Human Language Processing

And Comparing LMs to Humans

Aaron Mueller

CAS CS 505: Introduction to Natural Language Processing

Spring 2026

Boston University

Admin

- Course evaluation season is here!
  - If we can get a >75% response rate, I'll add 1 point of extra credit to everyone's final report grades, and I'll bring coffee/tea and donuts the last day of class

Disclaimer

- Comparing humans to LMs gets *very* hairy very quickly.
  - We are exposed to very different amounts of data
  - We are based on very different architectures
  - Questions like “Is AI conscious?” Or “Can AI systems know/understand anything?” are not the domain of today’s lecture

The Classical View of Cognition

- Information is represented by strings of symbols (discrete concepts), analogous to how variables are represented in programs
- *Example*: past tense formation

If verb is regular:

stem + **-ed**

walk -> walk**ed** jump -> jump**ed** crystallize -> crystalliz**ed**

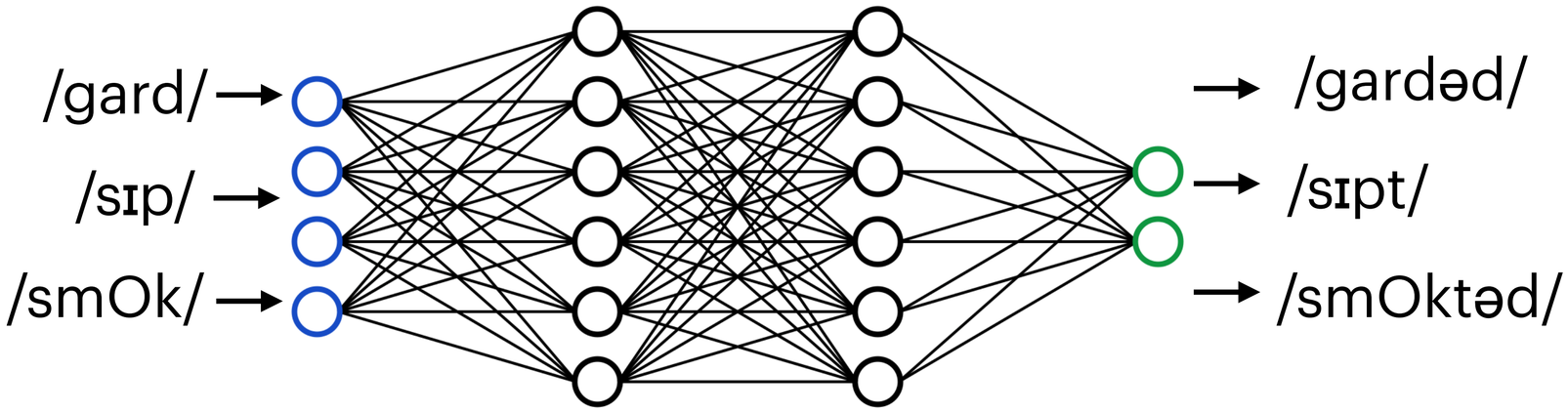
If verb is irregular:

Store as exception in a mental lexicon

“Words and rules” view

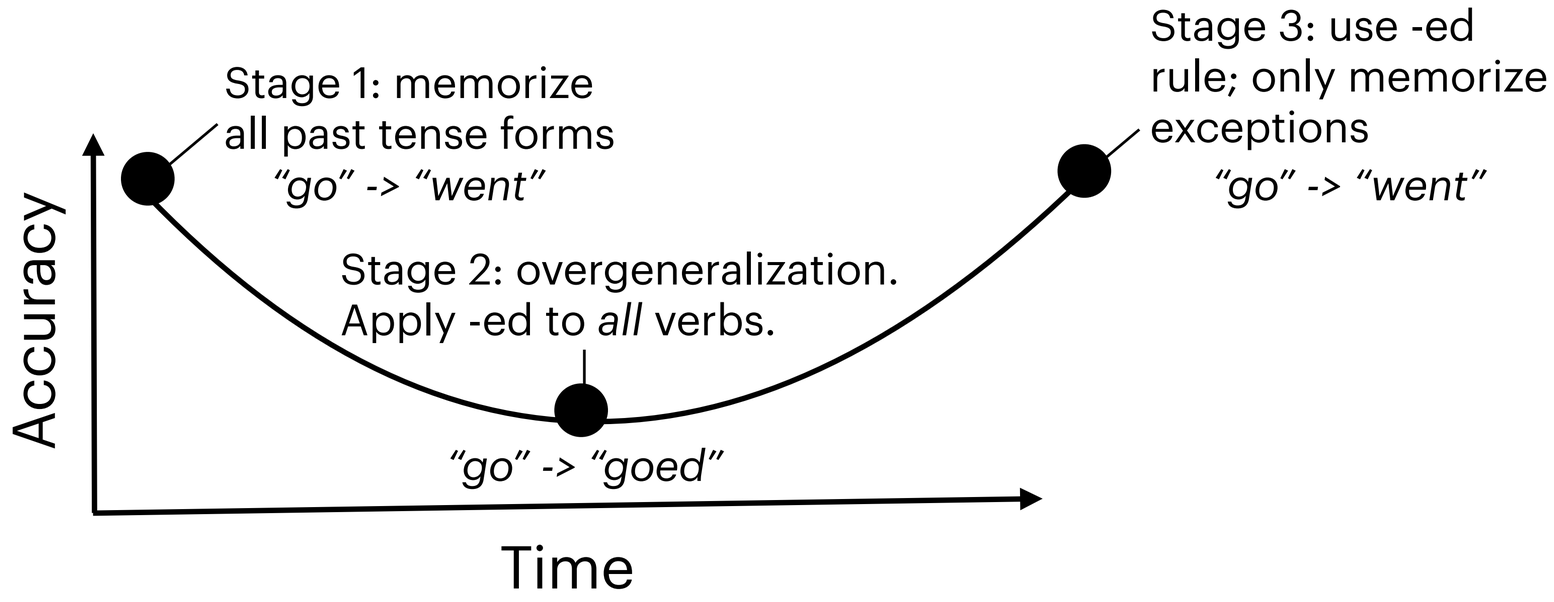
Connectionism

- Connectionists hold that neural networks may provide a new framework for understanding the nature of human cognition
    - No explicit symbols necessary: everything is continuous



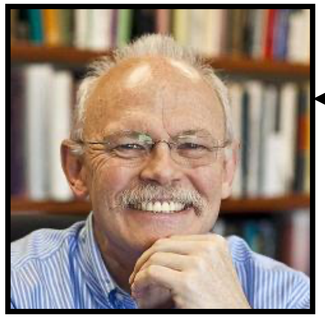
The Controversial History of ML in Cognitive Science

- Rumelhart & McClelland [1986]: trained a neural network to form English past tense verbs given the present tense verb
  - Generalized to verbs not in the training set!
  - Better yet: it captured the *trajectory* of learning in humans!



The Controversial History of ML in Cognitive Science

The Past Tense Debate

Rumelhart & McClelland: neural networks can be effective models of language learning and use. The model captures human learning trajectories without symbols.

Pinker & Prince: but the model fails to generalize to simple regular past tense verbs! Rules and symbols are necessary for robust generalization, and neural nets can only *approximate* this.


[Paraphrased by me.]

Why care?

- The past tense debate was a proxy war over a bigger question: does the mind manipulate symbols, or is it sub-symbolic/continuous?

- Modern LLMs reignite similar debates: do LLMs “understand” anything in a robust way, or are they sophisticated pattern matchers? Do they need symbols?

- Cognitive science and psycholinguistics can inform LM architectures, evaluation, and what “success” means

Outline

1. How do humans process a sentence?
   - How comparable are LM and human reading “abilities”?

2. Developmentally plausible language models

3. Training LMs to be cognitive models

Cloze Tasks

- **Cloze task**: readers a passage, ask them to fill in the blanks:

  I went to the _____ and bought milk and eggs. I knew it was going to rain, but I forgot to take my _____, and ended up getting soaked on the way _____.

- **Cloze probability:** the probability of a word occurring in a particular context:

  (a) My brother came inside to _____

  (b) The children went outside to _____

- "Play" is plausible in both sentences, but would be chosen by humans ≈90% of the time in (b) and would almost never be the first choice for (a) for most people.

Predictive Processing

- Humans don’t wait until they’ve read a whole sentence to start understanding it!
- We’re *always* filling in the gaps, and trying to predict what someone will say next so that we can get a head-start on processing it
- How do we know this?

Garden Path Sentences

"The old man the boat."

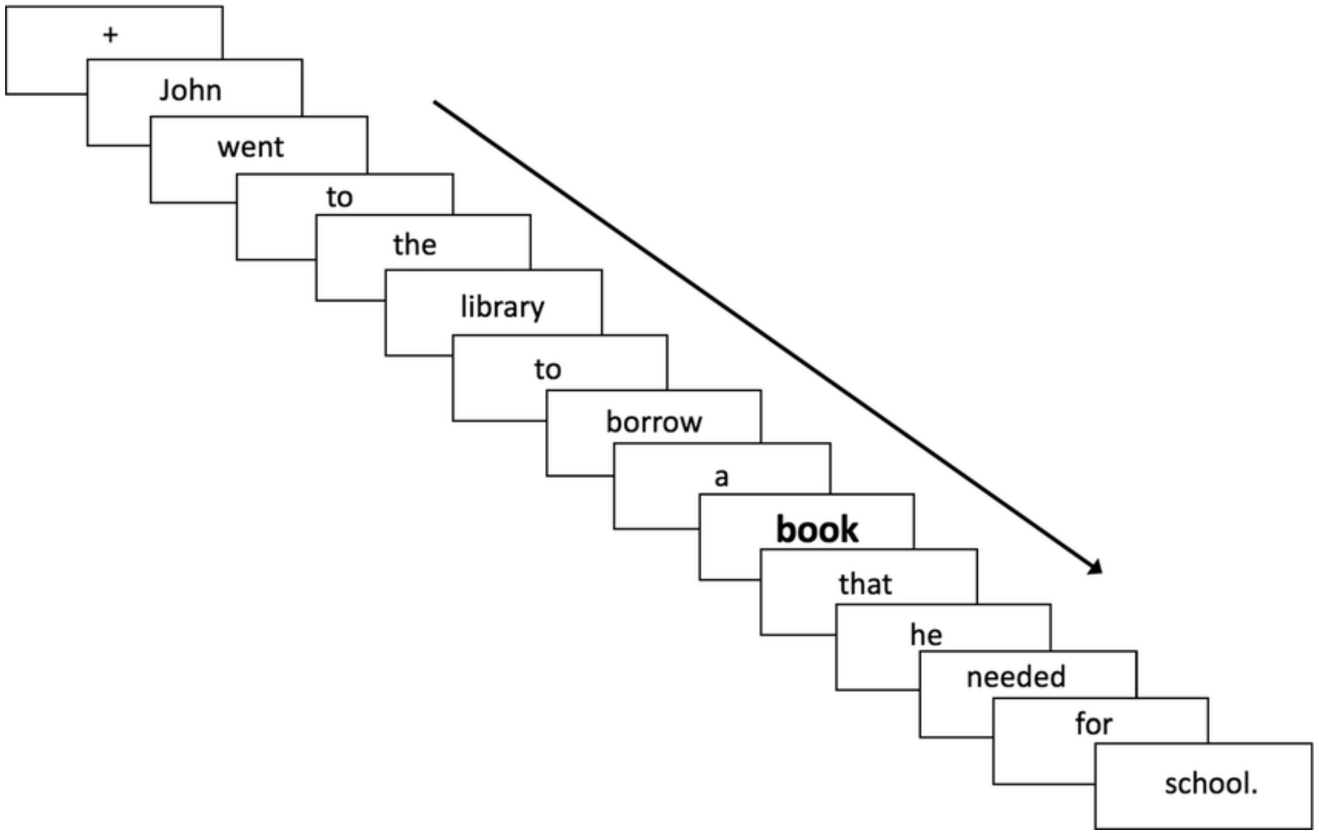
"The complex houses married and single soldiers and their families."

- Why is this so hard to parse?
- We read "The old man", and start predicting that a verb is likely to come next
  - But then we see a noun phrase!
  - We were "led down the garden path", so to speak; "man" is actually a verb
- The fact that these are hard to parse is often taken as evidence of predictive parsing

Self-paced Reading Task

- Give humans a sentence one word at a time, and ask them to press a button to see the next word

- Time how long the participant takes on each word



Human Performance in Self-paced Reading

Pretend you’re pressing a button to get each word:

```
The  shop  sold to  the  man
```

```
The  shop  sold to  the  bank  was  old
```

```
The  necklace  sold to the  bank  was old
```

People are very sensitive to word frequency in reading time tasks!

What happens to a self-paced reader when they face a garden-path effect?

Why is the third sentence not as likely to garden-path readers?

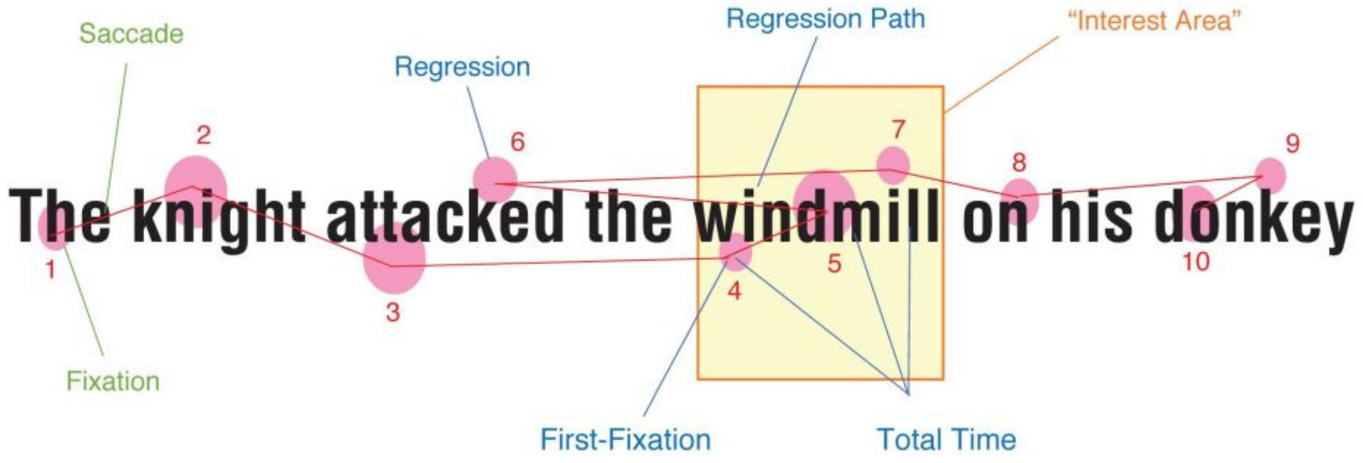
When does semantics come in?

- Are we constantly using semantics to help us process sentences, or only as needed?
- **Hypothesis 1:** we use syntactic heuristics
  - Get the parse first, then interpret sentence. Backtrack if it doesn't make sense
  - *Heuristic:* when an NP starts the sentence, it's the subject
- **Hypothesis 2:** rely on syntactic probabilities using head words
- **Hypothesis 3:** consider full semantics of a constituent as soon as it's built
- **Hypothesis 4:** consider full semantics of a constituent even before it's built

Eye Tracking

- Self-paced reading data isn't reliable enough to answer these questions.
- Other techniques like brain imaging are too coarse.

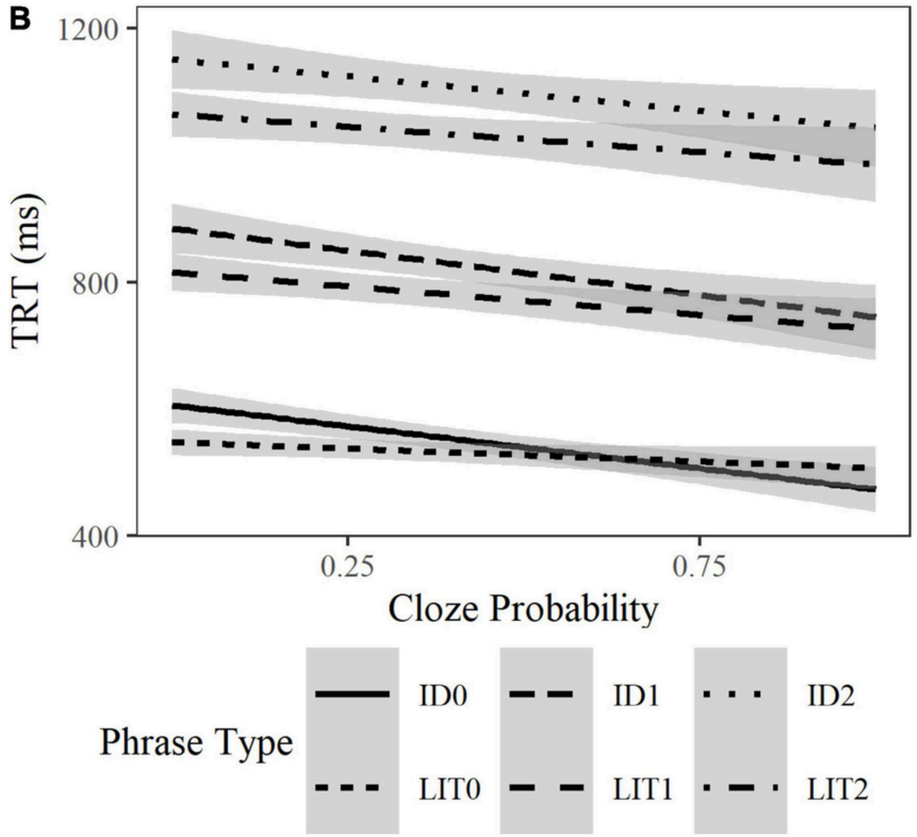
- We can **track people's eye movements** as they read a passage
  - E.g., garden path effects should result in backtracking
  - Especially hard-to-understand words or structures should be fixated upon



Reading Times vs. Probability

- Word reading times generally decrease as a function of cloze probabilities

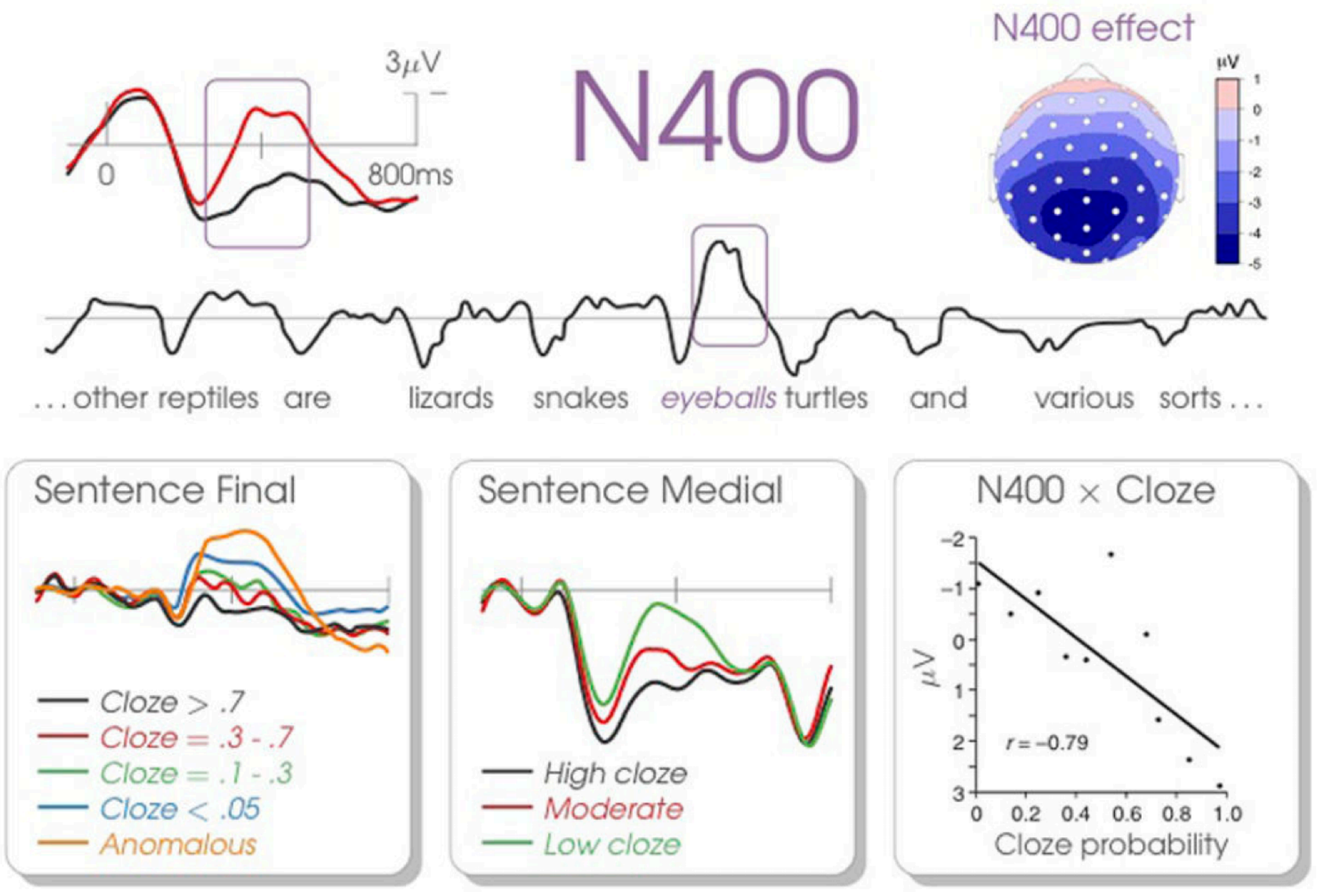
- The easier something is to read, the less time we spend looking at it
  - “Easier to read” -> more predictable



The N400

- We can take EEGs of humans and record various signals.

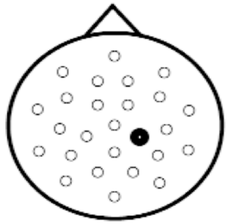
- One such signal is the **N400**: a signal which is highly correlated with semantic surprisal.
    - Spikes when a surprising word is encountered

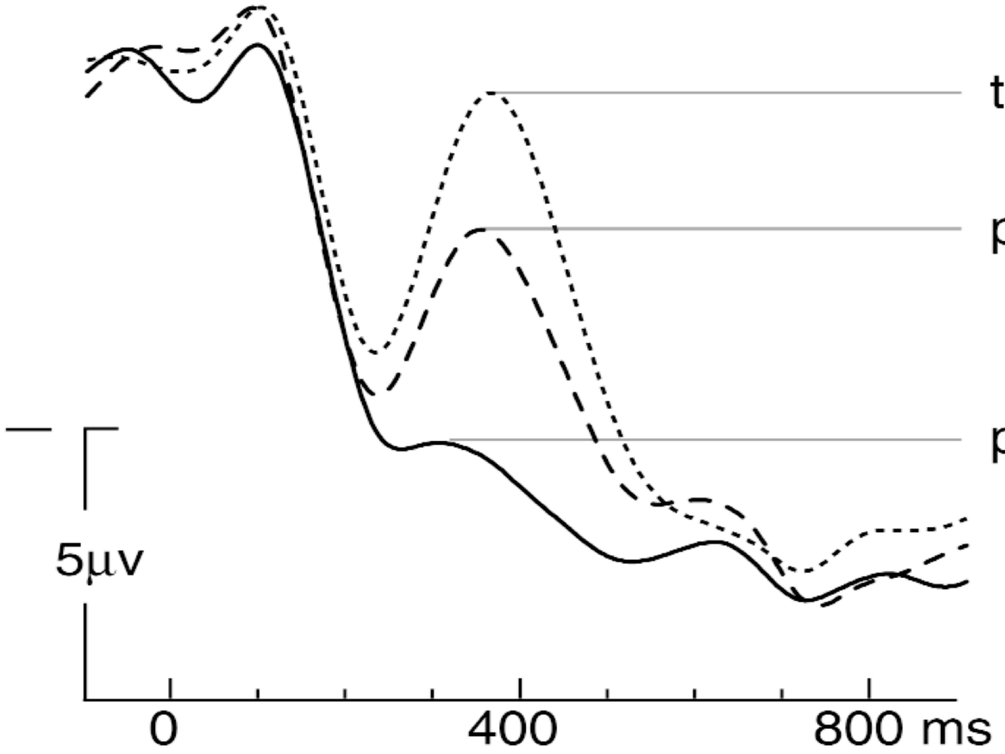


The N400

‘They wanted to make the hotel look more like a tropical resort.
So along the driveway they planted rows of ...’

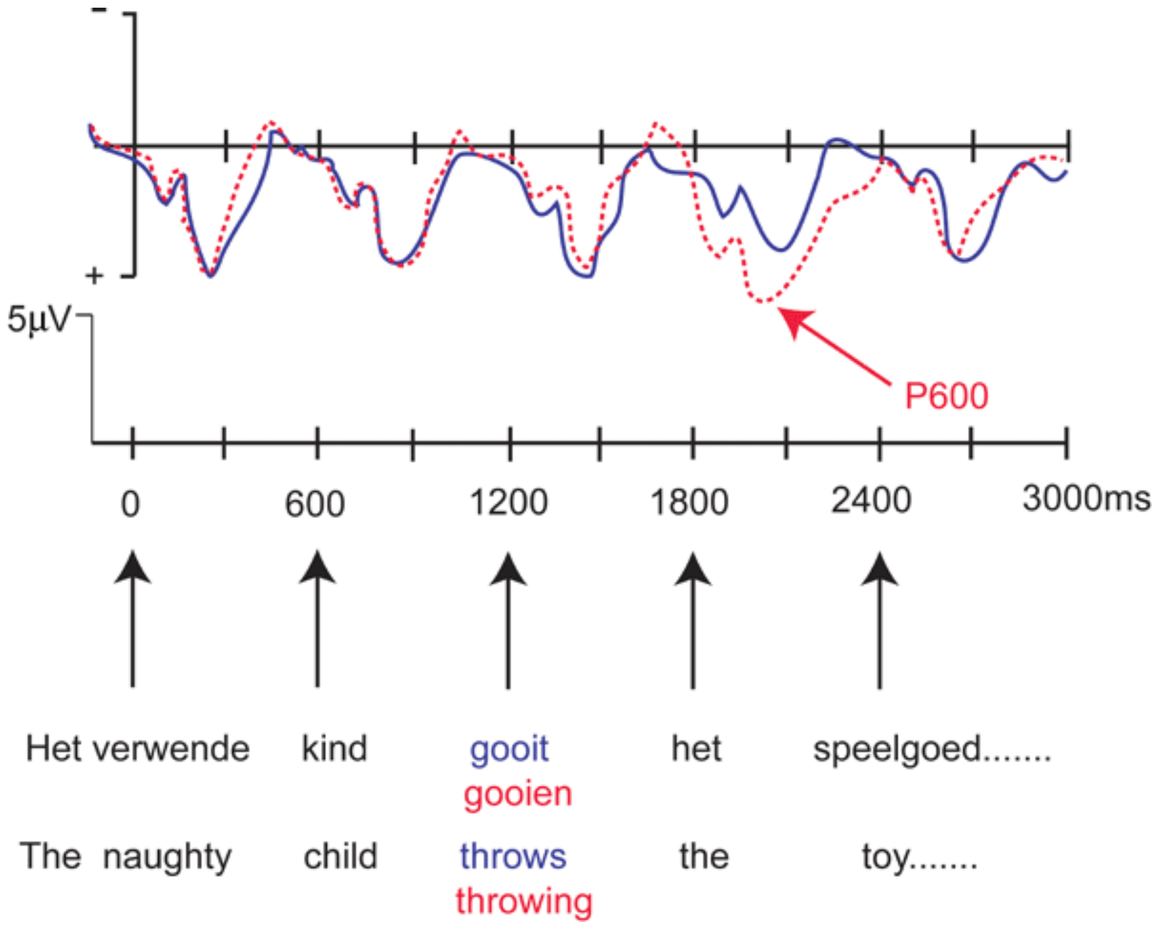
R. medial
central





The P600

- The N400 is often characterized as tracking *semantic* surprisal.
- Meanwhile, the P600 is often characterized as tracking *syntactic* surprisal.



Surprisal Theory

- **Hale [2001]; Levy [2008]**: the time required to comprehend a word is based on its predictability
    - Predictability is often quantified as **surprisal** (negative log-probability given context)

- We don't know the "real" surprisal, but we can estimate the surprisal of a word given context given a language model:

$$\text{Surprisal} = -\log_2 p(w_n \mid w_1, \ldots, w_{n-1})$$

Psychometric Fit

- Idea: train a model to predict likelihood of a word without considering its surprisal first. Then add the surprisal, and see how much better the model gets

Baseline: fit a regression on word length, word frequency, position in text, whether the previous word was fixated upon by readers

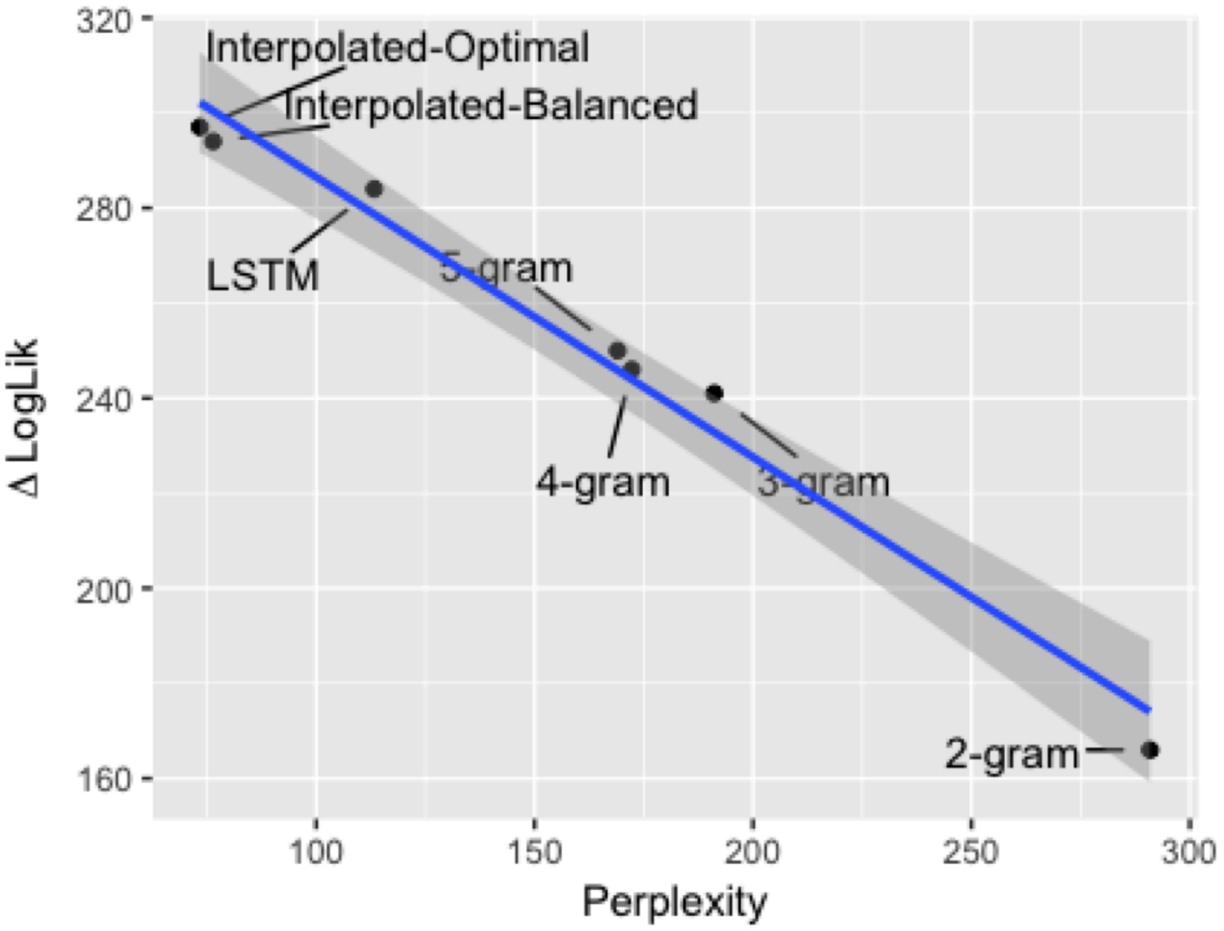
Full: fit a regression on the baseline variables *plus* LM surprisal on current and previous words

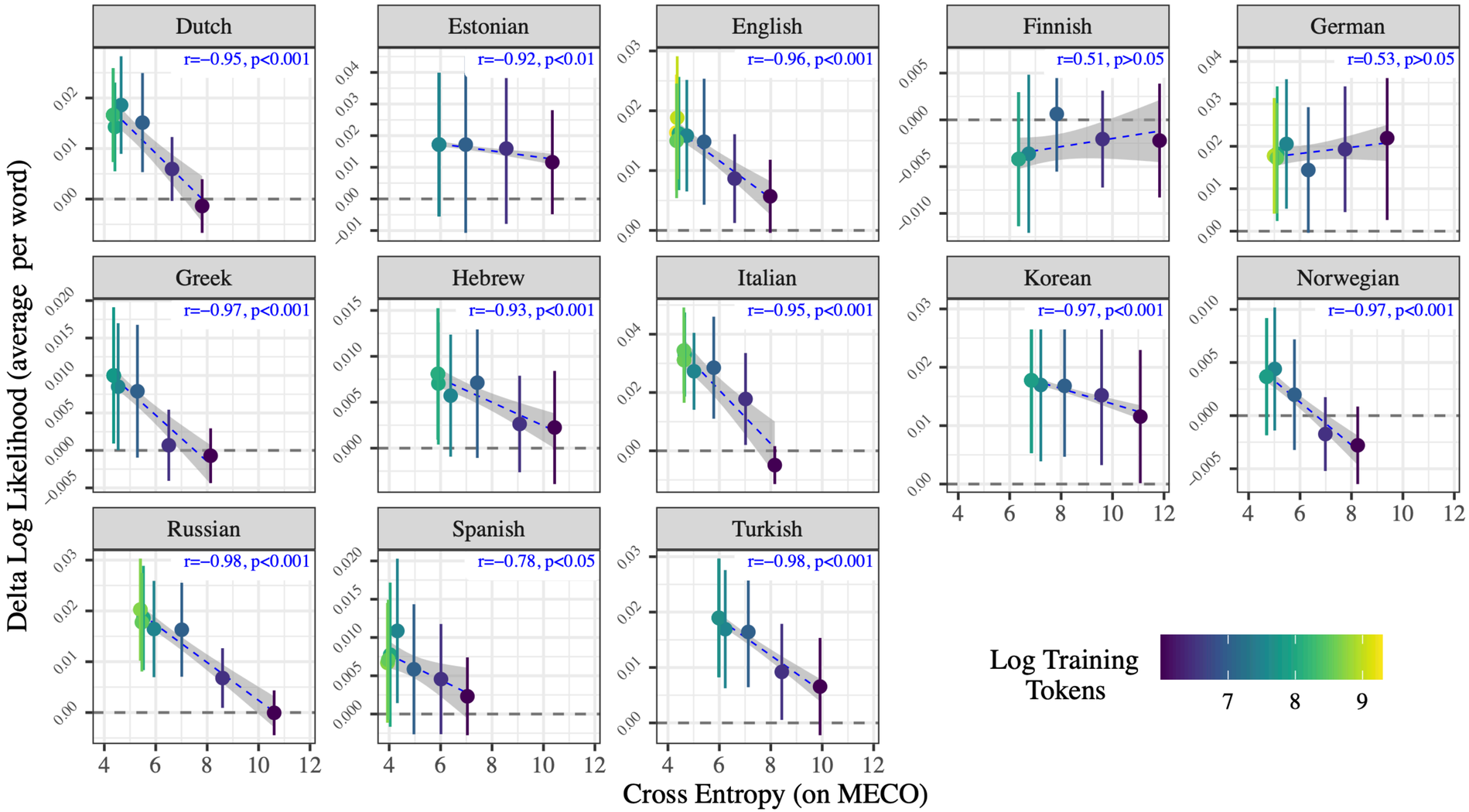
$$\Delta\text{LogLik} = L_{\text{full}} - L_{\text{baseline}}$$

Psychometric Fit

- ΔLogLik increases as a *linear function* of language model quality

- So it seems like the psychometric fit of LM surprisals to human reading times generally improves with the quality of LMs

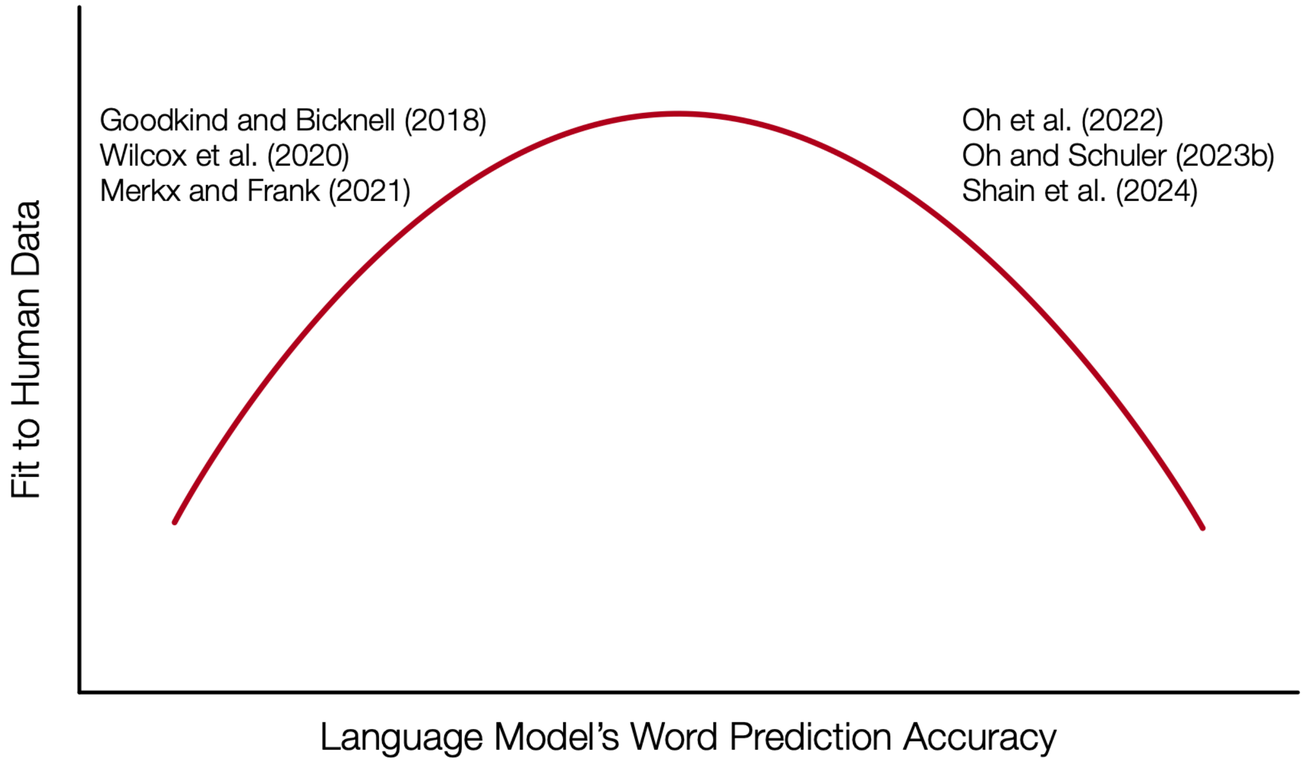




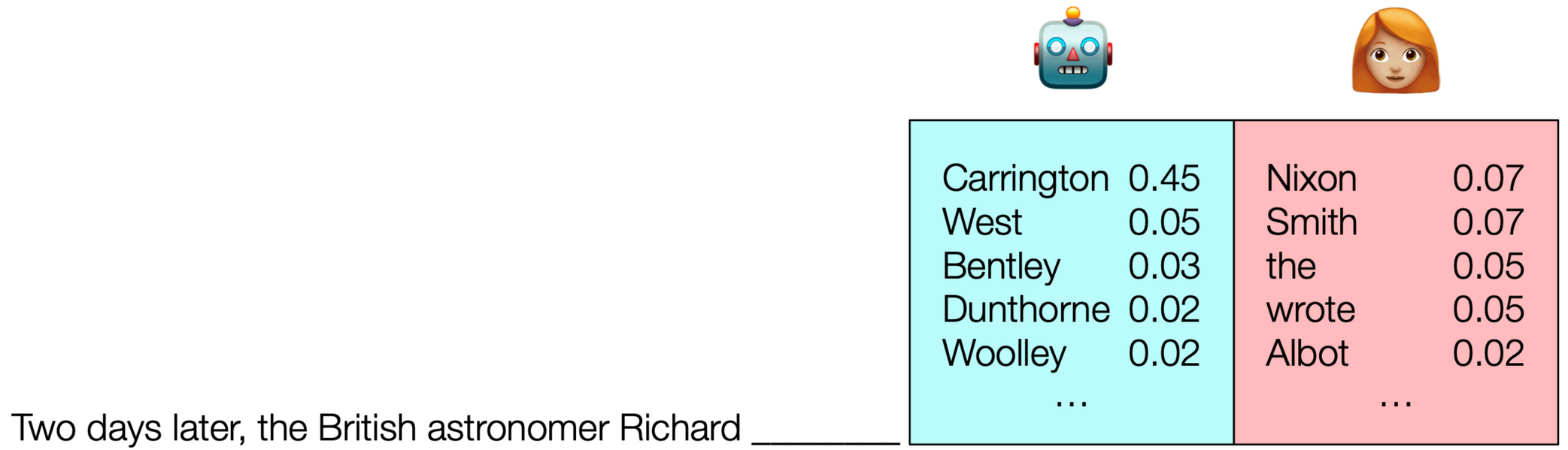
The predictions of surprisal theory hold in many languages!

Psychometric Fit

- But actually, it's complicated: as LMs improve, their psychometric fit improves up to a point, but then begins to decrease. Why?



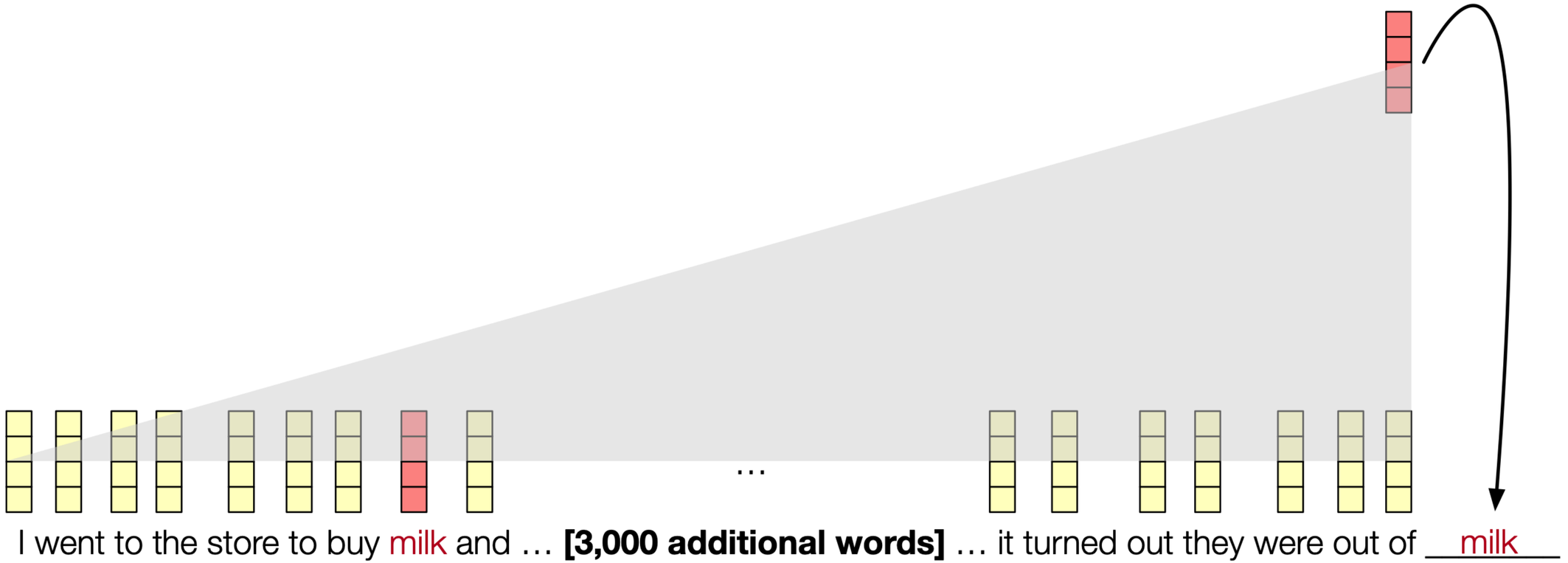
LMs are superhuman word predictors.



LMs trained on very large datasets assign high probability to (often correct) words that humans find surprising.

Think of rare names and technical terms.

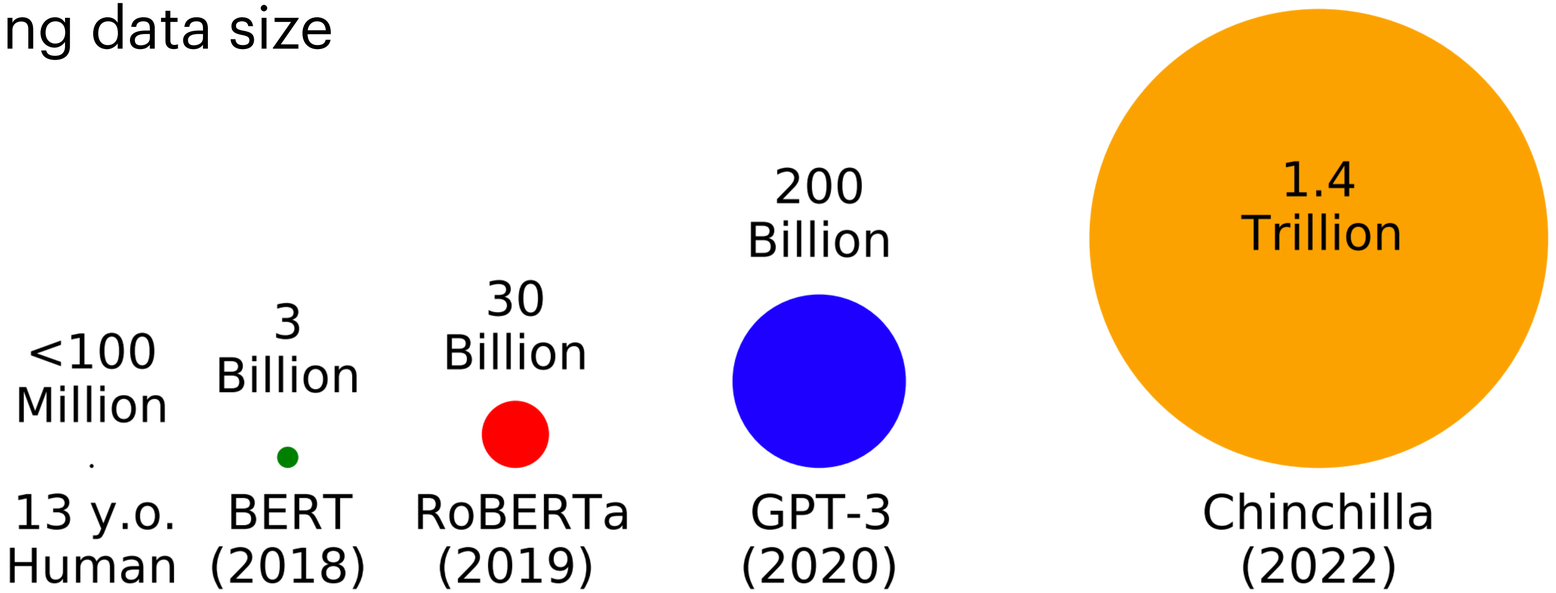
LMs have superhuman context memory.



Transformers can track extremely long-distance dependencies. Humans have bounded memory/attention.

Training More Human-like LMs

- The most obvious difference between LMs and humans is the massive difference in pre-training data size

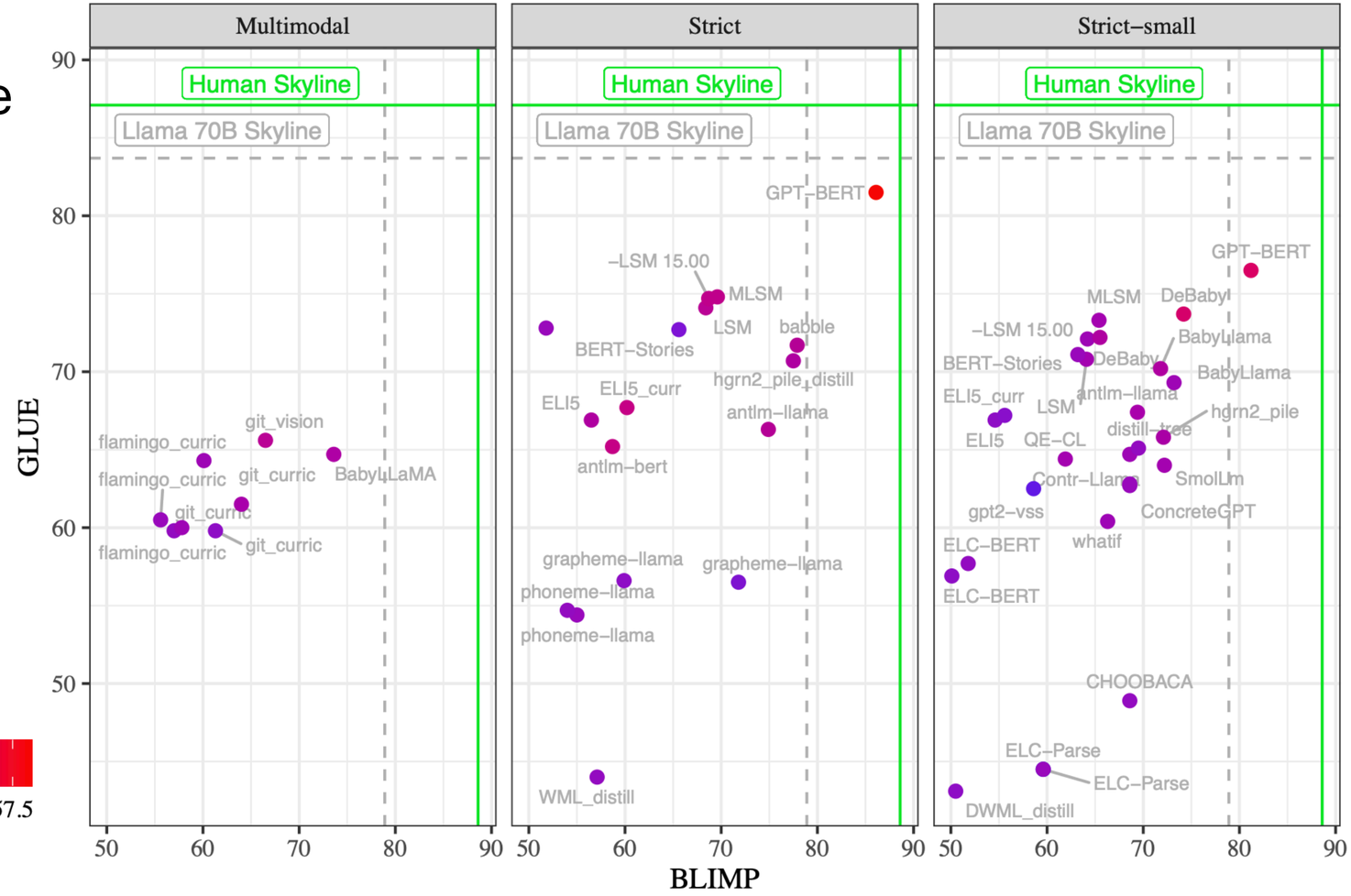


- One idea: train LMs on a more human-like amount of data!

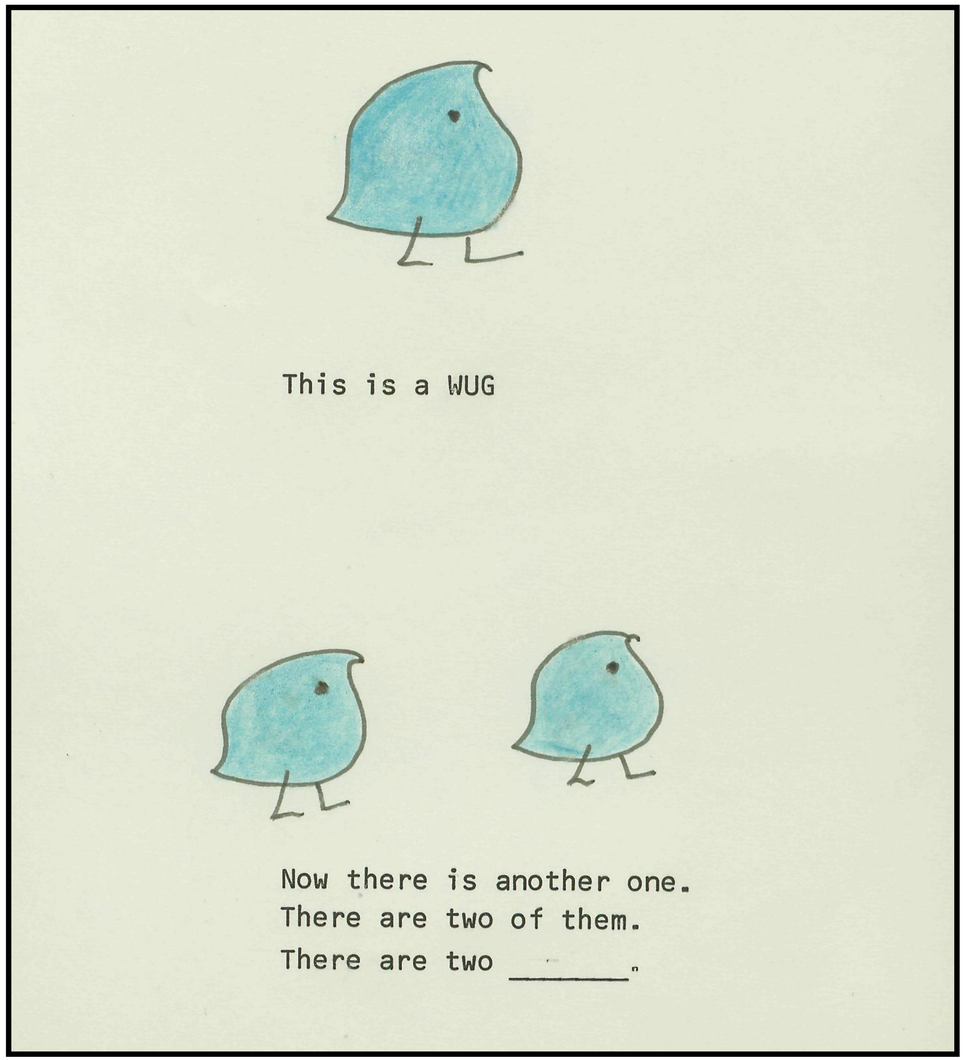
BabyLM Challenge

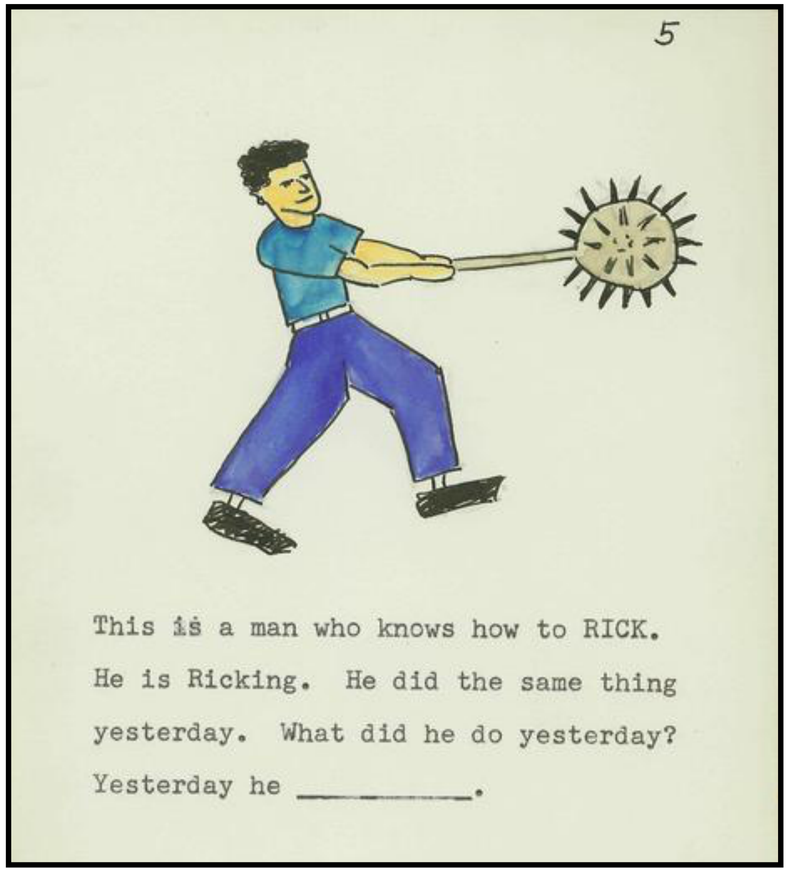
Sample-efficient pretraining on a developmentally plausible corpus

- Have researchers from the NLP community compete to train the best language models they can given <100M words of pre-training data



Evaluating Morphological Learning



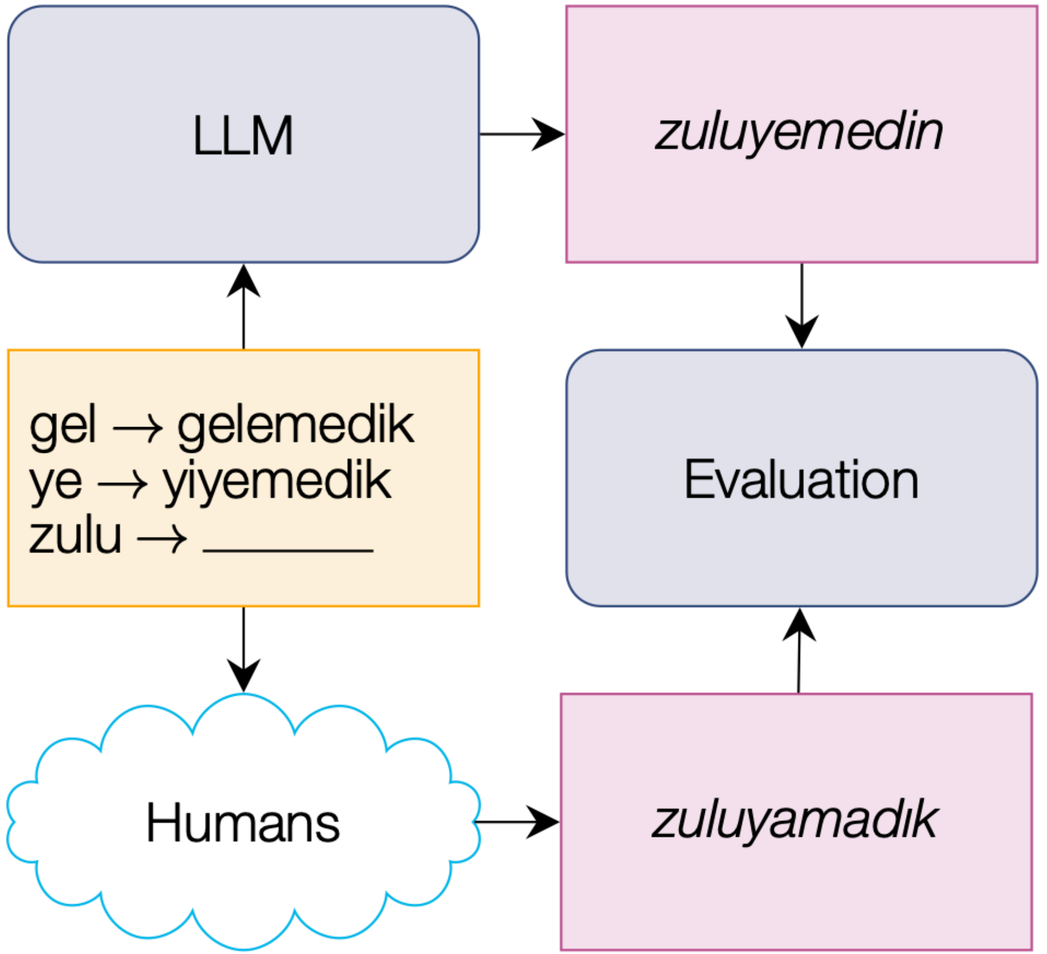


Evaluating Human Likeness

Wug Tests

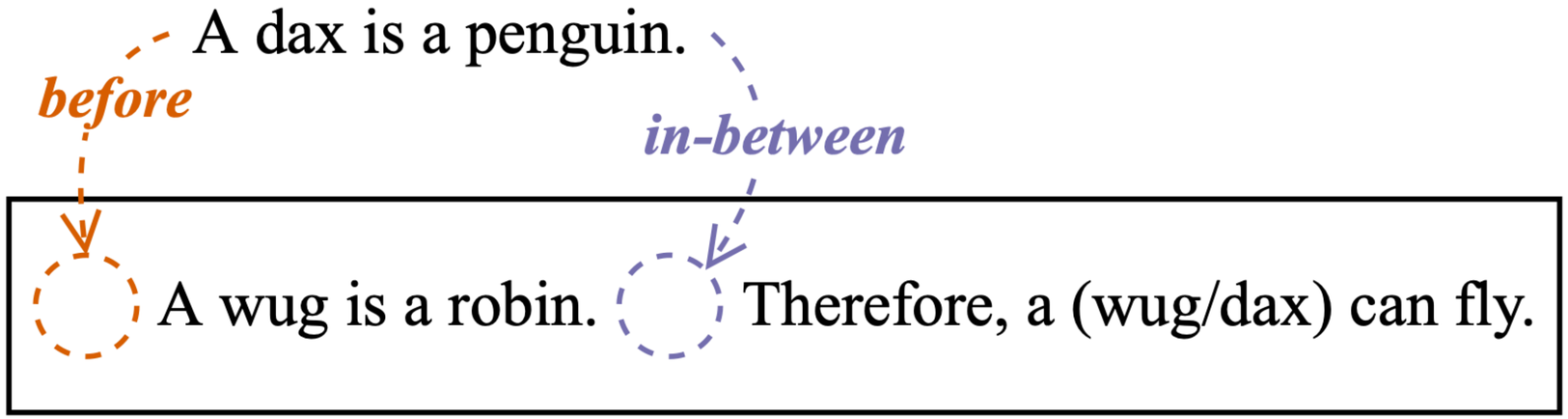
- Children learn how to inflect new plurals by 4 to 5 years old
  - Give them a nonce word (a novel word that doesn't exist), and ask them to pluralize it
  - Wug -> wugs

- Weissweiler et al. [2023] propose a benchmark to measure morphological generalization in LMs



Evaluating Human Likeness

Property Inheritance



- Children begin to understand property inheritance at around 3-4 years old.
- Misra et al. [2023] propose a benchmark to measure LMs' capabilities in handling property inheritance: COMPS

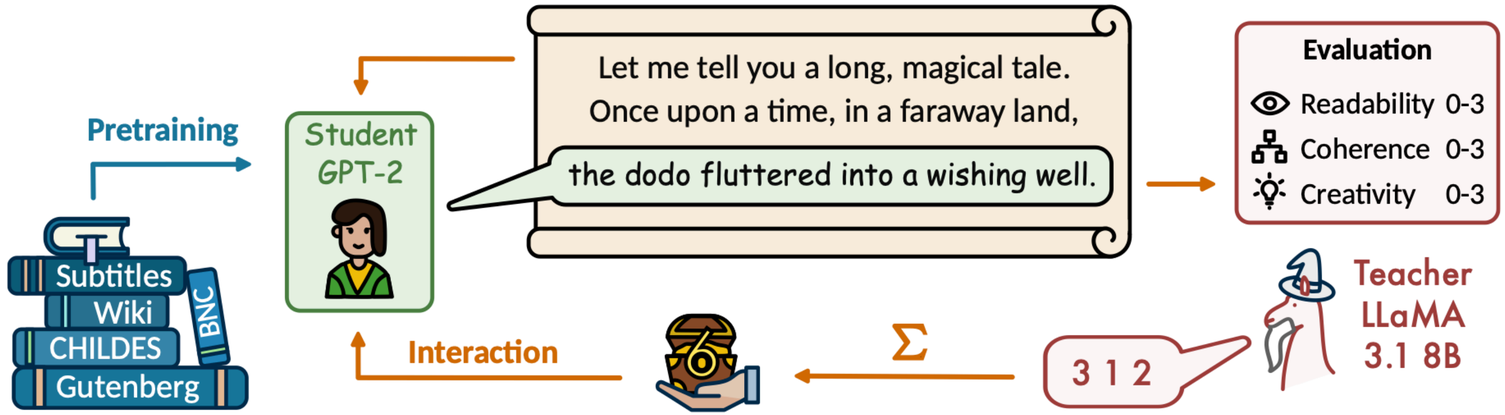
Model	Entity Tracking	WUG Adjective Nominalization	WUG Past Tense	COMPS	Reading	AoA
Baseline-gpt-bert-base-mixed (mntp)	39.9	41.2	27.1	59.7	6.3	22.3
Baseline-gpt-bert-base-causal-focus (causal)	30.9	63	26.8	58.3	5.7	10.7
Baseline-gpt-bert-base-causal-focus (mntp)	41.9	39.9	32.7	60	6.1	11.5
Baseline-gpt-bert-base-mixed (causal)	33.2	58.2	17.9	56.9	6.2	10.4
Baseline-gpt-bert-base-masked-focus (mntp)	41.5	35	23.2	58.3	6.3	9.6
Baseline-gpt-bert-base-masked-focus (causal)	31	51.1	25.1	55.7	6	12.8
blalm-100m-dynmod-bounded	22.2	47.5	37.5	58.3	0.8	8.6
simple_diffusion_cosine	40.8	49.6	15.4	56.4	7.4	-22
babylm-baseline-100m-gpt2	31.5	50.2	7.3	56.2	5.5	5.3
CLASS-IT_140M	19.9	60.1	9.6	56	0.3	11.8

Even the best BabyLM models are a long ways from human-like language learning.

What's missing from BabyLM?

- Children learn in multimodal and interactive settings
  - They receive feedback
  - They can ask questions
  - Language is grounded in actions, sounds, and visuals

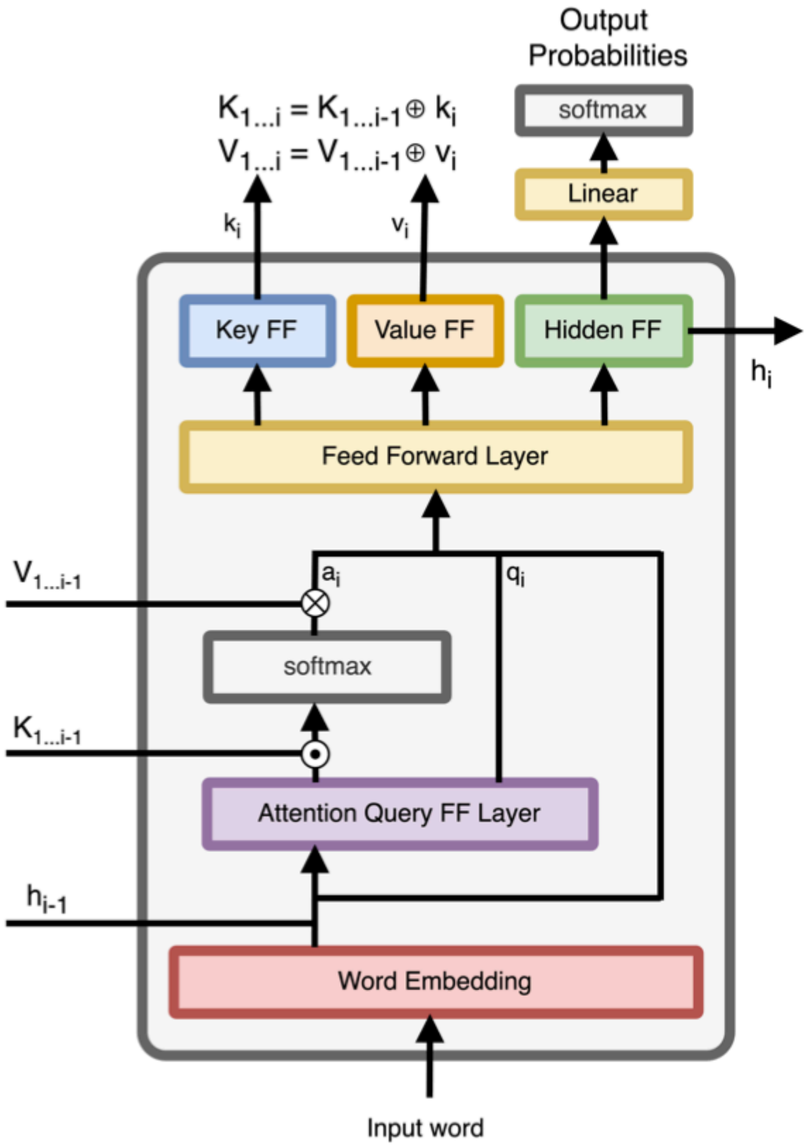
- Some are trying to fix this!



Addressing Superhuman Memory

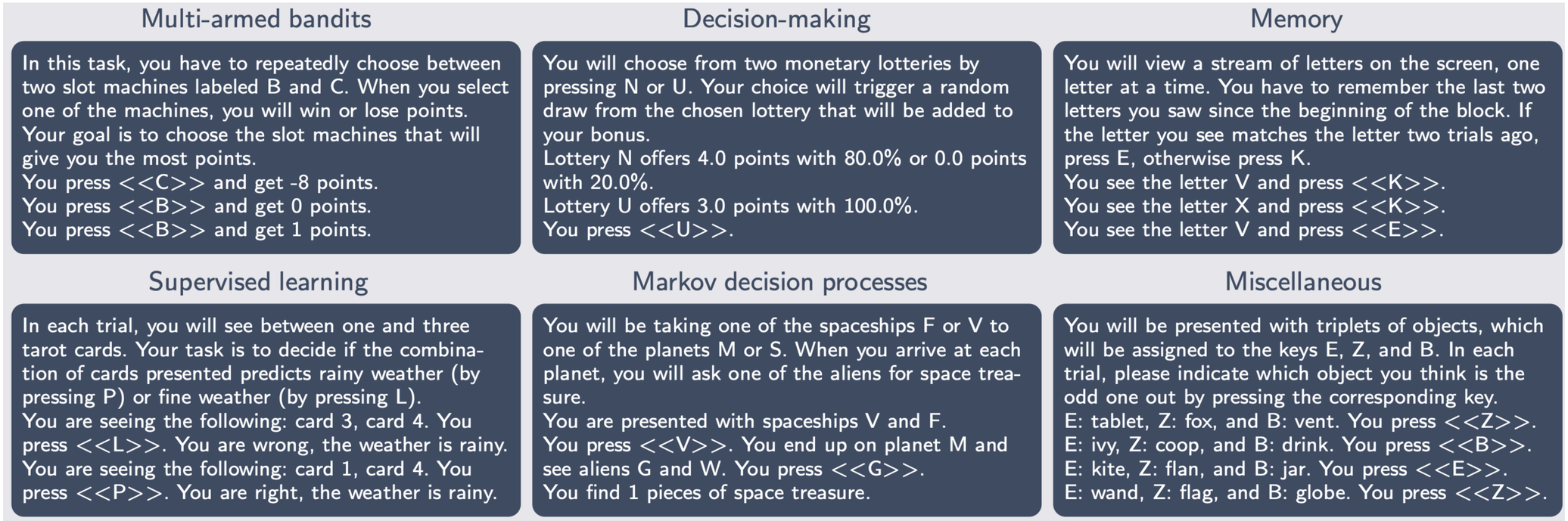
- Can add human-like memory constraints to Transformer-based models

- Better yet: maybe we should abandon Transformers if our goal is to model human language processing
    - Recurrent architectures are seen by many as more appropriate

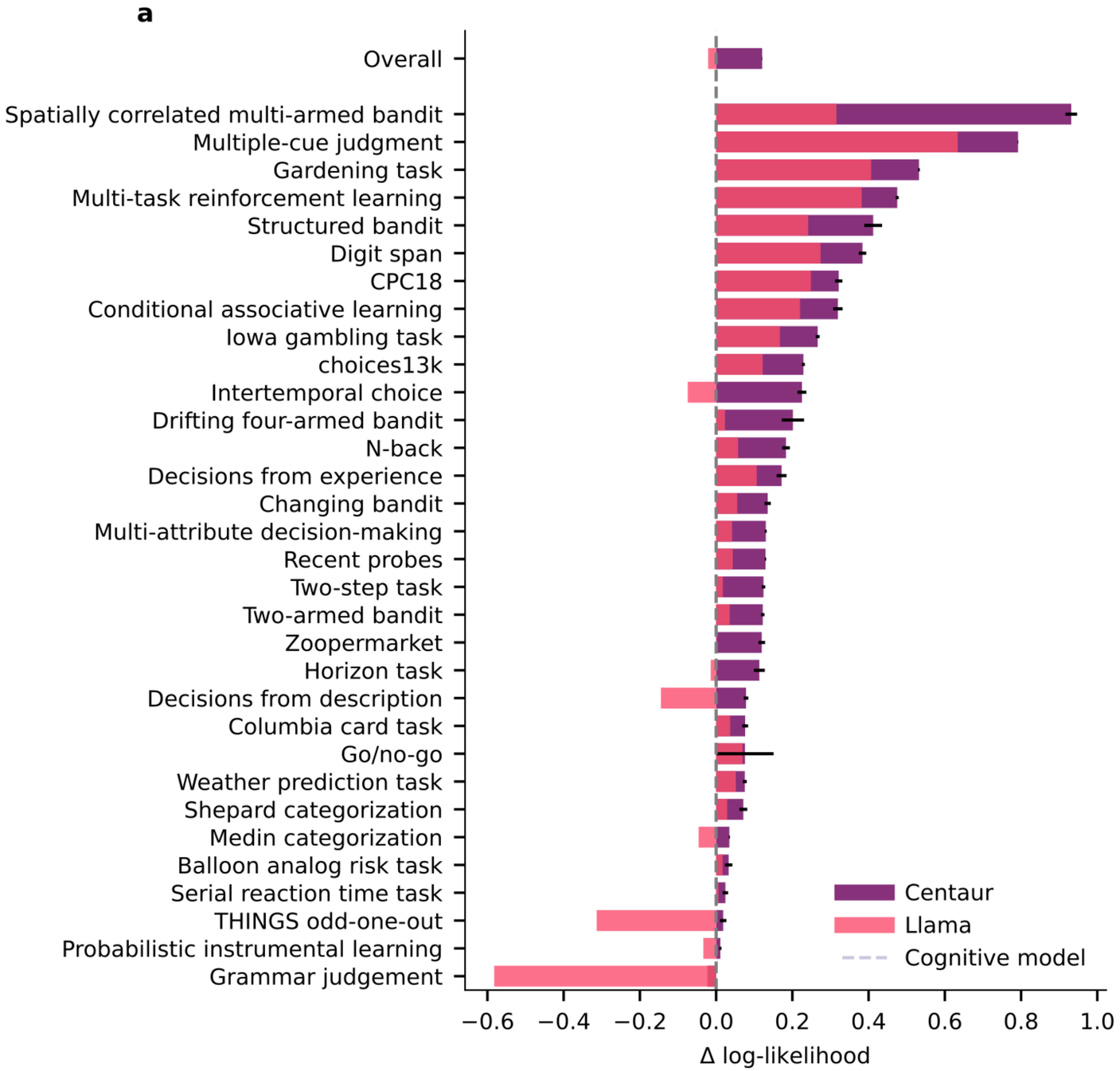


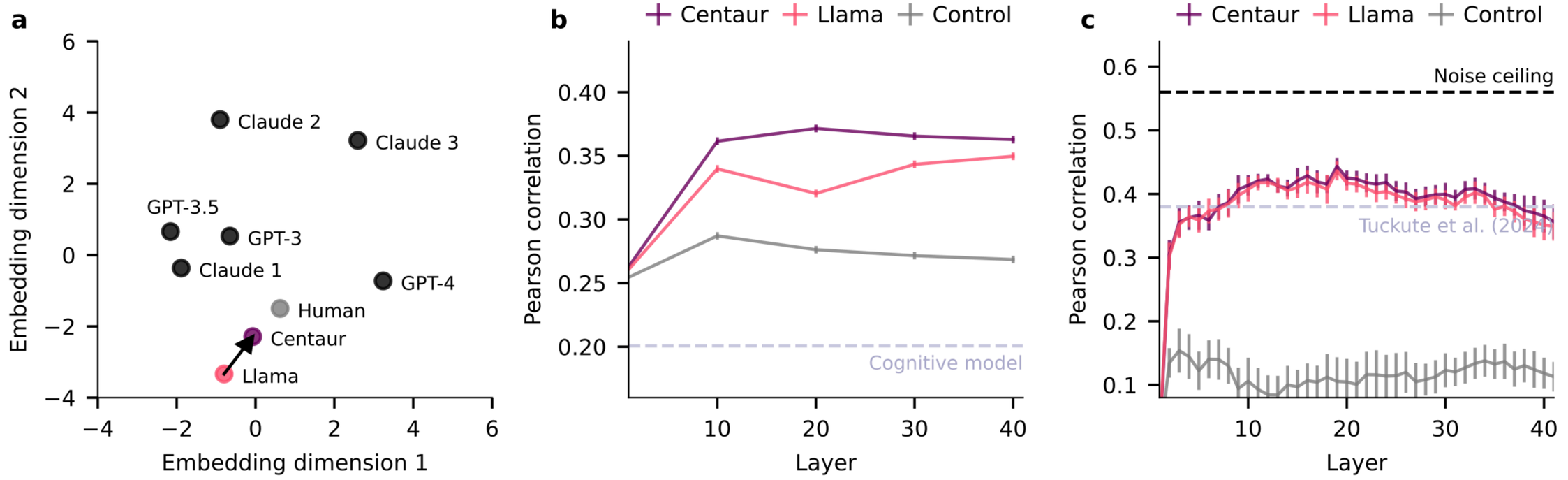
Fine-tuning for Human Likeness

- **Centaur:** a language model (Llama) fine-tuned on Psych-101
  - Data from 60,000 participants performing over 10M choices in 160 experimental settings



- Centaur’s output distributions better align with human choices across these task settings.
- For most tasks, Llama is fine, but Centaur is better.
- For others, Llama is not human-like at all, but Centaur is at least as good as a cognitive model.





- Human brain signals can be better decoded from the representations of Centaur compared to Llama
- More recent models are not necessarily closer to human representations

Critiques of Centaur

- Is Centaur actually simulating cognition (probably not), or just pattern-matching over psych experiment formats?
- Does fitting a behavior mean that a model uses human-like mechanisms?
- How do we resolve the difference between *behavioral* alignment vs. *mechanistic* alignment?

Where to?

- Grounded/multimodal learning
- Memory-constrained architectures
- Interpretability as a tool for cognitive modeling
  - Can we find circuits that correspond to human processing stages?

Summary

- Humans process language *predictively* and *incrementally*
- LM surprisal predicts human reading time measures surprisingly well
  - Better LMs can have better predictive power—but only up to a point
- LMs are superhuman in ways that make them imperfect cognitive models
- Matching human processing means matching human *constraints*, not just capabilities

Next Week

- Tue.: Medical NLP and interpretability
  - Guest speaker: Hiba Ahsan
  - From Northeastern, in Byron Wallace's group

- Thu.: Final project help session
  - I'll start with FAQs and examples of cool projects
  - We'll add Qs to a shared doc in class, and I'll address the most common ones live (and give individual groups with less frequent questions help afterwards)
  - Free-form coworking afterwards
  - Food/drinks provided if you all do course evals :)