

# Linguistics, LLMs, and the inference to the best explanation

Roni Katzir  
rkatzir@tauex.tau.ac.il

GLOWing Lecture  
January 23, 2026

# Stochastic parrots? A refutation of generative linguistics?

## On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender<sup>\*</sup>  
ebender@uw.edu  
University of Washington  
Seattle, WA, USA

Timnit Gehr<sup>\*</sup>  
timnitg@blackinai.org  
Black in AI  
Palo Alto, CA, USA

Angelina McMillan-Major  
amym@uw.edu  
University of Washington  
Seattle, WA, USA

Shmargaret Shmitchell  
shmargaret.shmitchell@gmail.com  
The Author

### ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: how big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [26], and most recently Switch-C [14], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks [52], we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [126], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [37]. As we outline in §3, increasing the environmental and financial costs of these models doubly penalizes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing outlined in §2, the first consideration should be the environmental cost.

## The neural architecture of language: integrative modeling converges on predictive processing

Martin Schrimp<sup>a,b,1,2</sup>, Idan Asher Blank<sup>a,b,2</sup>, Greta Tuckwell<sup>a,b,2</sup>, Carina Kaur<sup>a,b,2</sup>, Eghbal A. Hossein<sup>a,b</sup>, Nancy Kanwisher<sup>a,b</sup>, Joshua B. Tenenbaum<sup>a,b,3</sup>, and Evelina Fedorenko<sup>a,b,1,2</sup>

<sup>a</sup>Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; <sup>b</sup>McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02138; <sup>c</sup>Center for Brain, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139; and <sup>d</sup>Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095

Contributed by Nancy Kanwisher, September 3, 2021 (sent for review April 22, 2021; revised by Matthew M. Botvinick and Adele E. Goldberg)

The neuroscience of perception has recently been revolutionized with an integrative modeling approach in which computation, brain function, and behavior are linked across many datasets and many computational models. By revealing trends across models, this approach yields novel insights into cognitive and neural mechanisms in the target domain. We have present a systematic study taking this approach to higher-level cognition: human language processing, our species' signature cognitive skill. We find that the most powerful "transformer" models predict nearly 100% of explainable variance in neural responses to sentences and generalize across different datasets and imaging modalities (functional MRI and electrocorticography). Models' neural fits ("brain scores") and fits to behavioral responses are both strongly correlated with model accuracy on the next-word prediction task (but not other language tasks). Model architecture appears to substantially contribute to neural fit. These results provide computationally explicit evidence that predictive processing fundamentally shapes the language comprehension mechanisms in the human brain.

network models of language processing with skepticism [13, 14], given that these systems lack (and often deliberately attempt to do without) explicit symbolic representation—traditionally seen as a core feature of linguistic meaning. Recent ANNs models of language, however, have proven capable of at least approximating some aspects of symbolic computation and have achieved remarkable success on a wide range of applied natural language processing (NLP) tasks. The results presented here, based on this new generation of ANNs, suggest that a computationally adequate model of language processing in the brain may be closer than previously thought.

Because we build on the same logic in our analysis of language in the brain, it is helpful to review why the neural network-based integrative modeling approach has proven so powerful in the study of object recognition in the ventral stream. Crucially, our ability to robustly link computation, brain function, and behavior is supported not by testing a single model on a single dataset or a single kind of data, but by large-scale integrative benchmarking (4) that establishes consistent patterns of performance across many different ANNs applied to multiple neural and behavioral datasets, together with their

computational neuroscience | language comprehension | neural recordings (fMRI and EEG) | artificial neural networks | deep learning

## Modern language models refute Chomsky's approach to language

Steven T. Piantadosi<sup>a,b</sup>

<sup>a</sup>UC Berkeley, Psychology <sup>b</sup>Helen Wills Neuroscience Institute

The rise and success of large language models undermines virtually every strong claim for the innateness of language that has been proposed by generative linguistics. Modern machine learning has subverted and bypassed the entire theoretical framework of Chomsky's approach, including its core claims to particular insights, principles, structures, and processes. I describe the sense in which modern language models implement genuine theories of language, including representations of syntactic and semantic structure. I highlight the relationship between contemporary models and prior approaches in linguistics, namely those based on gradient computations and memorized constructions. I also respond to several critiques of large language models, including claims that they can't answer "why" questions, and skepticism that they are informative about real life acquisition. Most notably, large language models have attained remarkable success at discovering grammar without using any of the methods that some in linguistics insisted were necessary

Trends in  
Cognitive Sciences



### Feature Review

## Dissociating language and thought in large language models

Kyle Mahowald,<sup>1,2,3</sup> Anna A. Ivanova,<sup>3,4,5</sup> Idan A. Blank,<sup>3,4</sup> Nancy Kanwisher,<sup>4,6</sup> Joshua B. Tenenbaum,<sup>4,6</sup> and Evelina Fedorenko<sup>4,6</sup>

Large language models (LLMs) have come closest among all models to date to mastering human language, yet opinions about their linguistic and cognitive capabilities remain split. Here, we evaluate LLMs using a distinction between formal linguistic competence (knowledge of linguistic rules and patterns) and functional linguistic competence (understanding and using language in the world). We ground this distinction in human neuroscience, which has shown that formal and functional competence rely on different neural mechanisms. Although LLMs are exceptionally

### Highlights

Formal linguistic competence (getting the form of language right) and functional linguistic competence (using language to accomplish goals in the world) are distinct cognitive skills.

The human brain contains a network of

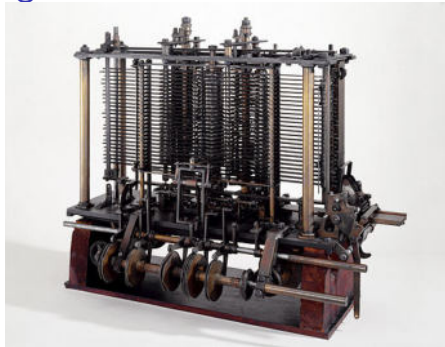
# Science vs engineering: flying



## Relevance?

- The biplane theory
- The proxy view

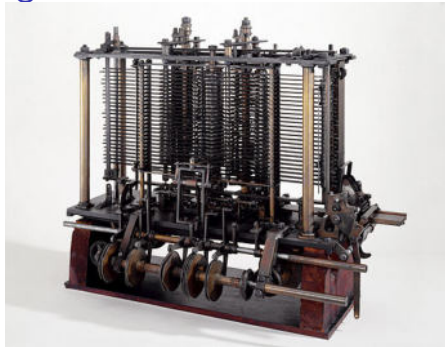
# Science vs engineering: thinking



## Relevance?

- The LLM theory
- The proxy view

# Science vs engineering: thinking



## Relevance?

- The LLM theory
- The proxy view

## Further discussion

Katzir (2023), Fox and Katzir (2024), Lan et al. (2024b), Ziv et al. (2025)

# Scientific explanation: a reminder

## Inference to the best explanation (ITTBE)

- Science aims to explain, not to approximate surface frequencies
- Laws are isolatable parts of best explanations
- A law might hold but be obscured on the surface



# Scientific explanation: a reminder

## Inference to the best explanation (ITTBE)

- Science aims to explain, not to approximate surface frequencies
- Laws are isolatable parts of best explanations
- A law might hold but be obscured on the surface

## One statement of ITTBE

$$\arg \min_{H \in \mathcal{H}} |H| + |D : H|$$



# A first step toward explaining human linguistic cognition

## Discrete unboundedness

A turtle is dancing on the roof



# A first step toward explaining human linguistic cognition

## Discrete unboundedness

A flying turtle is dancing on the roof

# A first step toward explaining human linguistic cognition

## Discrete unboundedness

A flying turtle is dancing angrily on the roof

# A first step toward explaining human linguistic cognition

## Discrete unboundedness

A flying turtle is dancing angrily on the roof or a three-tailed zebra is sitting in the corner

# A first step toward explaining human linguistic cognition

## Discrete unboundedness

If a flying turtle is dancing angrily on the roof or a three-tailed zebra is sitting in the corner then it is going to snow

# A first step toward explaining human linguistic cognition

## Discrete unboundedness

Kim claims that if a flying turtle is dancing angrily on the roof or a three-tailed zebra is sitting in the corner then it is going to snow

# A first step toward explaining human linguistic cognition

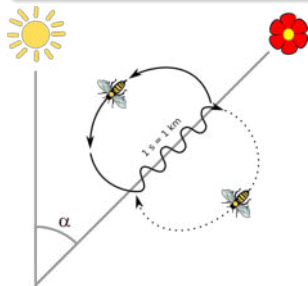
## Discrete unboundedness

Kim claims that if a flying turtle is dancing angrily on the roof or a three-tailed zebra is sitting in the corner then it is going to snow . . .

# A first step toward explaining human linguistic cognition

## Discrete unboundedness

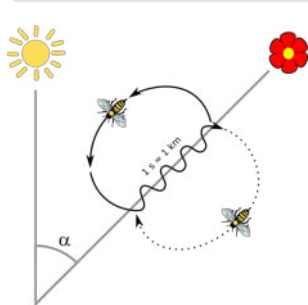
Kim claims that if a flying turtle is dancing angrily on the roof or a three-tailed zebra is sitting in the corner then it is going to snow . . .



# A first step toward explaining human linguistic cognition

## Discrete unboundedness

Kim claims that if a flying turtle is dancing angrily on the roof or a three-tailed zebra is sitting in the corner then it is going to snow . . .

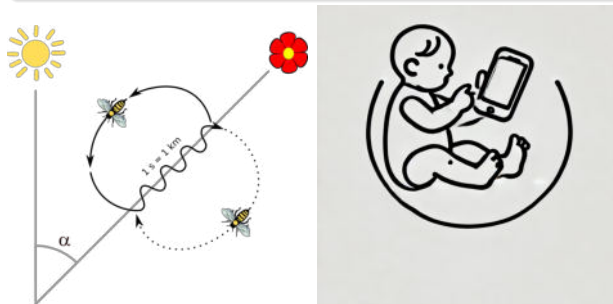




# A first step toward explaining human linguistic cognition

## Discrete unboundedness

Kim claims that if a flying turtle is dancing angrily on the roof or a three-tailed zebra is sitting in the corner then it is going to snow ...



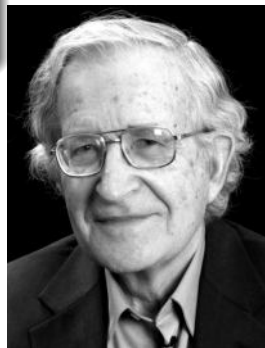
## The Computational Hypothesis

Certain aspects of the linguistic competence of a native speaker of a language can be characterized as a computer program.

# What everyone can agree on

## Reverse engineering human linguistic cognition

If part of linguistic competence is a computer program, we can try to figure out this program and reverse-engineer the computer on which it is run



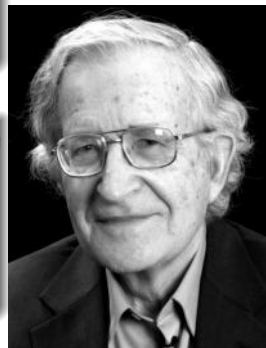
# What everyone can agree on

## Reverse engineering human linguistic cognition

If part of linguistic competence is a computer program, we can try to figure out this program and reverse-engineer the computer on which it is run

## Core parts of the computer

- Programming language
- Learning algorithm
- Processing mechanisms
- ...



# What everyone can agree on

## Reverse engineering human linguistic cognition

If part of linguistic competence is a computer program, we can try to figure out this program and reverse-engineer the computer on which it is run

## Core parts of the computer

- Programming language
- Learning algorithm
- Processing mechanisms
- ...

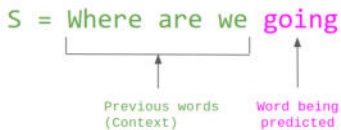
## Universal Grammar

All humans are born with the same programming language and learning algorithm



- 1 Large language models
- 2 Two architectural considerations
  - Competence vs. performance
  - Correctness vs. likelihood
- 3 Objective function
- 4 Representations
  - Alignment with the stimulus
  - Typology
- 5 Concluding remarks

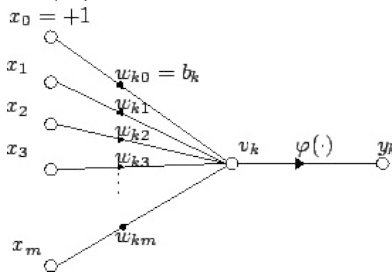
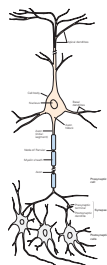
# Language models



$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$



# Artificial neural networks



WITTEN, K. OF  
MATHEMATICAL BIOLOGY  
VOLUME 5, 1956

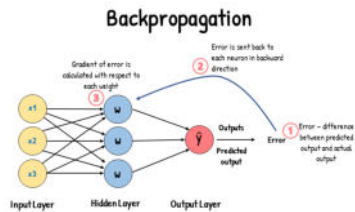
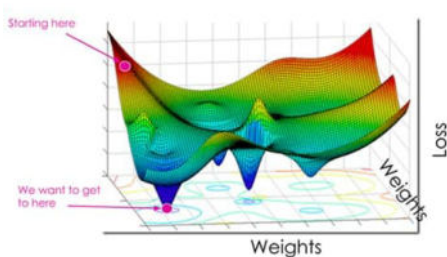
## A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH AND WALTER PITTS

FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE,  
DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE,  
AND THE UNIVERSITY OF CHICAGO.

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical terms for cells containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular cases among possible neurophysiological interconnections are redundant, in the sense that for every net behaving under one assumption, there exists another net which behaves identically for other and gives the same result, although perhaps not in the same time. Various examples of the calculus are discussed.

Training: trying to minimize  $|D : H|$





- 1 Large language models
- 2 Two architectural considerations
  - Competence vs. performance
  - Correctness vs. likelihood
- 3 Objective function
- 4 Representations
  - Alignment with the stimulus
  - Typology
- 5 Concluding remarks

## Familiar examples

### Center embedding

The mouse [the cat [the dog **[the bee stung]** bit] chased] ran

## Familiar examples

### Center embedding

The mouse [the cat [the dog [**the bee stung**] bit] chased] ran

### Agreement attraction

The little duck [that met the horses with the blue spots who were eating apples] **\*are**/✓**is** jumping

# Familiar examples

## Center embedding

The mouse [the cat [the dog [**the bee stung**] bit] chased] ran

## Agreement attraction

The little duck [that met the horses with the blue spots who were eating apples] **\*are**/✓**is** jumping

## Humans

The best explanation for human behavior distinguishes between competence and performance

# Familiar examples

## Center embedding

The mouse [the cat [the dog [**the bee stung**] bit] chased] ran

## Agreement attraction

The little duck [that met the horses with the blue spots who were eating apples] **\*are**/✓**is** jumping

## Humans

The best explanation for human behavior distinguishes between competence and performance

## LLMs

- No evidence for difficulty in LLMs
- And their errors are competence errors

- 1 Large language models
- 2 Two architectural considerations
  - Competence vs. performance
  - Correctness vs. likelihood
- 3 Objective function
- 4 Representations
  - Alignment with the stimulus
  - Typology
- 5 Concluding remarks

## Correct vs. probable

	<b>Probable</b>	<b>Improbable</b>
<b>Correct</b>	Kim danced	The lettuce the cabbage saw absconded
<b>Incorrect</b>	The key to the cabinets were missing	The a dog saw

## Correct vs. probable

	<b>Probable</b>	<b>Improbable</b>
<b>Correct</b>	Kim danced	The lettuce the cabbage saw absconded
<b>Incorrect</b>	The key to the cabinets were missing	The a dog saw

- The distinction between correctness and probability is fundamental to humans
- Well beyond language: art, moral judgments, science, ...
- LLMs only output probabilities
- The implications for the role of LLMs in society are worrying



- 1 Large language models
- 2 Two architectural considerations
  - Competence vs. performance
  - Correctness vs. likelihood
- 3 Objective function
- 4 Representations
  - Alignment with the stimulus
  - Typology
- 5 Concluding remarks

# A simple pattern

Puzzle

aabbbccddd

# A simple pattern

## Puzzle

aabbbccddd, abbcdd

# A simple pattern

## Puzzle

aabbbccddd, abbcdd, abcd

# A simple pattern

## Puzzle

aabbbccddd, abbcdd, abcd, aaaabcccd

# A simple pattern

## Puzzle

aabbbccddd, abbcdd, abcd, aaaabcccd, aabbcdd

# A simple pattern

## Puzzle

aabbbccddd, abbcdd, abcd, aaaabcccd, aabbcdd, aaabbbbcccd

# A simple pattern

## Puzzle

aabbbccddd, abbcdd, abcd, aaaabcccd, aabbcdd, aaabbbbcccd, aaabcccd



# A simple pattern

## Puzzle

aabbbccddd, abbcdd, abcd, aaaabcccd, aabbcdd, aaabbbbcccd, aaabcccd

## Solution

$a^i b^j c^i d^j$

# ChatGPT as scientist

Model: GPT-4

RB

Here are a few strings that follow a certain pattern: aabbbccddd, abbcdd, abcd, aaaabccccc, aabbcddd, aaabbbbbbccccdddd, aaabcccd. Can you list five other strings that follow the same pattern?



Sure! The pattern seems to be a string with increasing consecutive occurrences of each letter in the alphabet starting from 'a'. Here are five other strings that follow the same pattern:

1. aabccccddddeeeee
2. aaabbbbccccdd
3. aabccddeee
4. aabbcddde
5. aaabbbbccccccccddd

RB

Another string following the same pattern starts like this: aaaaabbc. Can you complete the string?



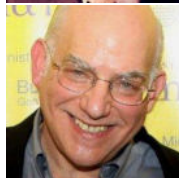
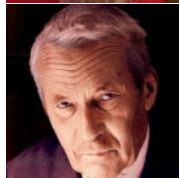
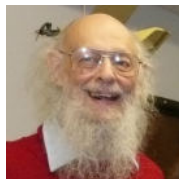
Certainly! Following the pattern of increasing consecutive occurrences of each letter in the alphabet starting from 'a', the completed string would be: aaaaabbcddddeeeeeee.



# The missing ingredient: simplicity

## LLMs miss half the picture

- The failure of ChatGPT is not an accident
- Recall: LLMs minimize  $|D : H|$
- To be good scientists, LLMs would need to perform inference to the best explanation



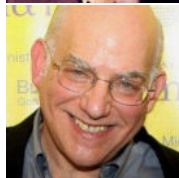
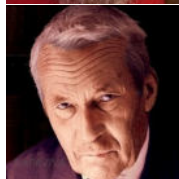
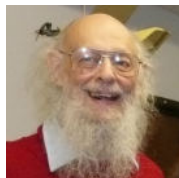
# The missing ingredient: simplicity

## LLMs miss half the picture

- The failure of ChatGPT is not an accident
- Recall: LLMs minimize  $|D : H|$
- To be good scientists, LLMs would need to perform inference to the best explanation

## Reminder: ITTBE

$$\arg \min_{H \in \mathcal{H}} |H| + |D : H|$$



# Neural networks can be better scientists with ITTBE<sup>1</sup>

## Minimum Description Length Recurrent Neural Networks

Nur Lan<sup>1,2</sup>, Michal Geyer<sup>2</sup>, Emmanuel Chémia<sup>1,3\*</sup>, Roni Katzir<sup>2\*</sup>

<sup>1</sup>Ecole Normale Supérieure, France

<sup>2</sup>Tel Aviv University, Israel

<sup>3</sup>EHESS, PSL University, CNRS

{nlan, chemia}@ens.fr

michalgeyer@mail.tau.ac.il

rkatzir@tauex.tau.ac.il

### Abstract

We train neural networks to optimize a Minimum Description Length score, that is, to balance between the complexity of the network and its accuracy at a task. We show that networks optimizing this objective function master tasks involving memory challenges and go beyond context-free languages. These learners master languages such as  $a^nb^n$ ,  $a^nb^nc^n$ ,  $a^nb^{2n}$ ,  $a^nb^mc^{n+m}$ , and they perform addition. Moreover, they often do so with 100% accuracy. The networks are small, and their inner workings are transparent. We thus provide formal proofs that their perfect accuracy holds not only on a given test set, but for any input sequence. To our knowledge, no other connectionist model has been shown to capture the underlying grammars for these languages in full generality.

corpora. Moreover, standard regularization techniques fall short in many cases, as we show below.

Second, even when successful, they tend to produce non-categorical results. That is, they output very high probabilities to target responses, but never 100%. Adequate, human-like generalization, on the other hand involves having both a probabilistic guess (which neural networks can do) and, at least in some cases, a clear statement of a categorical best guess (which neural networks cannot do).

Third, these networks are often very big, and it is generally very hard to inspect a given network and determine what it is that it actually knows (though see Lakretz et al., 2019, for a recent successful attempt to probe this knowledge in the context of linguistics).

Some of the challenges above arise from the

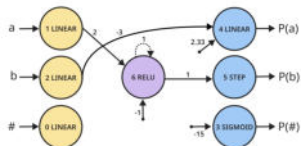


Figure 3: The network found by the MDL learner for the  $a^nb^n$  task, for a training set with data set size 500. See Theorem 4.1 for a description of how this network accepts any  $a^nb^n$  sequence and why it rejects any other sequence.

$a^nb^n$	Unit 6	Unit 4 <b><math>P(a)</math></b>	Unit 5 <b><math>P(b)</math></b>	Unit 3 <b><math>P(\#)</math></b>
Initial #	0	7/3 <b><math>\sim 1</math></b>	0 <b><math>\sim 0</math></b>	$\sigma(-15)$ <b><math>\sim 0</math></b>
$k^{th} a$	$k$	7/3 <b><math>\sim .7</math></b>	1 <b><math>\sim .3</math></b>	$\sigma(-15)$ <b><math>\sim 0</math></b>
$k^{th} b$ , $k < n$	$n - k$	-2/3 <b>0</b>	1 <b><math>\sim 1</math></b>	$\sigma(-15)$ <b><math>\sim 0</math></b>
$n^{th} b$	0	-2/3 <b>0</b>	0 <b>0</b>	$\sigma(-15)$ <b>1</b>

Table 2: Unit values (columns) during each phase of a valid  $a^nb^n$  sequence (rows). The second line for output units, given in bold, indicates the final normalized probability.

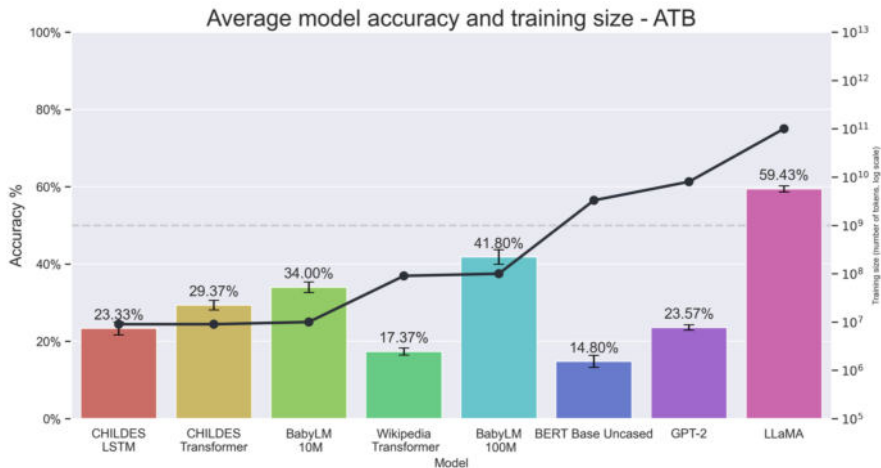
<sup>1</sup>Lan et al. (2022, 2024a); Abudy et al. (2025)

- 1 Large language models
- 2 Two architectural considerations
  - Competence vs. performance
  - Correctness vs. likelihood
- 3 Objective function
- 4 Representations
  - Alignment with the stimulus
  - Typology
- 5 Concluding remarks

# Some phenomena

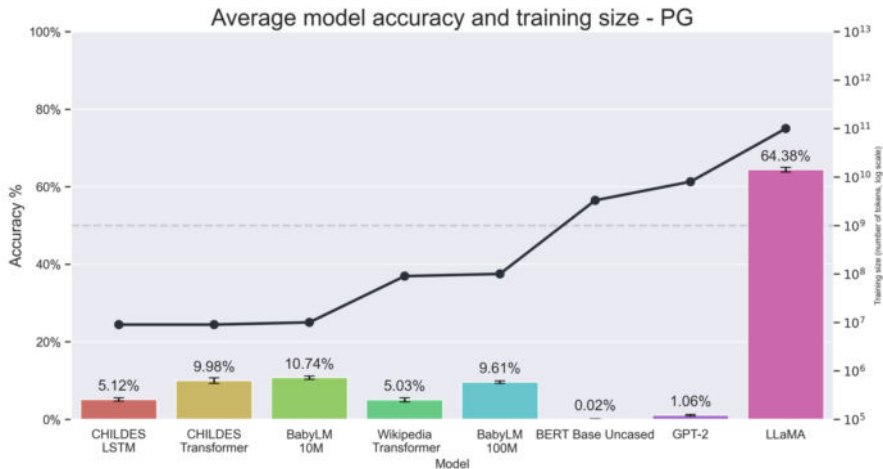
Phenomenon	Example ( <b>Grammatical</b> /* <b>Ungrammatical</b> )
Across-the-board movement (ATB)	<p>Which boy did you say that Kim hated and that Mary loved <u>yesterday</u>?</p> <p>* Which boy did you say that Kim hated and that Mary loved <u>Ann</u> yesterday?</p>
Parasitic gaps (PG)	<p>I know who John's talking to is going to annoy <u>soon</u>.</p> <p>* I know who John's talking to is going to annoy <u>you</u> soon.</p>
That-trace effects (TTE)	<p>* Who did you say that <u>loves</u> Sue?</p> <p>Who did you say that <u>Sue</u> loves?</p> <p>Who did you say <u>loves</u> Sue?</p> <p>Who did you say <u>Sue</u> loves?</p>

# Across-the-board movement



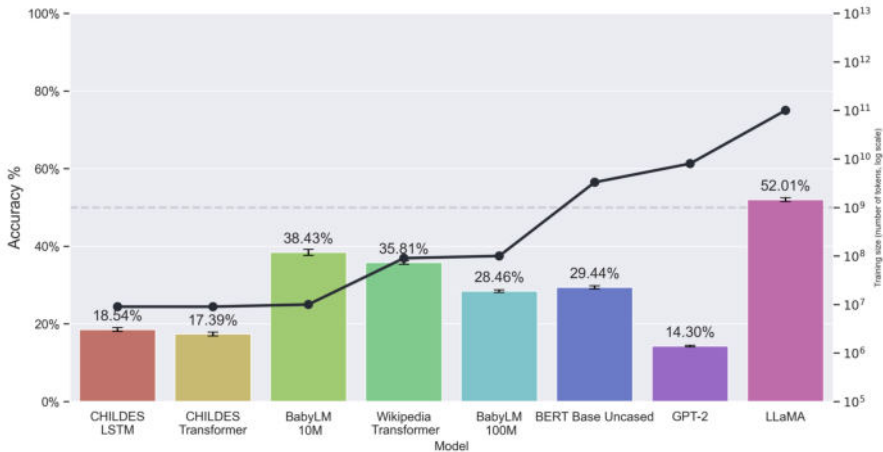


# Parasitic gaps



# That-trace effects

Average model accuracy and training size - TTE



- 1 Large language models
- 2 Two architectural considerations
  - Competence vs. performance
  - Correctness vs. likelihood
- 3 Objective function
- 4 Representations
  - Alignment with the stimulus
  - Typology
- 5 Concluding remarks

# Inference to the best explanation in light of the typology

## Multiple interacting factors

- Programming language
- Learning algorithm
- Processing mechanisms
- Communicative pressure
- ...



# Some key aspects of competence

## Constituency

- (1) [John will [talk [to [Kim]]]]
- a. Talk to Kim, John will
  - b. \* Talk to, John will Kim



# Some key aspects of competence

## Constituency

- (1) [John will [talk [to [Kim]]]]
- a. Talk to Kim, John will
  - b. \* Talk to, John will Kim

## Entailment

- (2) Kim spoke to ✓**every**/\***some** student who ever smoked



# Some key aspects of competence

## Constituency

- (1) [John will [talk [to [Kim]]]]
- a. Talk to Kim, John will
  - b. \* Talk to, John will Kim

## Entailment

- (2) Kim spoke to ✓ **every** / \***some** student who ever smoked



## Modularity

- (3) Kim only has (\***more than**) THREE children

- 1 Large language models
- 2 Two architectural considerations
  - Competence vs. performance
  - Correctness vs. likelihood
- 3 Objective function
- 4 Representations
  - Alignment with the stimulus
  - Typology
- 5 Concluding remarks



# Concluding remarks

## The LLM theory fails at explanation

- Architecture: competence vs. performance, correctness vs. likelihood
- They learn in a way that neglects  $|H|$
- Their programming language is non-human-like

# Concluding remarks

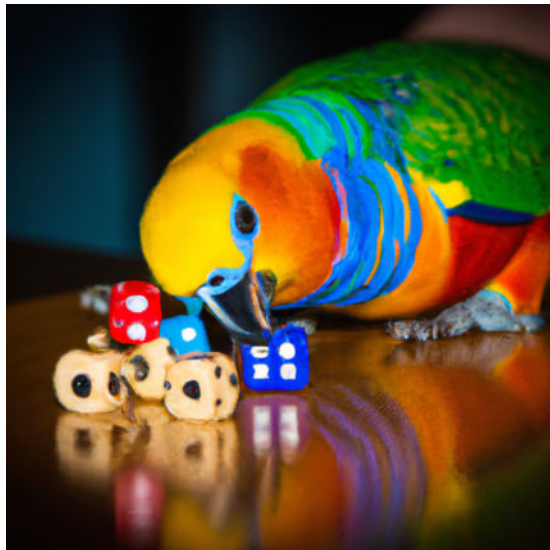
## The LLM theory fails at explanation

- Architecture: competence vs. performance, correctness vs. likelihood
- They learn in a way that neglects  $|H|$
- Their programming language is non-human-like

## The proxy view is potentially interesting, but...

- What target theory is it meant to support?
- So far, no challenges to linguistic theory

Thank you!



## Further readings I

- Abudy, Matan, Orr Well, Emmanuel Chemla, Roni Katzir, and Nur Lan. 2025. A minimum description length approach to regularization in neural networks. Ms., TAU and ENS, May 2025.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, 610–623. New York, NY, USA: Association for Computing Machinery.
- Berwick, Robert C., Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive Science* 35:1207–1242.
- Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.
- Chomsky, Noam. 1959. Review of B. F. Skinner. *Language* 35:26–58.
- Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

## Further readings II

- Fox, Danny, and Martin Hackl. 2006. The universal density of measurement. *Linguistics and Philosophy* 29:537–586.
- Fox, Danny, and Roni Katzir. 2024. Large language models and theoretical linguistics. *Theoretical Linguistics* 50:71–76.
- Futrell, Richard, and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences* 1–98.
- Katzir, Roni. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics* 17.
- Lan, Nur, Emmanuel Chemla, and Roni Katzir. 2024a. Bridging the empirical-theoretical gap in neural network formal language learning using minimum description length. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. Lun-Wei Ku, Andre Martins, and Vivek Srikumar, 13198—13210. Bangkok, Thailand: Association for Computational Linguistics.

## Further readings III

- Lan, Nur, Emmanuel Chemla, and Roni Katzir. 2024b. Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry* 1–56.
- Lan, Nur, Michal Geyer, Emmanuel Chemla, and Roni Katzir. 2022. Minimum description length recurrent neural networks. *Transactions of the Association for Computational Linguistics* 10:785–799.
- Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences* 28:517–540.
- McCulloch, Warren S., and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology* 5:115–133.

## Further readings IV

- Miller, George, and Noam Chomsky. 1963. Finitary models of language users. In *Handbook of Mathematical Psychology*, ed. R. Duncan Luce, Robert R. Bush, and Eugene Galanter, volume 2, 419–491. New York, NY: Wiley.
- Moro, Andrea. 2015. *The boundaries of babel: The brain and the enigma of impossible languages*. Current Studies in Linguistics. MIT Press, 2nd edition.
- Moro, Andrea, Matteo Greco, and Stefano F. Cappa. 2023. Large languages, impossible languages and human brains. *Cortex* 167:82–85.
- Piantadosi, Steven T. 2023. Modern language models refute Chomsky's approach to language. Ms., available at <https://ling.auf.net/lingbuzz/007180>, March 2023.

## Further readings V

- Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* 118:e2105646118.
- Solomonoff, Ray J. 1964. A formal theory of inductive inference, parts I and II. *Information and Control* 7:1–22, 224–254.
- Sprouse, Jon, Beracah Yankama, Sagar Indurkha, Sandiway Fong, and Robert C. Berwick. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review* 35:575–599.
- Ziv, Imry, Nur Lan, Emmanuel Chemla, and Roni Katzir. 2025. Large language models as proxies for theories of human linguistic cognition. Accepted, *Journal of Language Modelling*, February 2025.