller, C. (2003). Broca's area and the language instinct. Nature neuroscience, 6(7), 774-781.
- Ziv, I., Lan, N., Chemla, E., & Katzir, R. (2025). Biasless Language Models Learn Unnaturally: How LLMs Fail to Distinguish the Possible from the Impossible. arXiv preprint arXiv:2510.07178.
- Yang, X., Aoyama, T., Yao, Y., & Wilcox, E. (2025). Anything Goes? A Crosslinguistic Study of (Im) possible Language Learning in LMs. arXiv preprint arXiv:2502.18795.
- Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. Cognitive science, 44(3), e12814.
- Someya, T., Svete, A., DuSell, B., O'Donnell, T., Giulianelli, M., & Cotterell, R. (2025, July). Information locality as an inductive bias for neural language models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 27995-28013).
- Mansfield, J., & Kemp, C. (2023). The emergence of grammatical structure from inter-predictability.
- Hawkins, J. A. (2004). Efficiency and complexity in grammars. OUP Oxford.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. Cognition, 68(1), 1-76.
- Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. Proceedings of the National Academy of Sciences, 117(5), 2347-2353.
- Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. Language and Linguistics Compass, 6(10), 635-653.
- Potts, C. (2019). A case for deep learning in semantics: Response to Pater. Language, 95(1), e115-e124.
- Ross, J. R. (1972). The category squish: Endstation Hauptwort. In Chicago Linguistic Society (Vol. 8, pp. 316-328).
- Comrie, B. (1989). Language universals and linguistic typology: Syntax and morphology. University of Chicago press.
- Croft, W., & Poole, K. T. (2008). Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. Theoretical linguistics, 34(1), 1-37.
- Papadimitriou, I., Chi, E. A., Futrell, R., & Mahowald, K. (2021). Deep subjecthood: Higher-order grammatical features in multilingual BERT. arXiv preprint arXiv:2101.11043.
- Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program for compositional distributional semantics. Linguistic Issues in language technology, 9, 241-346.
- Belinkov, Y. (2018). On internal language representations in deep learning: An analysis of machine translation and speech recognition (Doctoral dissertation, Massachusetts Institute of Technology).