CS 784: Computational Linguistics
Lecture 19: Computational Psycholinguistics with Language Models (Part I - Inference)

Freda Shi

School of Computer Science, University of Waterloo
fhs@uwaterloo.ca

March 25, 2025

Pre-Class Questions

Language models are trained on large corpora of text.

How does this training process relate to the science of natural language?

How do you like the following statement?

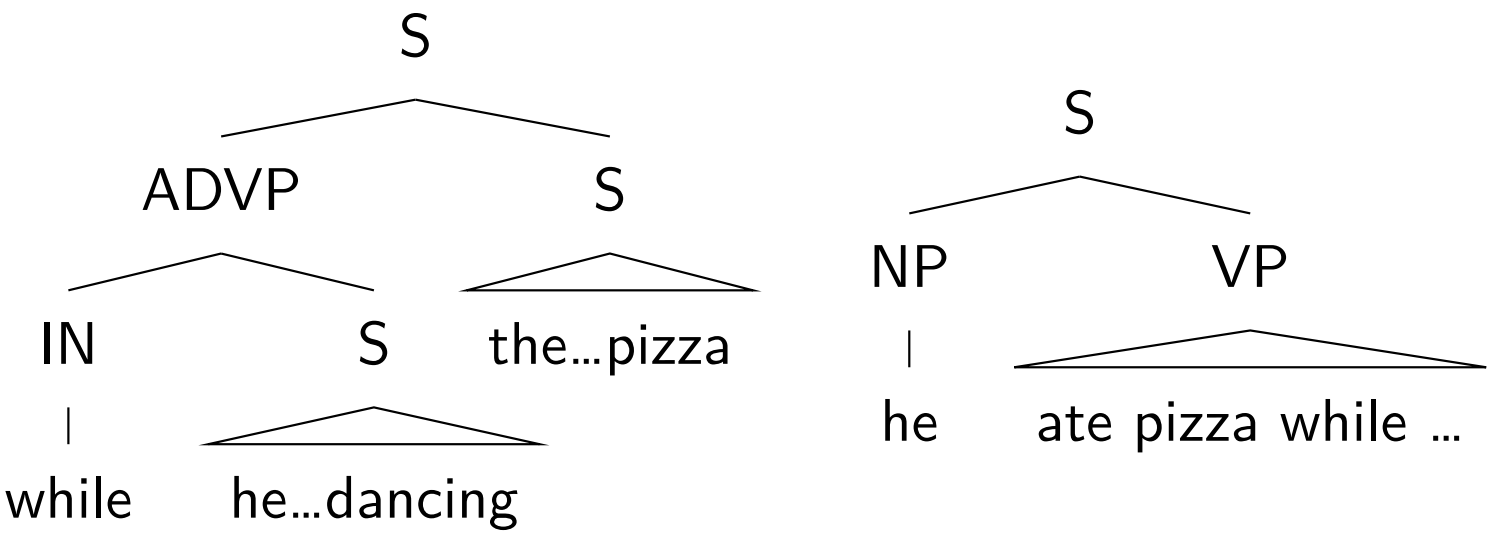
- Language models behave like humans in some senses, so studying language model training can help us understand **human language acquisition**.
- Language models behave like humans in some senses, so studying language model training can help us understand **adult human language processing**.
- Language models are just statistical models, so studying language model may reveal some facts about language, but not necessarily about how humans use them.

Poverty of the Stimulus

Children are not exposed to **rich enough** data with their linguistic environments to acquire **every** feature of their language without **innate language-specific cognitive biases**.

1. While he was dancing, the Ninja Turtle ate pizza.
2. He ate pizza while the Ninja Turtle was dancing.

Could *he* refer to the Ninja Turtle in both sentences?

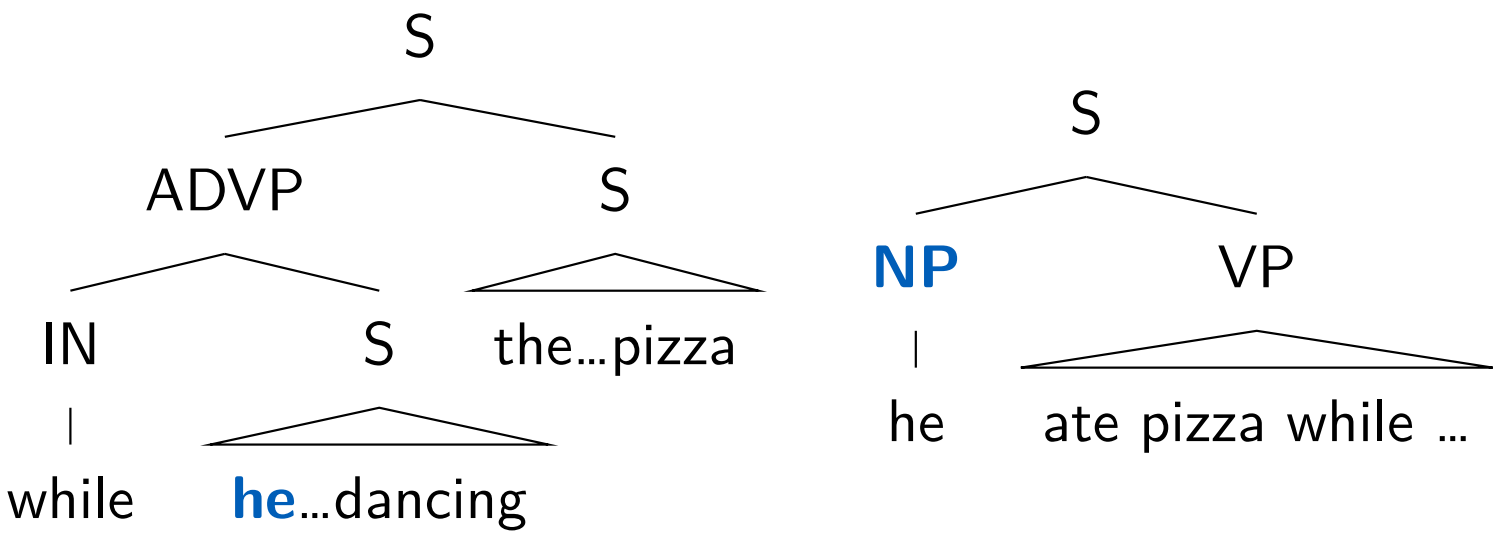


C-Command

A node A **C-commands** a node B if:

- Neither A nor B is the other's ancestor.
- Any node that dominates A also dominates B.

In other words, a node C-commands its siblings and all their descendants.



Crain and McKee (1985): the pronoun can precede its antecedent but cannot C-command it.

Poverty of the Stimulus

1. While he was dancing, the Ninja Turtle ate pizza.
2. He ate pizza while the Ninja Turtle was dancing.

The full argument for the poverty of the stimulus:

- It's **never made clear** to children that which interpretation is possible.
- Therefore, they must have some innate knowledge about the structure of language to recognize that the pronoun can refer to the Ninja Turtle in the first sentence but not in the second.

Is there possibly something wrong with this argument?

A (Possibly Overly) Critical Retrospect

Crain and McKee (1985): the pronoun can precede its antecedent but cannot C-command it.

Are we satisfied with this level of explanation? What could have been made clearer/stronger/more precise?

- Does this style of explanation apply to all interesting linguistic phenomena?
- Is C-command something specific for English or universal? How are the theories generalized for prediction?

Arguing against the Poverty of the Stimulus - Approach 1

The Poverty of the Stimulus

Children are not exposed to **rich enough** data with their linguistic environments to acquire **every** feature of their language without **innate language-specific cognitive biases**.

Language models successfully model almost all linguistic phenomena, with only the inductive bias of the Transformers.

Arguing against the Poverty of the Stimulus - Approach 2

The Poverty of the Stimulus

Children are not exposed to **rich enough** data with their linguistic environments to acquire **every** feature of their language without **innate language-specific cognitive biases**.

Grammar induction (i.e., unsupervised parsing) is quite accurate using large corpora of text and pretrained language models.

What if we achieve high accuracy in unsupervised parsing with only cognitively plausible data (e.g., 20 million words, roughly the amount of linguistic input to an 4- or 5-year-old human)?

Arguing against the Poverty of the Stimulus - Approach 3

The Poverty of the Stimulus

Children are not exposed to **rich enough** data with their linguistic environments to acquire **every** feature of their language without **innate language-specific cognitive biases**.

Training language models on small corpora of text (e.g., 20 million words, or child-directed speech) can still achieve high accuracy in next-word prediction in unseen context.

Poverty of the Stimulus: A Patch

Argument

LLMs have seen much larger corpora of text than children, so the data availability is not comparable, and therefore, LLMs are not a good model for human language acquisition.

Counterargument

LLMs indeed have much higher data availability than children, but we have our brains evolved through millions of years to process language.

If certain inductive biases have evolved in language models through training, then it supports a loose form of the poverty of the stimulus.

Language Models as Psycholinguistics Subjects

What is the goal of using language models as psycholinguistics subjects?

- Better understanding of human language processing.
- Better understanding of language models.
- Better understanding of the relationship between the two.

Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve

R. Thomas McCoy Shunyu Yao Dan Friedman Matthew Hardy Thomas L. Griffiths

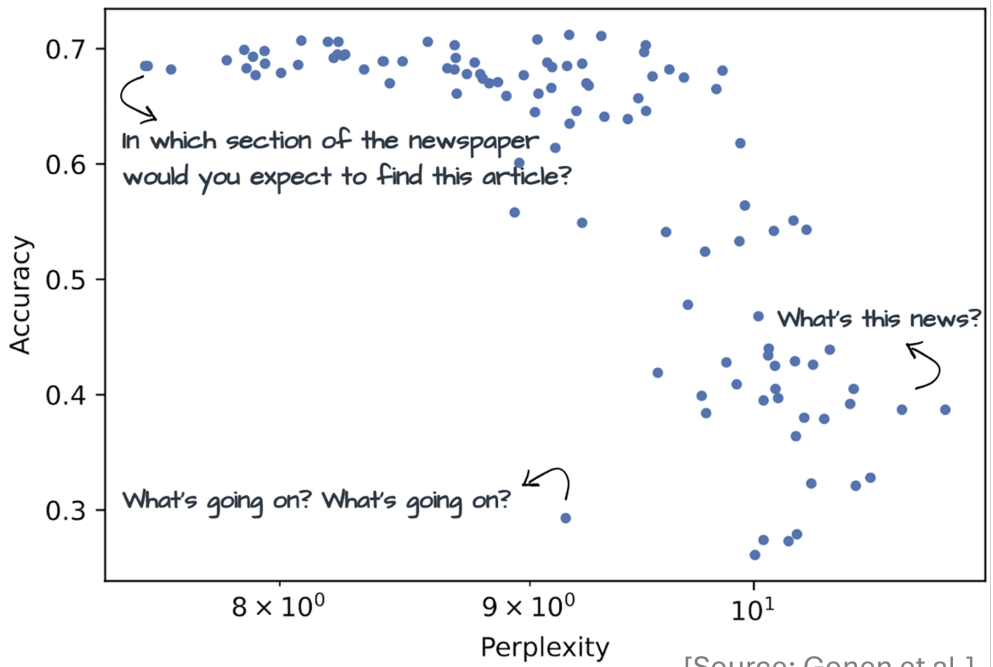
Princeton University

One-sentence summary:
To understand what language models are, we must understand what we have trained them to be.

Abstract:
The widespread adoption of large language models (LLMs) makes it important to recognize their strengths and limitations. We argue that in order to develop a holistic understanding of these systems we need to consider the problem that they were trained to solve: next-word prediction over Internet text. By recognizing the pressures that this task exerts we can make predictions about the strategies that LLMs will adopt, allowing us to reason about when they will succeed or fail. This approach—which we call the teleological approach—leads us to identify three factors that we hypothesize will influence LLM accuracy: the probability of the task to be performed, the probability of the target output, and the probability of the provided input. We predict that LLMs will achieve higher accuracy when these probabilities are high than when they are low—even in deterministic settings where probability should not matter. To test our predictions, we evaluate two LLMs (GPT-3.5 and GPT-4) on eleven tasks, and we find robust evidence that LLMs are influenced by probability in the ways that we have hypothesized. In many cases, the experiments reveal surprising failure modes. For instance, GPT-4's accuracy at decoding a simple cipher is 51% when the output is a high-probability word sequence but only 13% when it is low-probability. These results show that AI practitioners should be careful about using LLMs in low-probability situations. More broadly, we conclude that we should not evaluate LLMs as if they are humans but should instead treat them as a distinct type of system—one that has been shaped by its own particular set of pressures.

Language Model and Probability

Language models perform well on high-probability tasks and poorly on low-probability tasks.



[Source: Gonen et al.]

Linear functions

Multiply by 9/5 and add 32.

Input: 328
Correct: 622.4
✓ **GPT-4:** 622.4

Multiply by 7/5 and add 31.

Input: 328
Correct: 490.2
✗ **GPT-4:** 457.6

[Source: McCoy et al.]

A machine-learning explanation: without generalization guarantees, models perform well on in-domain data, and poorly on out-of-domain data.

Mixed-Effects Model

To systematically understand the behavior of language models, we need to control for various factors, and possibly conduct cross-model analysis.

$$y \sim \underbrace{x_1 + x_2 + x_3}_{\text{fixed effects}} + \underbrace{(1 + x_1 \mid \text{subject})}_{\text{per-subject random effects}} + \underbrace{(1 \mid \text{item})}_{\substack{\text{per-item} \\ \text{random effects}}}$$

- y: the dependent variable (e.g., reaction time, accuracy).
- x_1, x_2, x_3: independent predictor variables of data or participant (e.g., word frequency, word length, participant age), similarly to the fixed effects in linear regression.
- **Subject**: language models (analogous to human participants in psycholinguistics).
- **Item**: data example (e.g., a sentence).

Interpreting Mixed Effects Models

$$\text{Accuracy} \sim \underbrace{\text{ArgForm} + \text{Perplexity}}_{\text{Fixed Effects}} + \underbrace{(1 + \text{Perplexity} \mid \text{LLM})}_{\text{Random Effects}}$$

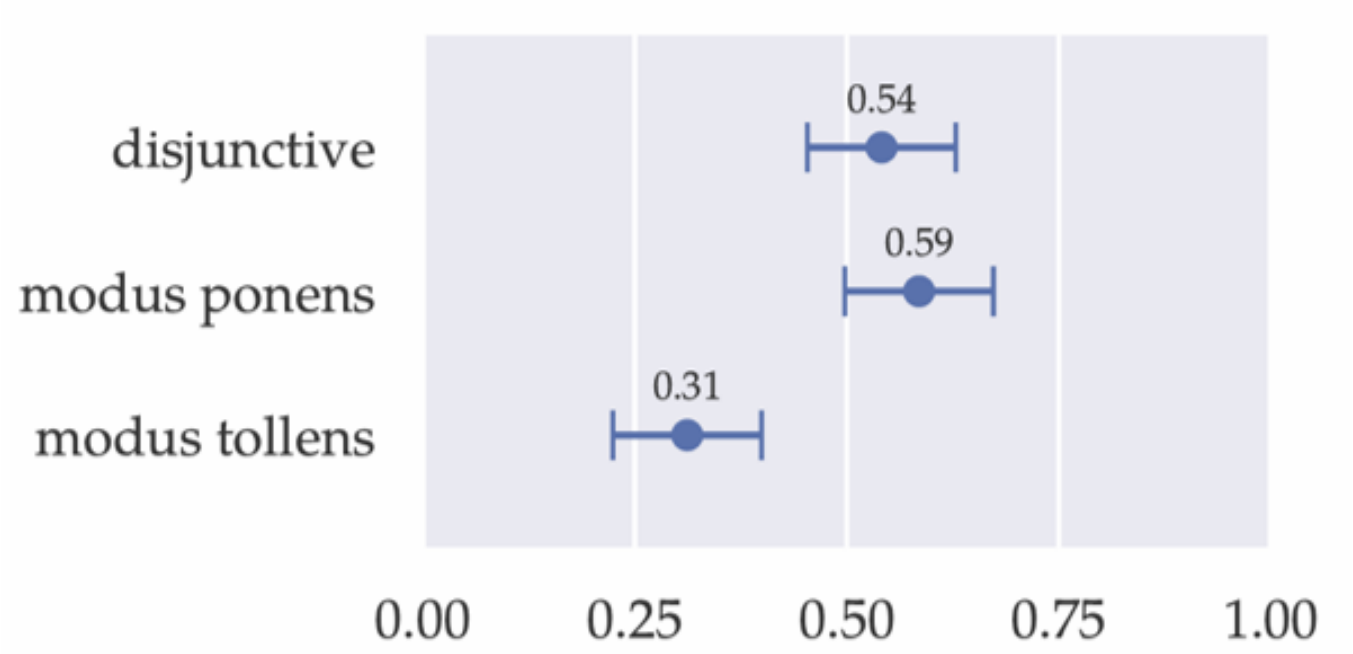
- **ArgForm** (categorical): the argument form of the reasoning question.

$$\{\varphi \vee \psi, \neg\varphi\} \vdash \psi, \qquad (\vee^{\mathsf{L}})$$
$$\{\neg\varphi \to \psi, \neg\varphi\} \vdash \psi, \qquad (\to^{\mathsf{L}}\text{; modus ponens})$$
$$\{\varphi \vee \psi, \neg\psi\} \vdash \varphi, \qquad (\vee^{\mathsf{R}})$$
$$\{\neg\varphi \to \psi, \neg\psi\} \vdash \varphi. \qquad (\to^{\mathsf{R}}\text{; modus tollens})$$

- **Perplexity** (continuous): the perplexity of the language model.
- **LLM** (categorical): the language model.

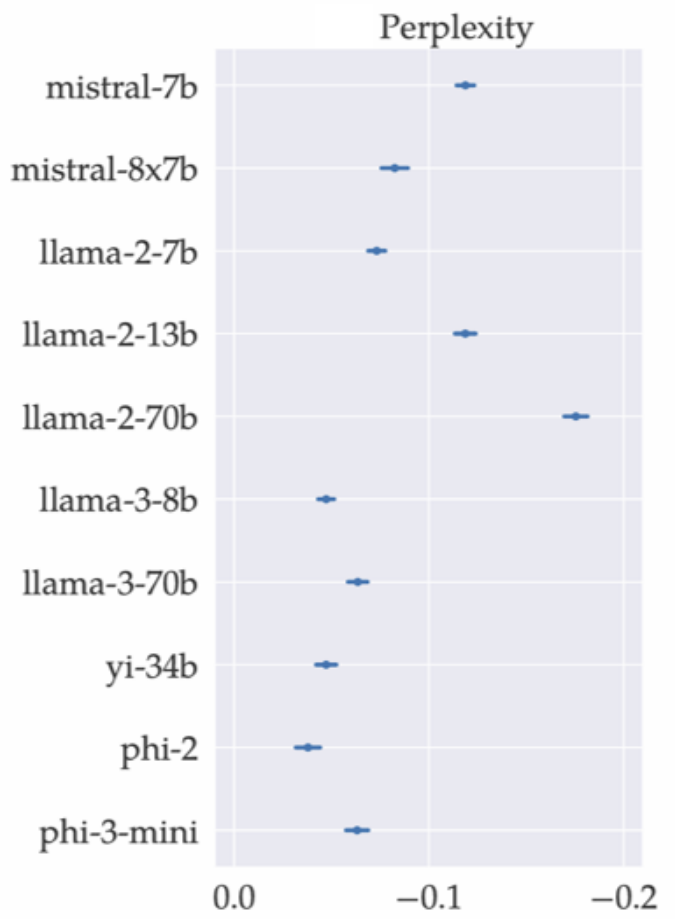
Interpreting Results: Argument Forms

$$\text{Accuracy} \sim \underbrace{\text{ArgForm} + \text{Perplexity}}_{\text{Fixed Effects}} + \underbrace{(1 + \text{Perplexity} \mid \text{LLM})}_{\text{Random Effects}}$$



Interpreting Results: Perplexity

$$\text{Accuracy} \sim \underbrace{\text{ArgForm} + \text{Perplexity}}_{\text{Fixed Effects}} + \underbrace{(1 + \text{Perplexity} \mid \text{LLM})}_{\text{Random Effects}}$$



Next

Computational Psycholinguistics with Language Models

(Part II - Training)

Please complete the course survey: