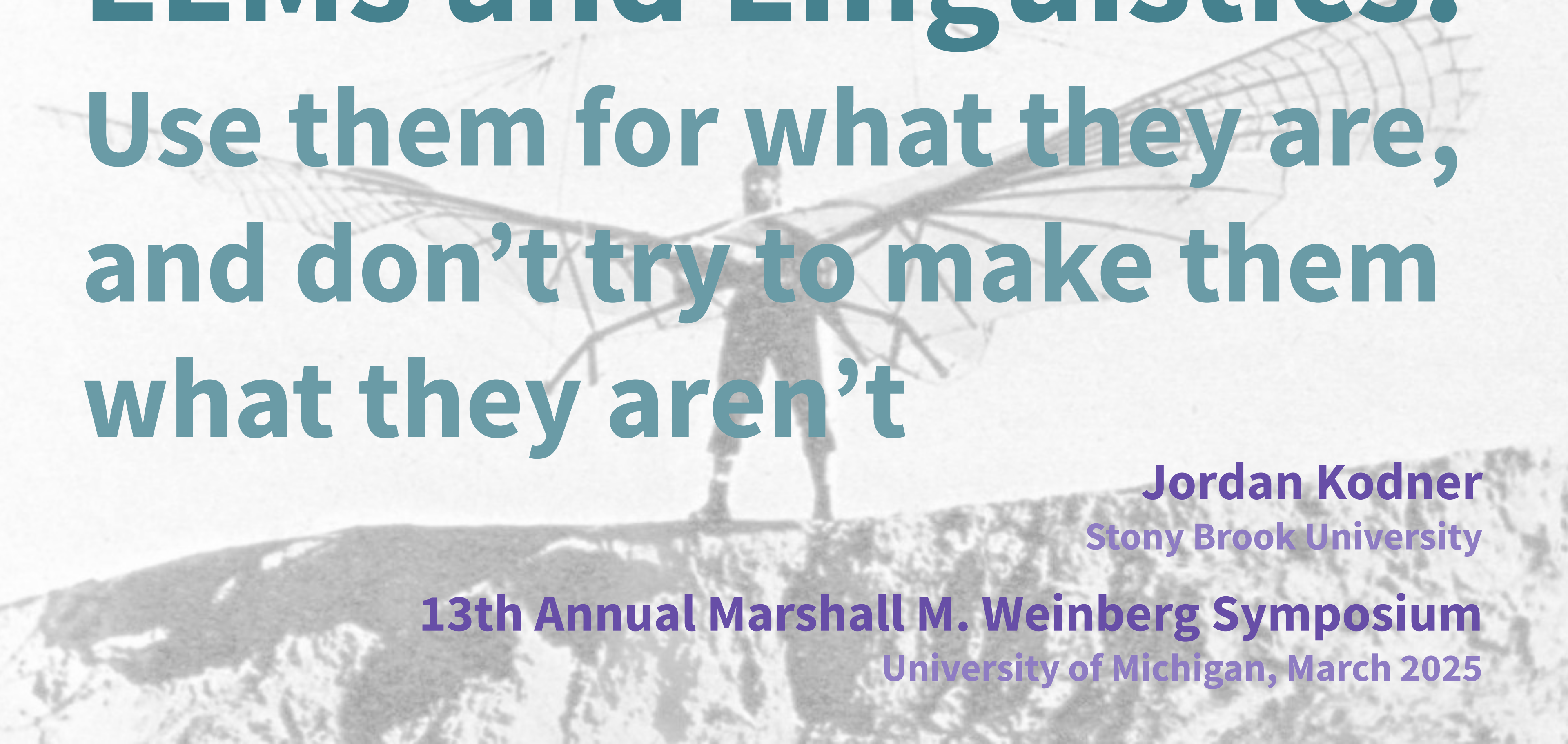


LLMs and Linguistics:

Use them for what they are, and don’t try to make them what they aren’t

Jordan Kodner
Stony Brook University

13th Annual Marshall M. Weinberg Symposium
University of Michigan, March 2025

Today's Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

What is a "Large Language Model?"

Most narrowly (*in sensu stricto*)

- Very large prompt-based models (2022-onward) - ChatGPT, Llama, Claude...

Broader

- Large transformer models (2018-onward) - BERT, GPT-2, T4...

Most broadly (*in sensu lato*)

- The full range of deep learning models for language data (2015-onward)

I do not use this term to refer to more distantly related research areas such as neural networks aiming for lower-level biological plausibility

Today’s Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

- My primary interest in cognitive science is working out the mental processes that underlie our capacity for language, especially language acquisition

Today’s Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

- My primary interest in cognitive science is working out the mental processes that underlie our capacity for language, especially language acquisition

→ My motivating question for today: Are LLMs insightful models of linguistic cognition?

- There’s an ever-growing body of literature arguing “yes”[1]
- There’s diversity in what “yes” means, but paper frequently tell us how LLMs reveal how humans acquire language
- Or enhance or disprove other approaches making claims about our cognitive capacity for language

[1] Pater (2019), Potts (2019), Baroni (2021), Sagae (2021), Wilcox et al. (2022), Warstadt & Bowman (2022), Petersen & Potts (2023), Piantadosi (2023), Portelance & Jasbi (2023), Timkey & Linzen (2023), Ambridge & Blything (2024), Kallini et al. (2024), Lavechin et al. (2024), Shah et al. (2024), Xu et al. (2025)...

Today’s Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

- Putting my cards on the table, I don’t find LLMs to be compelling models of human language cognition, nor do I find them directly informative
- That said, they are some of the best distributional pattern extractors we have, so they can tell us other things about language.

Today's Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

- Putting my cards on the table, I don't find LLMs to be compelling models of human language cognition, nor do I find them directly informative
- That said, they are some of the best distributional pattern extractors we have, so they can tell us other things about language.

Hence the title of the talk:

Use them for what they are

[distributional pattern extractors],

and don't try to make them what they aren't

[models of language cognition]

Outline for Today

“If it quacks like a duck…”

Reassessing the interpretation of behavioral probes for hierarchical syntax

Back to the 1980s

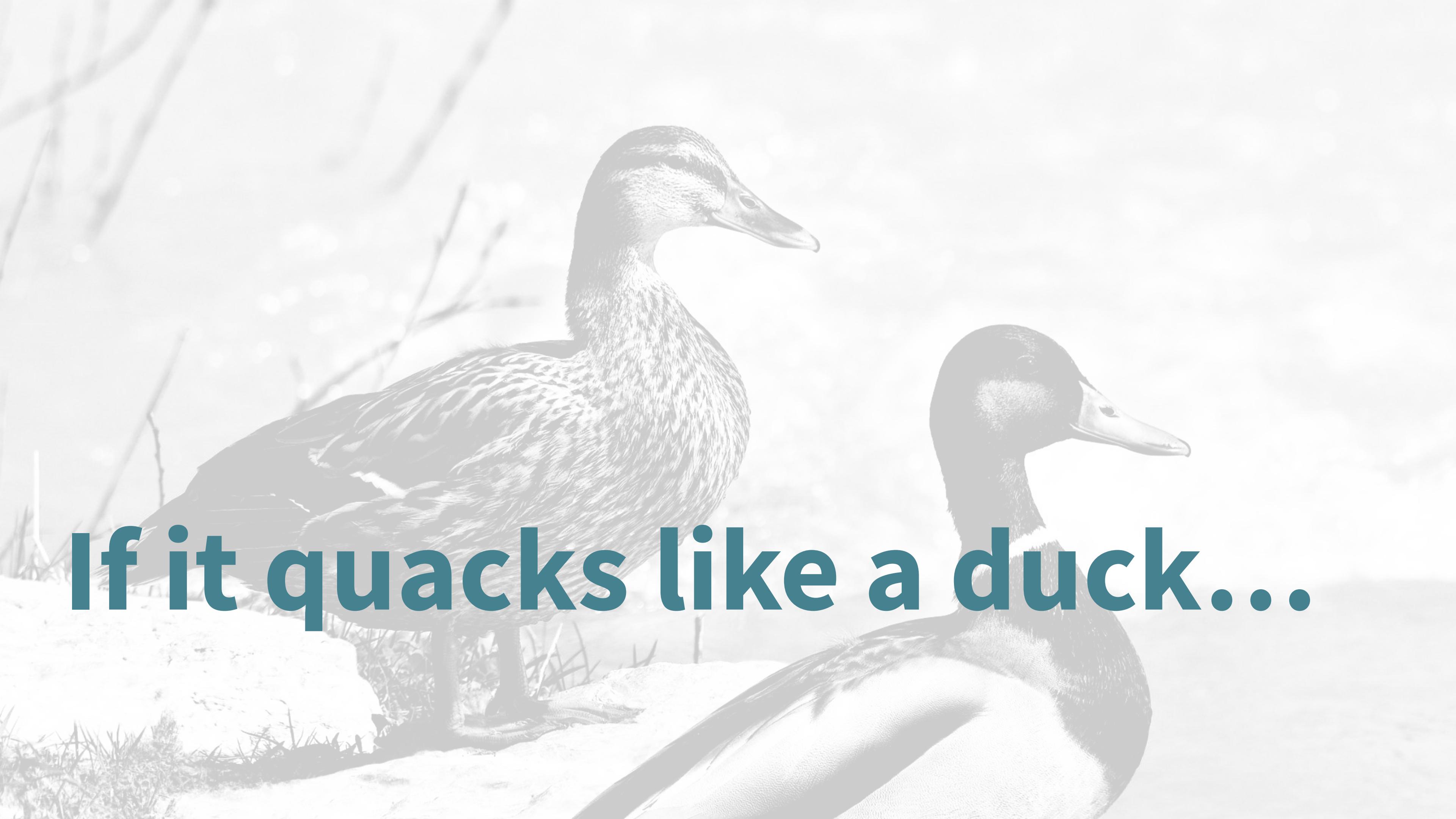
Innate characteristics of neural network models of morphological inflection

Red Herring

Looking beyond behavior to more fundamental issues

Flying High

LLMs are like airplanes to us birds, and that isn’t always a bad thing



If it quacks like a duck…

Behavioral Probes

A leading methodology in interpreting LLMs for language tasks

- **If a model behaves like a human on linguistic tasks,
  it tells us something about the model → it "knows" what it takes to succeed**

Behavioral Probes

A leading methodology in interpreting LLMs for language tasks

- **If a model behaves like a human on linguistic tasks,
it tells us something about the model → it “knows” what it takes to succeed**
- **This often comes with a corollary conclusion with implications for cog sci[1]
If a model behaves like a human on linguistic tasks,
it tells us something about the human → humans solve the task like the model**

[1] e.g., Huebner et al. (2021), Warstadt & Bowman (2022), Evanson et al. (2023), Millière (2024),
Negative-leaning results: Zhang et al (2023), Constantinescu et al. (2025)

Behavioral Probes

A leading methodology in interpreting LLMs for language tasks

- If a model behaves like a human on linguistic tasks,
  it tells us something about the model → it "knows" what it takes to succeed
- This often comes with a corollary conclusion with implications for cog sci[1]
  If a model behaves like a human on linguistic tasks,
  it tells us something about the human → humans solve the task like the model

"If it looks like a duck, swims like a duck, and quacks like a duck, then it's a duck. 🦆🦆"

[1] e.g., Huebner et al. (2021), Warstadt & Bowman (2022), Evanson et al. (2023), Millière (2024),
Negative-leaning results: Zhang et al (2023), Constantinescu et al. (2025)

Behavioral Probes for Hierarchical Syntax

Significant amount of work in this area over the last several years

Linzen et al. (2016), Chowdhury & Zamparelli (2018), Marvin & Linzen (2018), McCoy et al. (2018), Gulordava et al. (2018), Guruprasad et al. (2019), McCoy et al. (2020), Zhang et al. (2020), Papadimitriou et al. (2021), Huebner et al. (2021), Warstadt & Bowman (2022), Wilcox et al (2023), Yedetore et al. (2023), Ahuja et al. (2024)... and many more!

Behavioral Probes for Hierarchical Syntax

Significant amount of work in this area over the last several years

Linzen et al. (2016), Chowdhury & Zamparelli (2018), Marvin & Linzen (2018), McCoy et al. (2018), Gulordava et al. (2018), Guruprasad et al. (2019), McCoy et al. (2020), Zhang et al. (2020), Papadimitriou et al. (2021), Huebner et al. (2021), Warstadt & Bowman (2022), Wilcox et al (2023), Yedetore et al. (2023), Ahuja et al. (2024)... and many more!

Pre-built benchmarking data sets are increasingly popular

CoLA (2019)[1]
BLiMP (2020)[2]
SyntaxGym (2020)[3]
Zorro (2021)[4]
CausalGym (2024)[5]
...

[1] Warstadt et al. (2019), [2] Warstadt & Bowman (2020), [3] Hu et al. (2020), [4] Huebner et al. (2021), [5] Arora et al. (2024)

The Benchmark of Linguistic Minimal Pairs

Benchmark of Linguistic Minimal Pairs (BLiMP; Warstadt et al., 2020)

- By far the most cited grammaticality benchmark
- Also included in shared tasks with cognitive implications as an explicit goal[1]

[1] https://babylm.github.io/

The Benchmark of Linguistic Minimal Pairs

Benchmark of Linguistic Minimal Pairs (BLiMP; Warstadt et al., 2020)

- By far the most cited grammaticality benchmark
- Also included in shared tasks with cognitive implications as an explicit goal[1]
- Pairs of grammatical/ungrammatical sentences covering 12 phenomena
- Thousands of pairs generated automatically from templates
- LLM *M* succeeds on a sentence pair $(s_{\text{gram}}, s_{\text{ungram}})$ iff $P_M(s_{\text{gram}}) > P_M(s_{\text{ungram}})$

Sample Sentence Pair from BLiMP's `adjunct_island` Phenomenon

Grammatical: `Who should Derek hug after shocking Richard?`

Ungrammatical: `Who should Derek hug Richard after shocking?`

[1] https://babylm.github.io/

What do the Probes Actually Test?

Many probes do not actually test the intended structural patterns[1]

BLiMP Phenomenon: SUBJECT-VERB AGREEMENT

Long distance subject verb agreement is the classic example[2]

Grammatical:

Most glasses scare Martin.   Some patients who dislike Kendra negotiate.

Ungrammatical:

Most glasses scares Martin.   Some patients who dislike Kendra negotiates.

Number-mismatched distractor noun

[1] Vázquez Martínez et al. (2023) [2] e.g., Linzen et al. (2016), Marvin & Linzen (2018), Lepori et al. (2020), Hu et al. (2020), Ahuja et al. (2024)

What do the Probes Actually Test?

Many probes do not actually test the intended structural patterns[1]

BLiMP Phenomenon: SUBJECT-VERB AGREEMENT

Long distance subject verb agreement is the classic example[2]

Grammatical:

`Most glasses scare Martin.    Some patients who dislike Kendra negotiate.`

Ungrammatical:

`Most glasses scares Martin.   Some patients who dislike Kendra negotiates.`

- **The target noun is always the first noun**

→ **A linear rule like "the last verb agrees with the first noun" works perfectly!**

[1] Vázquez Martínez et al. (2023) [2] e.g., Linzen et al. (2016), Marvin & Linzen (2018), Lepori et al. (2020), Hu et al. (2020), Ahuja et al. (2024)

What do the Probes Actually Test?

Many probes do not actually test the intended structural patterns[1]

BLiMP Phenomenon: SUBJECT-VERB AGREEMENT

Long distance subject verb agreement is the classic example[2]

Grammatical:

```
    Most glasses scare Martin.   Some patients who dislike Kendra negotiate.
```

Ungrammatical:

```
    Most glasses scares Martin.  Some patients who dislike Kendra negotiates.
```

- **The target noun is always the first noun**

→ **A linear rule like "the last verb agrees with the first noun" works perfectly!**

- **The noun and verb are adjacent in ⅔ of test sentences!**

[1] Vázquez Martínez et al. (2023) [2] e.g., Linzen et al. (2016), Marvin & Linzen (2018), Lepori et al. (2020), Hu et al. (2020), Ahuja et al. (2024)

What do the Probes Actually Test?

Many probes do not actually test the intended structural patterns[1]

BLiMP Phenomenon: ANAPHORA AGREEMENT

Grammatical:

Sherry can't forget herself.	Every story would disagree with itself.

Ungrammatical:

Sherry can't forget himself.	Every story would disagree with himself.

- **These require that the final word agrees in number/gender with the first noun**

→ **Noun and anaphor can be identified with a linear (i.e., non-hierarchical) rule**

[1] Vázquez Martínez et al. (2023)

What do the Probes Actually Test?

Many probes do not actually test the intended structural patterns[2]

BLiMP Phenomenon: ANAPHORA AGREEMENT

Grammatical:

`Sherry can't forget herself.`   `Every story would disagree with itself.`

Ungrammatical:

`Sherry can't forget himself.`   `Every story would disagree with himself.`

- **These require that the final word agrees in number/gender with the first noun**

→ **Noun and anaphor can be identified with a linear (i.e., non-hierarchical) rule**

- **This is really an exercise in matching names with their conventional gender**

[1] Vázquez Martínez et al. (2023)

How big of a problem is this?

Simple linear rules achieve 84.35% overall

- Many tests can be solved perfectly by the application of simple linear rules
- A few of these are silly and should be easily ruled out with further probing
- But the linear rules cannot be ruled out on this data set

How big of a problem is this?

Simple linear rules achieve 84.35% overall

- Many tests can be solved perfectly by the application of simple linear rules
- A few of these are silly and should be easily ruled out with further probing
- But the linear rules cannot be ruled out on this data set

A 5-gram model achieves 61.2%

- These do not encode long distance dependencies or hierarchical structure
- Even in the original paper,
  they achieve ≥75% on 23/67 sub-phenomena and surpass GPT-2 on 8/67

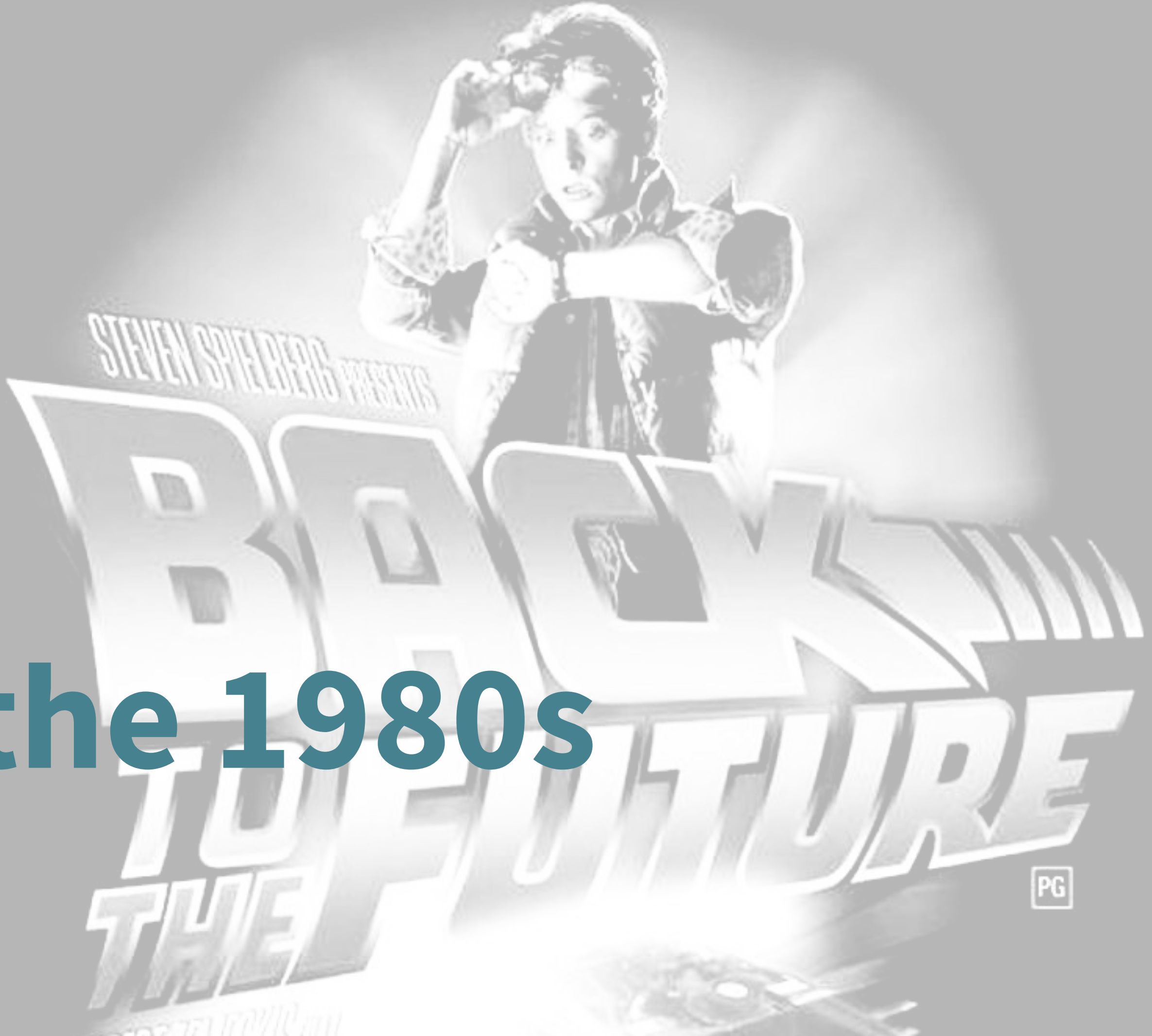
→ About ⅓ of tests are a priori uninformative about syntactic representations

Conclusions

Both the data set and the “duck test” logic are off the mark 🦆

- BLiMP is especially popular, and agreement is still the most popular example, but there’s a plethora of data sets and papers out there
- When a model performs well, it may or may not be for the reasons that the experimenter assumes → A well-known point across the sciences
- See Kodner & Gupta (2020), Lan et al. (2024), Vázquez et al. (*forthcoming*), etc. for (re)evaluations of some other influential papers and data sets

Back to the 1980s



What about Learning?

Do LLMs follow learning trajectories like humans do?

- Behavioral probes, internal probes, etc., focus on representation

  "Do LLMs represent language like humans do? What can that tell us?"
- But the process of learning is also a critical piece of the cognitive puzzle

  "Do LLMs learn language like humans do? What can that tell us?"

The Past Tense Debate in the 1980s and 1990s

There were many points of contention

- How does English past tense work? (hence the name)
- Are irregulars and regulars represented and processed differently?
- How much of linguistic cognition is specific vs. domain-general?
- Can we do without symbolic representations at all?

Increasing Generality

The Past Tense Debate in the 1980s and 1990s

There were many points of contention

- How does English past tense work? (hence the name)
- Are irregulars and regulars represented and processed differently?
- How much of linguistic cognition is specific vs. domain-general?
- Can we do without symbolic representations at all?

Increasing Generality

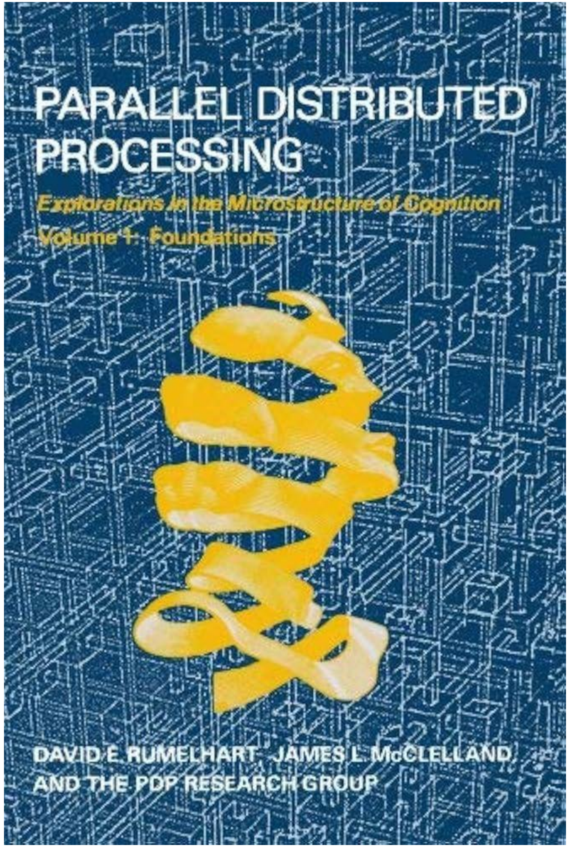
Huge implications for linguistics and cognitive science!

The First Exchange

Rumelhart and McClelland (1986)

- A connectionist approach
- Argued for models of the mind that forgo a middle level of abstraction with rules and symbols
- Adopted distributed representations where rules and symbols do not actually factor into the mental computation → the appearance of rules is emergent

The authors are explicit on both points

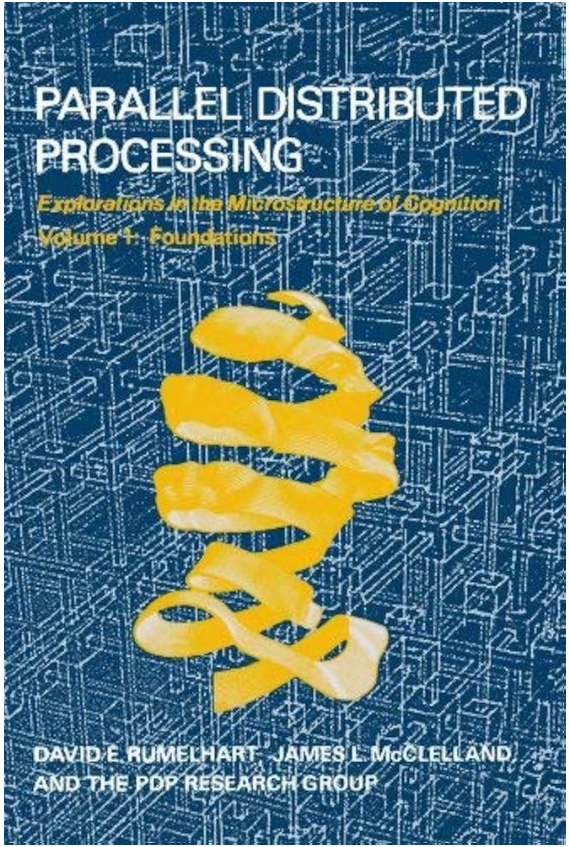

PARALLEL DISTRIBUTED PROCESSING
Explorations in the Microstructure of Cognition
Volume 1: Foundations
DAVID E. RUMELHART, JAMES L. McCLELLAND, AND THE PDP RESEARCH GROUP

The First Exchange

Rumelhart and McClelland (1986)

- A connectionist approach
- Argued for models of the mind that forgo a middle level of abstraction with rules and symbols
- Adopted distributed representations where rules and symbols do not actually factor into the mental computation → the appearance of rules is emergent
- Implemented with artificial neural networks (ANNs)

A clear forerunner to modern discussions about LLMs and cognitive science


PARALLEL DISTRIBUTED PROCESSING
Explorations in the Microstructure of Cognition
Volume 1: Foundations
DAVID E. RUMELHART, JAMES L. McCLELLAND, AND THE PDP RESEARCH GROUP

The First Exchange

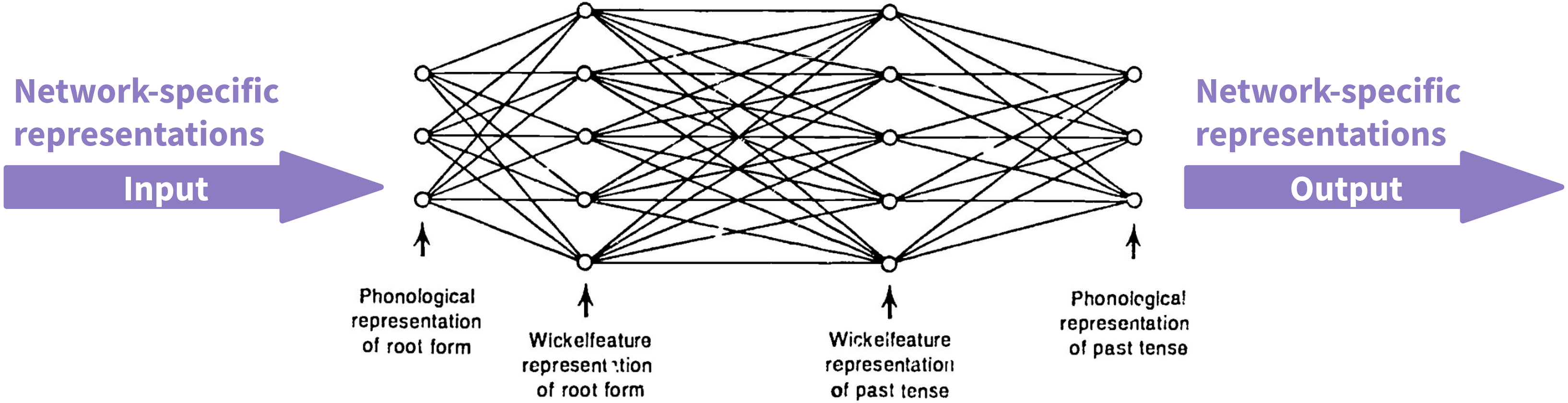
Rumelhart and McClelland (1986)

✔ They actually implemented a neural model of past tense inflection

✔ Crucially, it worked! (at least better than many thought possible in 1986)

✘ Required complicated task-specific input/output representations

✔ But it was among the first to include hidden layers and sigmoid activation



Network-specific representations

Input

Network-specific representations

Output

Phonological representation of root form
Wickelfeature representation of root form
Wickelfeature representation of past tense
Phonological representation of past tense

An Aside: Desiderata for an Inflection Learning Model

A satisfactory computational cognitive model should achieve:

An Aside: Desiderata for an Inflection Learning Model

A satisfactory computational cognitive model should achieve:

1. Learning on a realistic quantity of data with a realistically sized lexicon

Smaller than you'd think!

- Most morphological patterns appear by age 3-4, some much earlier[1]
- When we have a lexicon of several hundred to over 1000 "base words,"[2] only a fraction of which are verbs (or nouns)[3]
- We get around ~10M word tokens per year as input[4]

[1] Brown (1973), Deen (2005), Kim & Sundara (2020), [2] Fenson et al. (1994), [3] Bornstein et al. (2004), [4] Chan (2008), Gilkerson et al. (2017)

An Aside: Desiderata for an Inflection Learning Model

A satisfactory computational cognitive model should achieve:

2. **An asymmetric tendency towards over-regularization vs. over-irregularization**

For English Past Tense,

Over-regularization	**over-application of *-ed*** **(e.g., *feel-*feeled*)**
Over-irregularization	**other patterns** **(e.g., *beep-*bept*, *fry-*frew*)** **cf. *sleep-slept*, *fly-flew***

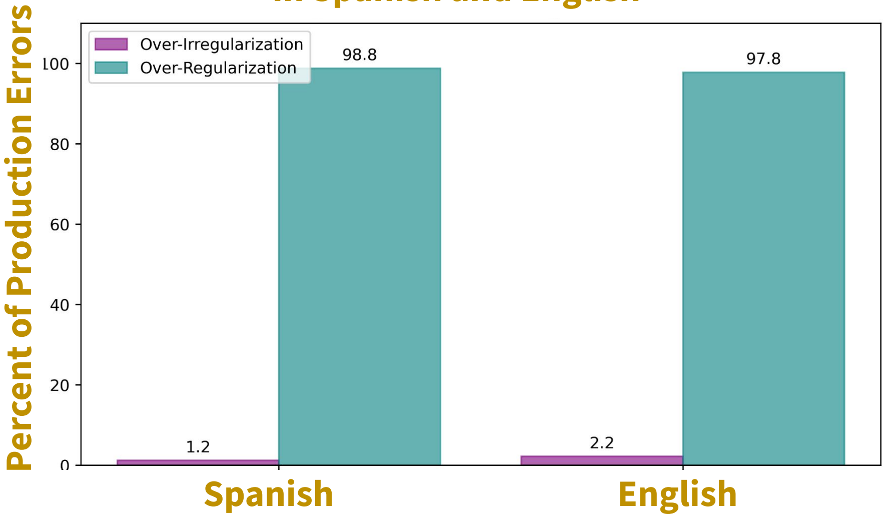
An Aside: Desiderata for an Inflection Learning Model

A satisfactory computational cognitive model should achieve:

2. An asymmetric tendency towards over-regularization vs. over-irregularization

Over-regularization is far more common than over-irregularization

Over-Regularization vs Over-Irregularization in Spanish and English[1]



[1] Clahsen et al. (2002), Maslen et al. (2004)

An Aside: Desiderata for an Inflection Learning Model

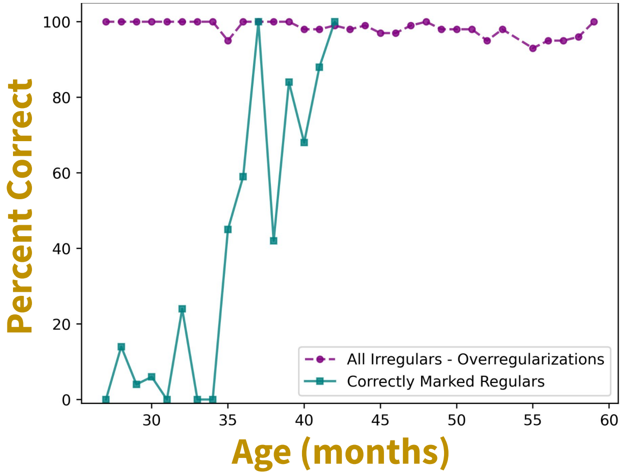
A satisfactory computational cognitive model should achieve:

3. **Developmental regression** **(“*u*-shaped learning”) where appropriate**
4. **An early preference for “base forms” (roots, bare stems; language-specific)**

Observed regression pattern

1. **Performance on irregulars is initially high**
2. **Declines around the time regular forms are consistently marked**
3. **Then later improves**

Adam’s Developmental Regression[1]



[1] Marcus et al. 1992

An Aside: Desiderata for an Inflection Learning Model

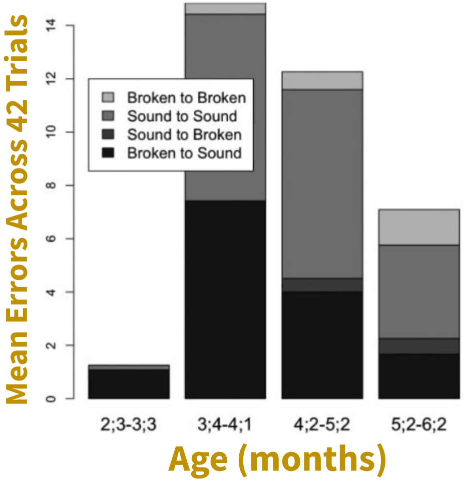
A satisfactory computational cognitive model should achieve:

3. **Developmental regression** (“*u*-shaped learning”) where appropriate
4. An **early preference for “base forms”** (roots, bare stems; language-specific)

Not all morphological patterns show developmental regressions, but many do.

This example from Arabic also shows the over-(ir)regularization asymmetry

Pluralization Errors in Ravid & Farah (1999)[1]



[1] Plot from Dawty-Hesterberg & Pierrehumbert (2015)

The First Exchange

Pinker & Prince (1988)

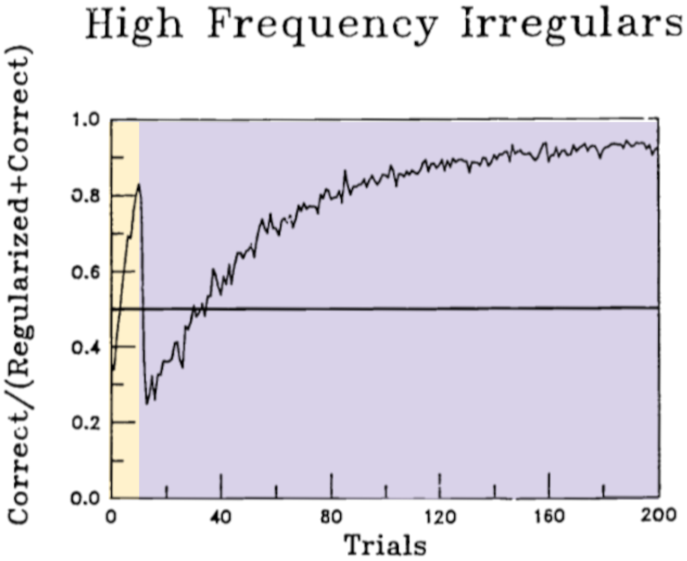
- Heavily criticized all aspects of Rumelhart & McClelland 1986
- Let’s focus on their discussion of the relationship between R&M and empirical observations about morphological development
- And skip arguments related to the particulars of R&M’s architecture

The First Exchange

Pinker & Prince (1988) - Spurious Developmental Regression

- R&M actually do report observing a developmental regression
- But they only achieve it by manipulating the training data in an unusual way

1. Expose the system to 10 high frequency verbs 10 times each (80% irregular)
2. Expose the system to 420 lower frequency verbs 190 times each (44% irregular)



The First Exchange

Pinker & Prince (1988) - Too much Over-Irregularization

- The model produced many over-irregularizations (and related issues)
- Failure to reproduce the strong asymmetry despite its favorable training
- Outputted few base forms (which would be a more plausible failure)

Some R&M over-irregularizations

shape-shipt

sip-sept

Some doubled outputs

type-typeded

step-steppeded

Some gibberish outputs

tour-toureder

mail-membled

The First Exchange

Pinker & Prince (1988) - Too much Over-Irregularization

- The model produced many over-irregularizations (and related issues)
- Failure to reproduce the strong asymmetry despite its favorable training
- Outputted few base forms (which would be a more plausible failure)

Some R&M over-irregularizations

shape-shipt

sip-sept

Gibberish Outputs

tour-toureder

mail-membled

Some doubled outputs

type-typeded

step-steppeded

Foreshadowing…

The Past Tense Debate Revisited

- The Past Tense Debate continued throughout the 1980s and 1990s[1]
- Studies extended past English past tense to German noun plurals, etc.
- The flurry of work saw many advances in neural architectures and training
- But the criticisms from the non-connectionist side were persistent because the shortcomings were persistent

[1] See McClelland & Patterson (2002), Pinker & Ullman (2002), Pinker (2006), and Seidenberg & Plaut (2014) for surveys from both perspectives

The Past Tense Debate Revisited

- The Past Tense Debate continued throughout the 1980s and 1990s[1]
- Studies extended past English past tense to German noun plurals, etc.
- The flurry of work saw many advances in neural architectures and training
- But the criticisms from the non-connectionist side were persistent because the shortcomings were persistent

Cue the deep learning revolution of the mid-2010s…

[1] See McClelland & Patterson (2002), Pinker & Ullman (2002), Pinker (2006), and Seidenberg & Plaut (2014) for surveys from both perspectives

The Past Tense Debate Revisited

NNMIs - Neural Network Models of Inflection

- A range of deep learning approaches to morphology learning: from seq2seq models in the late 2010s,[1] to character-level transducers[2] and transformers[3]
- Report high, often saturated performance on morphological inflection for many languages - see the SIGMORPHON inflection shared tasks[4]

Does high accuracy imply cognitive reality? (the duck test 🦆)

[1] e.g., Kirov & Cotterell (2018), [2] e.g., Clematide et al (2022), [3] e.g., Wu et al. (2021), [4] https://sigmorphon.github.io/sharedtasks/

The Past Tense Debate Revisited

NNMIs - Neural Network Models of Inflection

- A range of deep learning approaches to morphology learning: from seq2seq models in the late 2010s,[1] to character-level transducers[2] and transformers[3]
- Report high, often saturated performance on morphological inflection for many languages - see the SIGMORPHON inflection shared tasks[4]

Does high accuracy imply cognitive reality? (the duck test 🦆)

Some say so. Note the title of Kirov & Cotterell (2018):

“Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate”

[1] e.g., Kirov & Cotterell (2018), [2] e.g., Clematide et al (2022), [3] e.g., Wu et al. (2021), [4] https://sigmorphon.github.io/sharedtasks/

The NLPization of Cognitive Model Evaluation

Modern NNMIs are evaluated almost solely on raw performance

- K&C’s paper is almost entirely focused on practical improvements

  An encoder-decoder model allowing string inputs/outputs of arbitrary length

  Can be applied beyond English past tense to other inflection tasks

  Touted high test accuracy
- Little attempt in the shared tasks to provide interesting error analysis

The NLP quantitative train-test paradigm exists for very good reasons, but it is not appropriate for asking the questions that I am posing today

The NLPization of Cognitive Model Evaluation

Modern NNMIs are evaluated almost solely on raw performance

- K&C’s paper is almost entirely focused on practical improvements
  An encoder-decoder model allowing string inputs/outputs of arbitrary length
  Can be applied beyond English past tense to other inflection tasks
  Touted high test accuracy
- Little attempt in the shared tasks to provide interesting error analysis
- It has fallen on us and other groups to actually ask the cognitive questions
  Corkery et al. (2019), Gorman et al. (2019), McCurdy et al. (2020),
  Belth et al. (2021)[1], Kodner & Khalifa (2022), K et al. (2022, 2023a-c, *under review*), Payne & K (*under review* x2)

[1] Work completed while Belth was at

Morphological Inflection as an NLP Task

Training Time

(lemma, inflected form, feature set) triples

swim	swam	V;PST
eat	eats	V;PRS;3;SG
cat	cats	N;PL
…	…	…

Testing Time

(lemma, feature set) pairs → predict the inflected forms

swim	?	V;PRS;3;SG
box	?	N;PL
cat	?	N;SG
…	…	…

Morphological Inflection as an NLP Task

Training Time (lemma, inflected form, feature set) triples

swim	swam	V;PST
eat	eats	V;PRS;3;SG
cat	cats	N;PL
…	…	…

Testing Time (lemma, feature set) pairs → predict the inflected forms

swim	?	V;PRS;3;SG	→	swims
box	?	N;PL	→	boxes
cat	?	N;SG	→	cat
…	…	…		…

Building More Insightful Evaluations

Interpolation is Easy; Extrapolation is Hard

- **We performed several studies to test different kinds of morphological generalization (Kodner et al. 2022, *et seq*)**

Building More Insightful Evaluations

Interpolation is Easy; Extrapolation is Hard

- We performed several studies to test different kinds of morphological generalization (Kodner et al. 2022, *et seq*)
- Naïve sampling strategies introduce a bias towards easier test sets
- Correcting this:

→ It is challenging generalize across pieces of a paradigm

→ Performance is much lower under smaller training set sizes (hundreds)

Modern NNMI Errors in an Acquisition Setting

Kodner, Payne, Khalifa, & Liu (2023, CogSci and *under review*)

- **Three languages well-studied developmentally**
  Arabic (noun plurals), German (noun plurals), English (verb past tense)

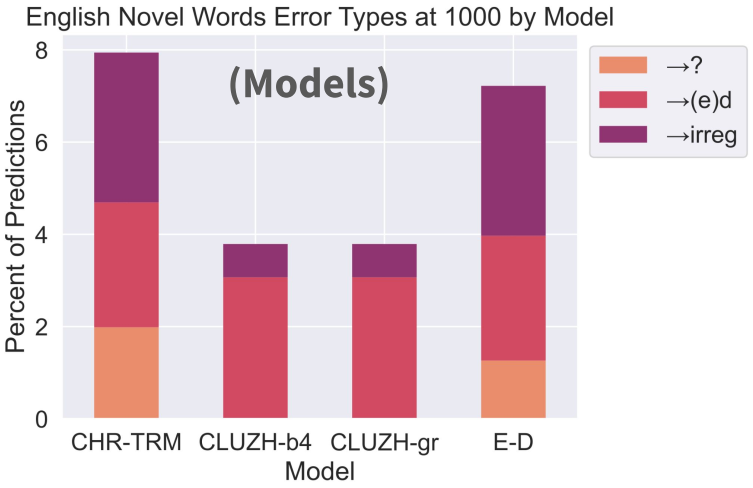
- **Trained with a sequences of nested training sets 100, 200…1000**
  To simulate incremental learning in a batch setting

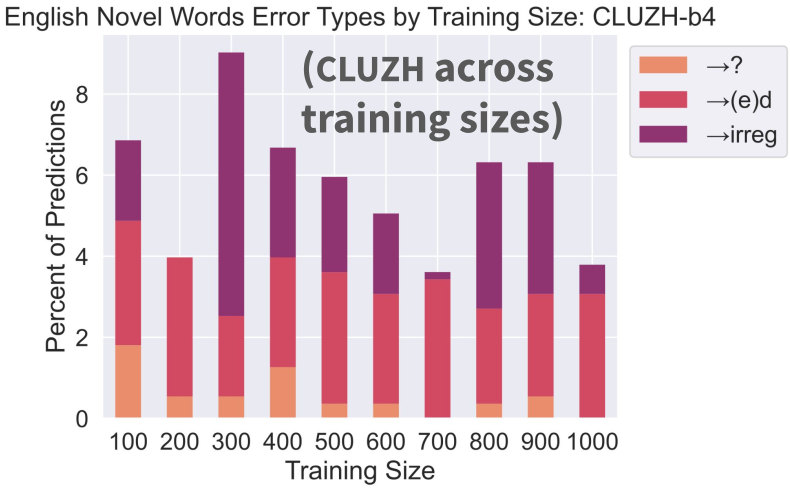
- **We looked at three NNMIs**
  A character transducer CLUZH (Clematide et al., 2022)
  A character transformer CHR-TRM (Wu et al., 2021)
  An LSTM encoder-decoder E-D (Kirov & Cotterell, 2018)

Evaluating English Over-(Ir)Regularization

Performance on the test set (novel words)

- Manually annotated predictions and evaluated against gold data
- All systems over-irregularize proportionately far more than child learners
- No system shows a *u*-shaped learning pattern

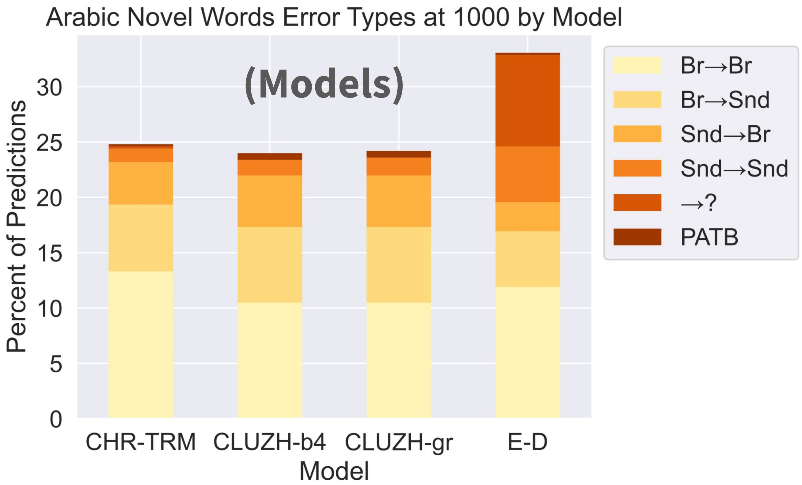


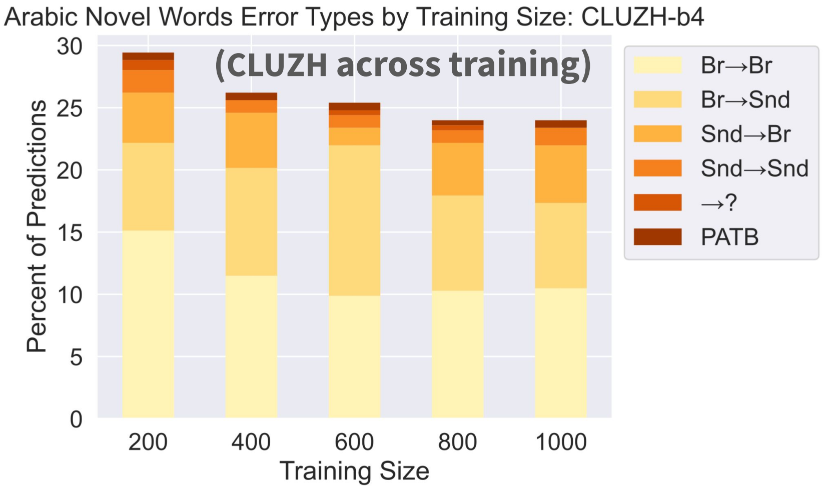


Evaluating Arabic Over-(Ir)Regularization

Performance on the test set (novel words)

- **Manually annotated predictions and evaluated against gold data**
- **All systems over-irregularize proportionately far more than child learners**
- **No system shows a *u*-shaped learning pattern**

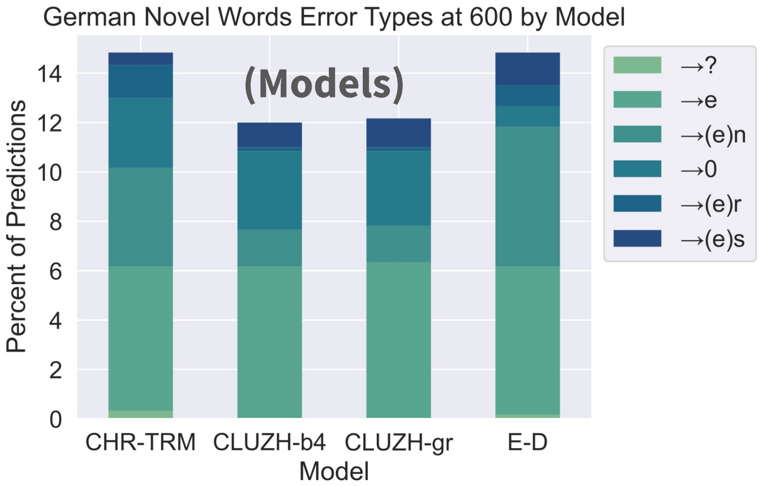


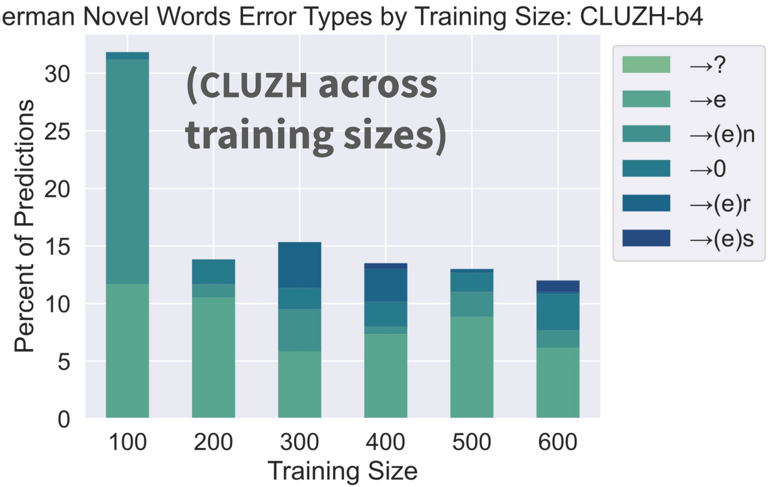


Evaluating German Over-(Ir)Regularization

Performance on the test set (novel words)

- **Manually annotated predictions and evaluated against gold data**
- **Systems excessively favor -*e* plurals and zero plurals**





Evaluating English Over-(Ir)Regularization

Performance on the training data (known words)

- All models achieve superhuman performance

  Perfect performance on nearly all seeds, even at training size 100

  True for German as well. Arabic was a little more challenging

Evaluating English Over-(Ir)Regularization

Performance on the training data (known words)

- All models achieve superhuman performance
  Perfect performance on nearly all seeds, even at training size 100
  True for German as well. Arabic was a little more challenging

- Exception: CHR-TRM failed on seed 0 training size 100 (nonsense) and made three errors on seed 2 training size 700

Over-Regularization	Over-Irregularization
*fall-*falled*	*treat-*trot*
*went-*gooed*	

What do errors actually look like?

They’re actually pretty fun to look at!

- **The CHR-TRM character transformer is prone to producing interesting errors.**
- **We found lots of irregularization,**

whisk-*whought

ping-*pong

peak-*pook

snow-*snew

What do errors actually look like?

They’re actually pretty fun to look at!

- The CHR-TRM character transformer is prone to producing interesting errors.
- We found lots of irregularization, metathesis,

*whisk-*whought*
*ping-*pong*
*peak-*pook*
*snow-*snew*

*bark-*braked*
*clink-*clikned*
*own-*won*
*sharpen-*shaprened*

What do errors actually look like?

They’re actually pretty fun to look at!

- The CHR-TRM character transformer is prone to producing interesting errors.
- We found lots of irregularization, metathesis, and R&M-style gibberish

*whisk-*whought*
*ping-*pong*
*peak-*pook*
*snow-*snew*

*bark-*braked*
*clink-*clikned*
*own-*won*
*sharpen-*shaprened*

*bleed-*bleded*
*go-*toyed*
*materialize-*materioolzed*
*sink-*snurk*

What do errors actually look like?

They’re actually pretty fun to look at!

- The CHR-TRM character transformer is prone to producing interesting errors.
- We found lots of irregularization, metathesis, and R&M-style gibberish

*whisk-*whought*
*ping-*pong*
*peak-*pook*
*snow-*snew*

*bark-*braked*
ned

*bleed-*bleded*
*go-*toyed*
*materialize-*materioolzed*
*sink-*snurk*

Recall, P&P specifically cite *-eded* errors in their 1988 criticism of R&M

Conclusions

Some Inherent Characteristics of NNMIs

- The connectionists of the Past Tense Debate often rightly argued that their shortcomings were plausibly due to their early-stage finicky architectures.
- Modern NNMIs achieve impressively high performance on standard tasks, yet they still show the same empirical deficits as models from the ‘80s and ‘90s

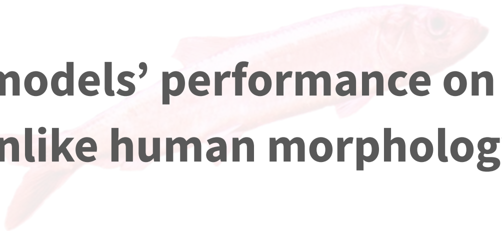
The consistent failure of neural models of morphology inflection to match these developmental behaviors suggests something inherent about these models as a class of learner.

Red Herring

A Quick Recap

What we’ve seen so far

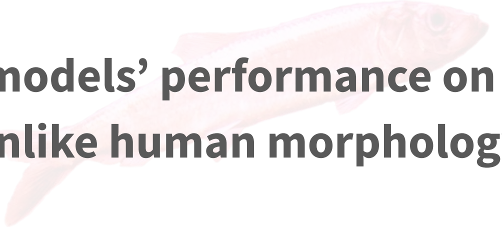
- **Weak benchmarking inflates models’ performance on syntax tasks**
- **NNMIs consistently perform unlike human morphology learners**



A Quick Recap

What we’ve seen so far

- Weak benchmarking inflates models’ performance on syntax tasks
- NNMIs consistently perform unlike human morphology learners

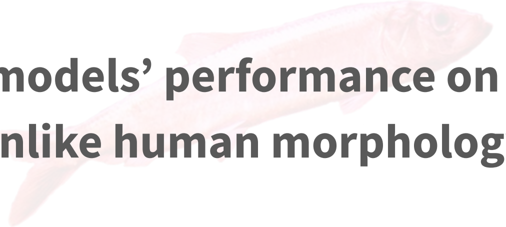


So is this why I reject LLMs (in the broad sense) as models of linguistic cognition?

A Quick Recap

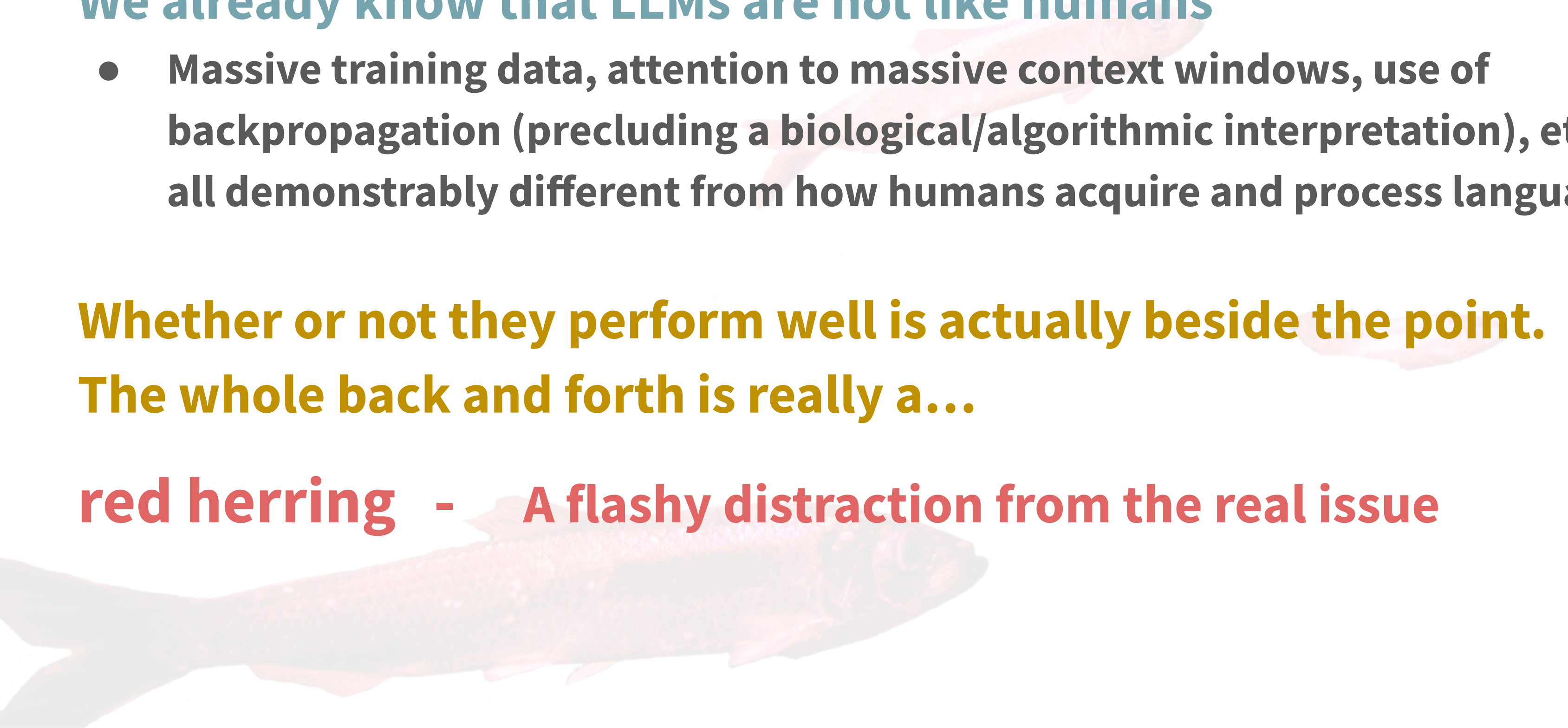
What we’ve seen so far

- Weak benchmarking inflates models’ performance on syntax tasks
- NNMIs consistently perform unlike human morphology learners



So is this why I reject LLMs (in the broad sense) as models of linguistic cognition? Actually, no!

Red Herring



We already know that LLMs are not like humans

- Massive training data, attention to massive context windows, use of backpropagation (precluding a biological/algorithmic interpretation), etc., are all demonstrably different from how humans acquire and process language

Whether or not they perform well is actually beside the point.
The whole back and forth is really a…

red herring - A flashy distraction from the real issue

Affirming the Consequent

As framed for cognitive science and ANNs (Guest & Martin, 2023)

Fallacious reasoning

If an ANN predicts [cognitive measure],
Then it is [cognitive system]

Few state this quite so explicitly in writing.
It’s usually more subtle or dressed in caveats.

Affirming the Consequent

As framed for cognitive science and ANNs (Guest & Martin, 2023)

Fallacious reasoning

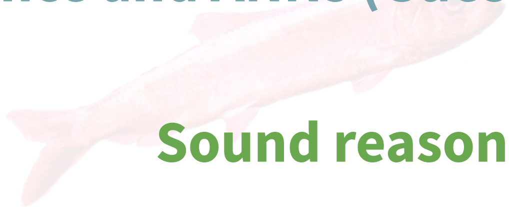
If an ANN predicts [cognitive measure],
Then it is [cognitive system]

Sound reasoning

If an ANN is [cognitive system]
Then it predicts [cognitive measure]

Affirming the Consequent

As framed for cognitive science and ANNs (Guest & Martin, 2023)



Fallacious reasoning

If an ANN predicts [cognitive measure],
Then it is [cognitive system]

Sound reasoning

If an ANN is [cognitive system]
Then it predicts [cognitive measure]

And it’s contrapositive:

If an ANN does not predict [cognitive measure]
Then it is not [cognitive system]

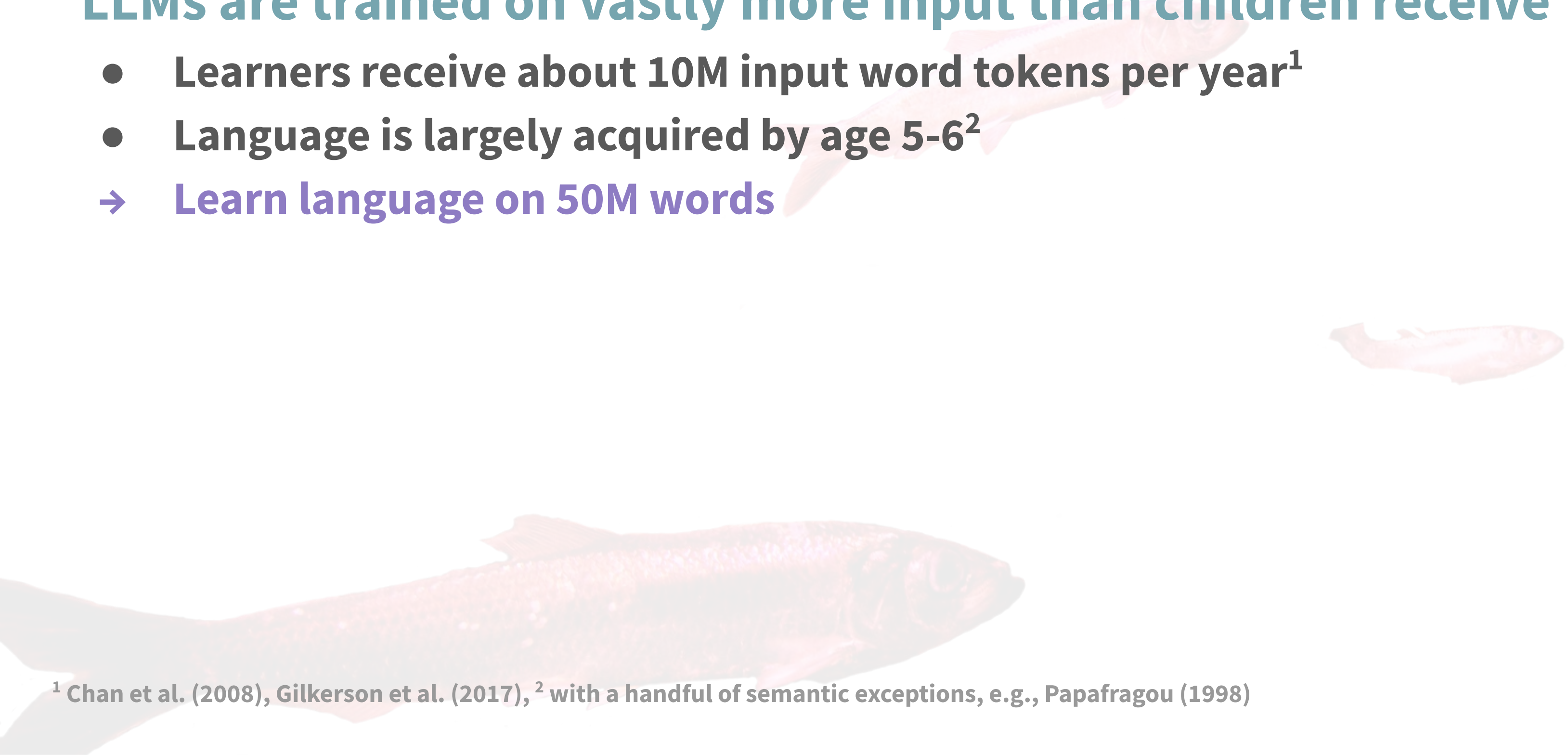
I used this in my criticism NNMIs

Training Size

LLMs are trained on vastly more input than children receive

- Learners receive about 10M input word tokens per year[1]
- Language is largely acquired by age 5-6[2]

→ Learn language on 50M words



[1] Chan et al. (2008), Gilkerson et al. (2017), [2] with a handful of semantic exceptions, e.g., Papafragou (1998)

Training Size

LLMs are trained on vastly more input than children receive

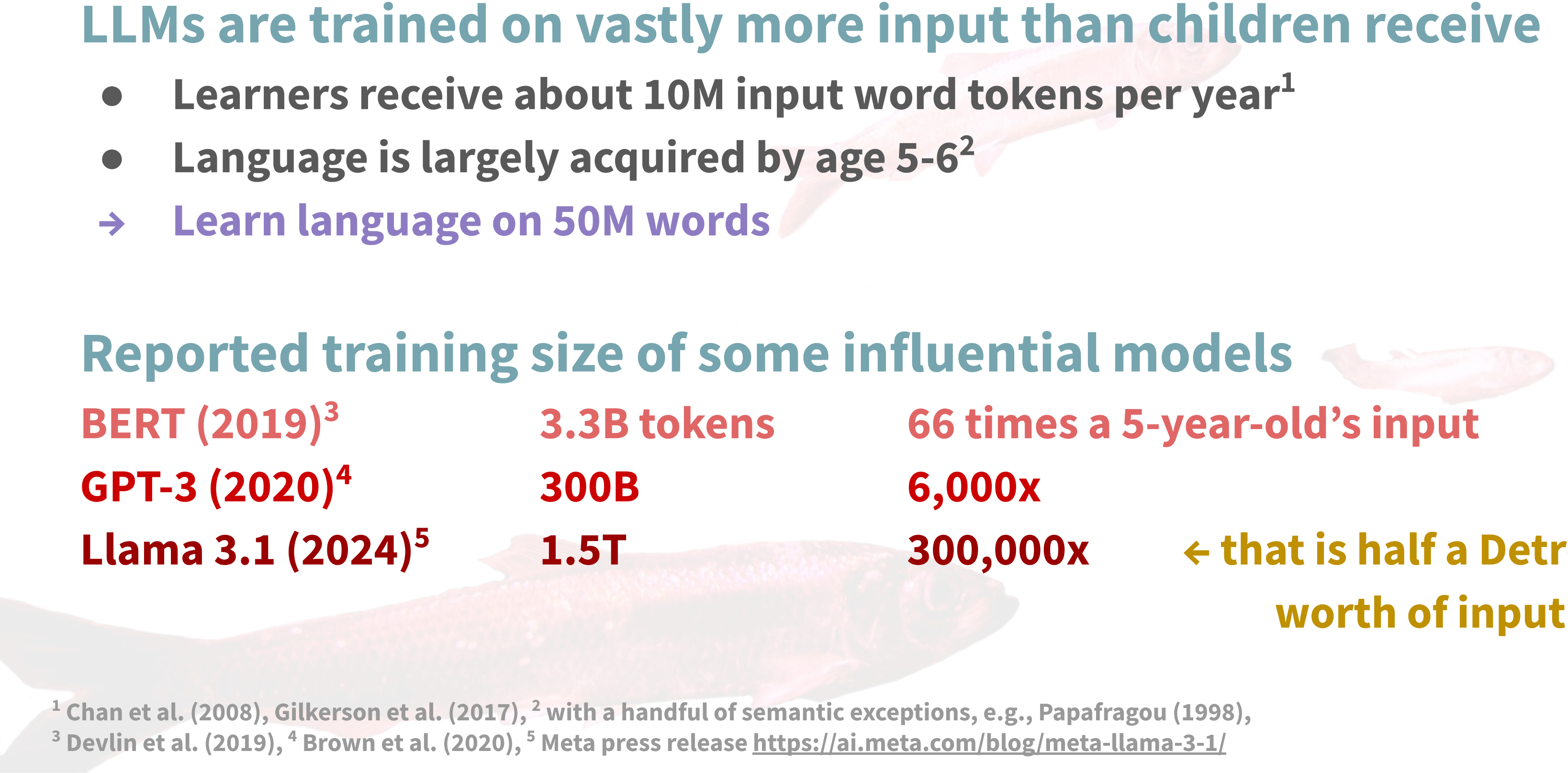
- Learners receive about 10M input word tokens per year[1]
- Language is largely acquired by age 5-6[2]

→ Learn language on 50M words

Reported training size of some influential models

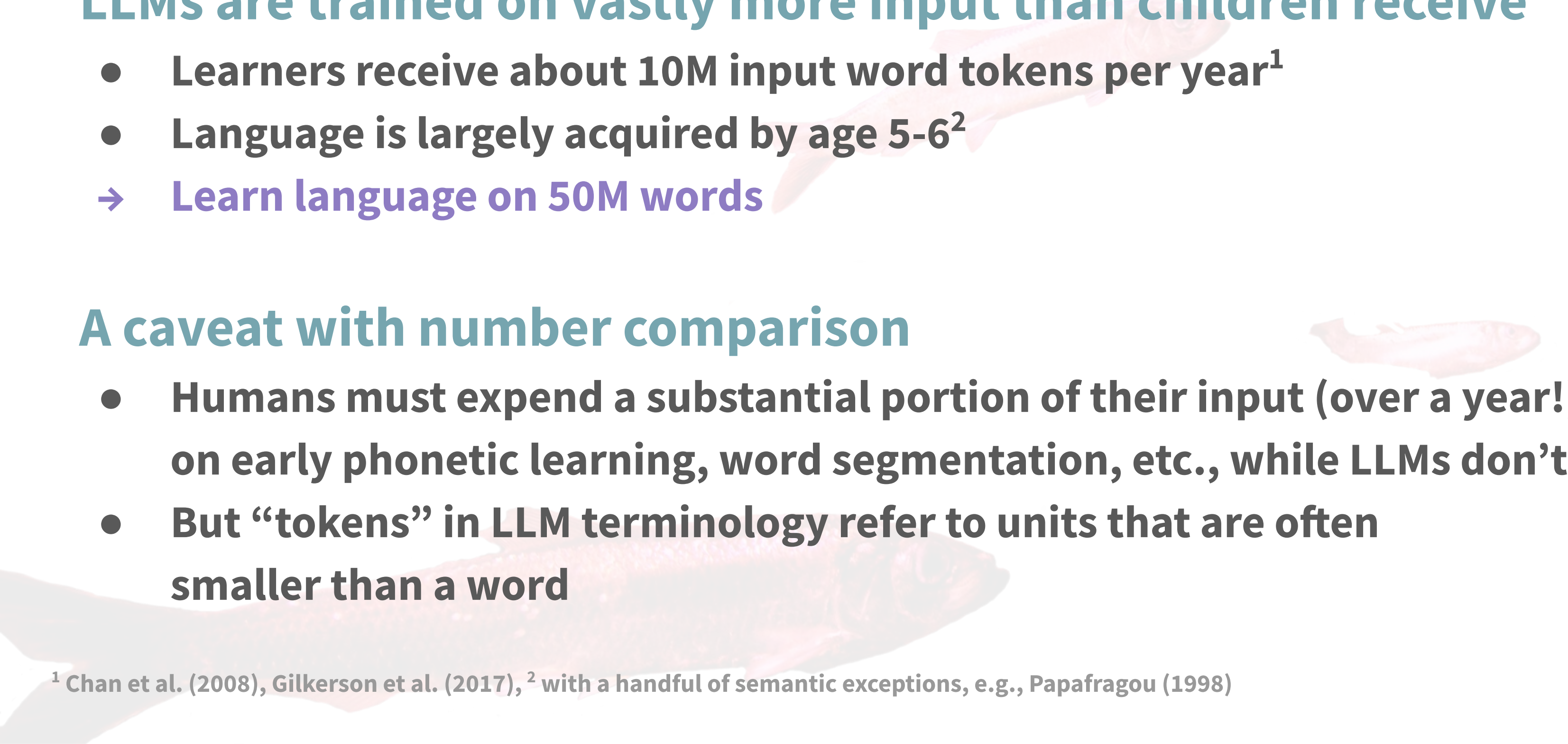
BERT (2019)[3]	3.3B tokens	66 times a 5-year-old's input
GPT-3 (2020)[4]	300B	6,000x
Llama 3.1 (2024)[5]	1.5T	300,000x

← that is half a Detroit's worth of input!



[1] Chan et al. (2008), Gilkerson et al. (2017), [2] with a handful of semantic exceptions, e.g., Papafragou (1998), [3] Devlin et al. (2019), [4] Brown et al. (2020), [5] Meta press release https://ai.meta.com/blog/meta-llama-3-1/

Training Size



LLMs are trained on vastly more input than children receive

- Learners receive about 10M input word tokens per year[1]
- Language is largely acquired by age 5-6[2]

→ Learn language on 50M words

A caveat with number comparison

- Humans must expend a substantial portion of their input (over a year!) on early phonetic learning, word segmentation, etc., while LLMs don’t
- But “tokens” in LLM terminology refer to units that are often smaller than a word

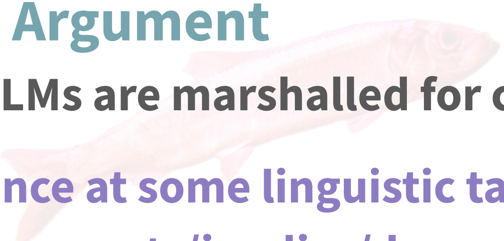
[1] Chan et al. (2008), Gilkerson et al. (2017), [2] with a handful of semantic exceptions, e.g., Papafragou (1998)

Training Size

The Poverty of the Stimulus Argument

- **The most common way that LLMs are marshalled for cognitive claims[1]**

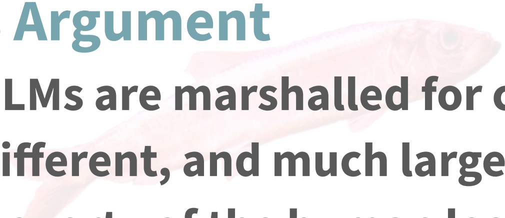
  “If this LLM displays competence at some linguistic task that was previously unseen outside of humans, it suggests/implies/demonstrates that humans could also do so with little or no in-built linguistic biases”



[1] e.g., McCoy et al. (2018), Baroni (2021), Huebner et al. (2021), Warstadt & Bowman (2022), Wilcox et al. (2023), Piantadosi (2023), Portelance & Jasbi (2023)

Training Size

The Poverty of the Stimulus Argument



- The most common way that LLMs are marshalled for cognitive claims[1]
- But if an LLM is trained on a different, and much larger stimulus, what can it tell us about the poverty of the human learner's stimulus?

[1] e.g., McCoy et al. (2018), Baroni (2021), Huebner et al. (2021), Warstadt & Bowman (2022), Wilcox et al. (2023), Piantadosi (2023), Portelance & Jasbi (2023)

Training Size

The Poverty of the Stimulus Argument

- The most common way that LLMs are marshalled for cognitive claims[1]
- But if an LLM is trained on a different, and much larger stimulus, what can it tell us about the poverty of the human learner's stimulus?

In Chris Potts's recent LSA (Linguistic Society of America conference) keynote[2]:

- He argues that an LLM can recognize PIPPs spanning finite-clauses even though they only occur 58 times in the C4 corpus (150B word tokens)

→ "There is no logical requirement for such in-built mechanisms" [to solve the Poverty of the Simulus]

[1] e.g., McCoy et al. (2018), Baroni (2021), Huebner et al. (2021), Warstadt & Bowman (2022), Wilcox et al. (2023), Piantadosi (2023), Portelance & Jasbi (2023), [2] Potts (2025, LSA Keynote https://www.youtube.com/watch?v=DBorepHuKDM)

Training Size

The Poverty of the Stimulus Argument

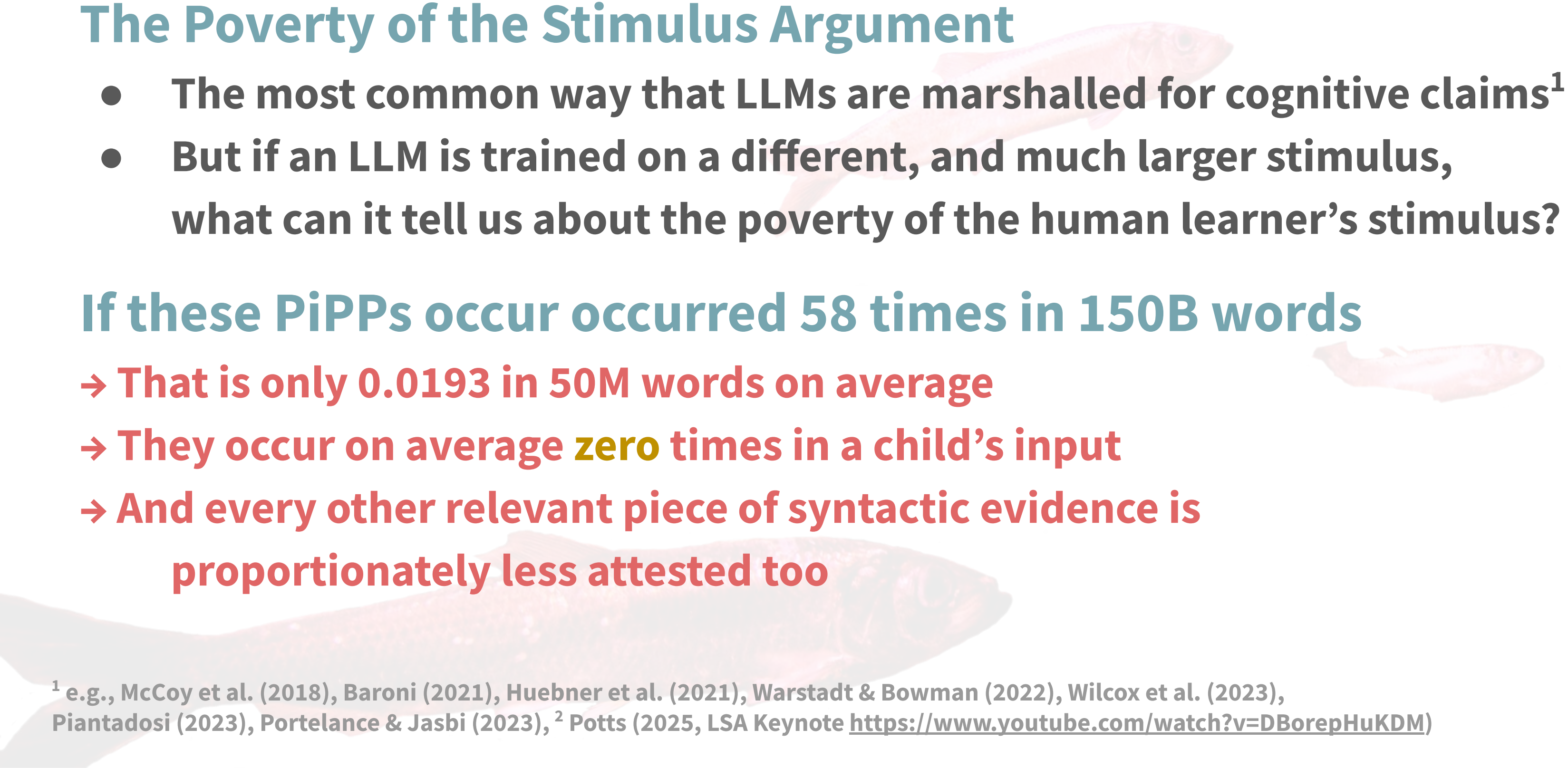
- The most common way that LLMs are marshalled for cognitive claims[1]
- But if an LLM is trained on a different, and much larger stimulus, what can it tell us about the poverty of the human learner’s stimulus?

If these PiPPs occur occurred 58 times in 150B words

→ That is only 0.0193 in 50M words on average

→ They occur on average zero times in a child’s input

→ And every other relevant piece of syntactic evidence is proportionately less attested too

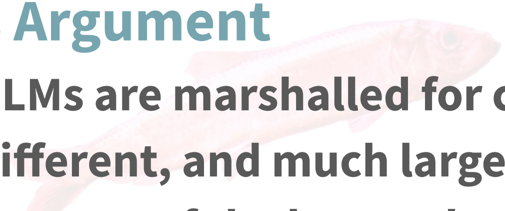


[1] e.g., McCoy et al. (2018), Baroni (2021), Huebner et al. (2021), Warstadt & Bowman (2022), Wilcox et al. (2023), Piantadosi (2023), Portelance & Jasbi (2023), [2] Potts (2025, LSA Keynote https://www.youtube.com/watch?v=DBorepHuKDM)

Training Size

The Poverty of the Stimulus Argument

- The most common way that LLMs are marshalled for cognitive claims[1]
- But if an LLM is trained on a different, and much larger stimulus, what can it tell us about the poverty of the human learner's stimulus?



I care about "The Poverty of THE Stimulus,"

not the Poverty of some other Stimulus

[1] e.g., McCoy et al. (2018), Baroni (2021), Huebner et al. (2021), Warstadt & Bowman (2022), Wilcox et al. (2023), Piantadosi (2023), Portelance & Jasbi (2023), [2] Potts (2025, LSA Keynote https://www.youtube.com/watch?v=DBorepHuKDM)

Context Windows

The number of recent tokens that an LLM can “remember”

- Can contain both the user input and the LLM’s own output
- Rapidly increasing context lengths have been a major contributor to improvements in LLM performance

Context windows for recent LLMs

GPT-3 (2020)[1]	2,000 tokens	
GPT-4 Turbo (2024)[2]	128,000	← *The War of the Worlds* is about 63k words
Gemini 1.5 Pro (2024)[3]	2,000,000	← All seven books of the *Harry Potter* series together are about 1.1 million words

[1] Brown et al. (2020), [2] https://help.openai.com/en/articles/8555510-gpt-4-turbo-in-the-openai-api,
[3] https://developers.googleblog.com/en/new-features-for-the-gemini-api-and-google-ai-studio/

The Optimization Function

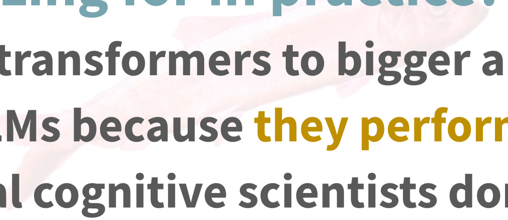
What are researchers optimizing for in practice?

- **NLP has moved from RNNs to transformers to bigger and bigger transformers to very large prompt-based LLMs because they perform better and better!**

The Optimization Function

What are researchers optimizing for in practice?

- NLP has moved from RNNs to transformers to bigger and bigger transformers to very large prompt-based LLMs because they perform better and better!
- Why have many computational cognitive scientists done the same thing? What about a 2020s LLM makes it more desirable than a late 2010s LLM?



The Optimization Function

What are researchers optimizing for in practice?

- NLP has moved from RNNs to transformers to bigger and bigger transformers to very large prompt-based LLMs because they perform better and better!
- Why have many computational cognitive scientists done the same thing? What about a 2020s LLM makes it more desirable than a late 2010s LLM?
- GPT-4 (2023) has much larger context windows than BERT (2019)

→ If GPT-4 outperforms BERT on a language processing task, what do we learn? That humans have really wide context windows?

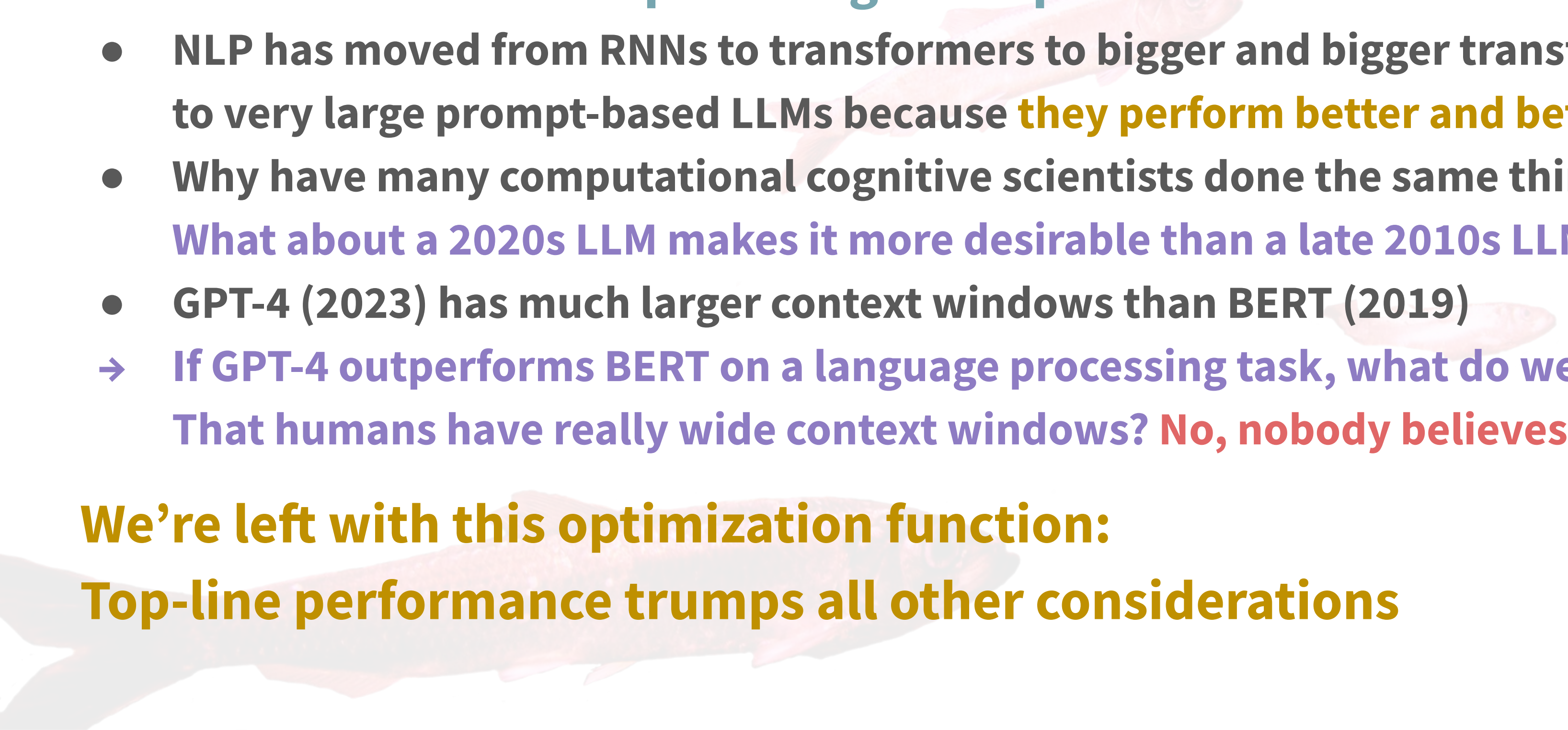
The Optimization Function

What are researchers optimizing for in practice?

- NLP has moved from RNNs to transformers to bigger and bigger transformers to very large prompt-based LLMs because **they perform better and better!**
- Why have many computational cognitive scientists done the same thing? What about a 2020s LLM makes it more desirable than a late 2010s LLM?
- GPT-4 (2023) has much larger context windows than BERT (2019)

→ If GPT-4 outperforms BERT on a language processing task, what do we learn? That humans have really wide context windows? No, nobody believes that.

We're left with this optimization function:
Top-line performance trumps all other considerations



But what about…

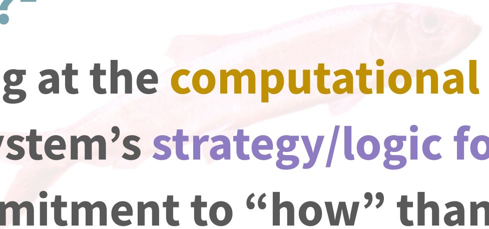
…an appeal to Marr's Levels?[1]

- LLMs are meant to be operating at the **computational level**
- This describes the cognitive system's **strategy/logic for mapping** from **input** to **output**, with less commitment to "how" than the algorithmic level

[1] Marr (1982)

But what about…

…an appeal to Marr's Levels?[1]



- LLMs are meant to be operating at the **computational level**
- This describes the cognitive system's **strategy/logic for mapping** from **input** to **output**, with less commitment to "how" than the algorithmic level

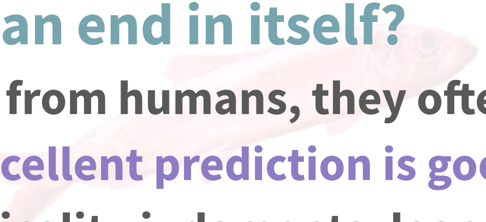
But we already know that the LLMs' input is not the human input, and we've just discussed uncertainties and problems with the output, and the strategy/logic of the mapping in LLMs is notoriously opaque

[1] Marr (1982)

But what about…

…an appeal to prediction as an end in itself?

- Even if LLMs differ irrevocably from humans, they often predict human behavior really well. Maybe excellent prediction is good enough?
- Predicting behavior (grammaticality judgments, learning trajectories, neural(🧠) activity, etc.) is an important part of cognitive science of language. It provides crucial evidence for the mechanisms of the mind



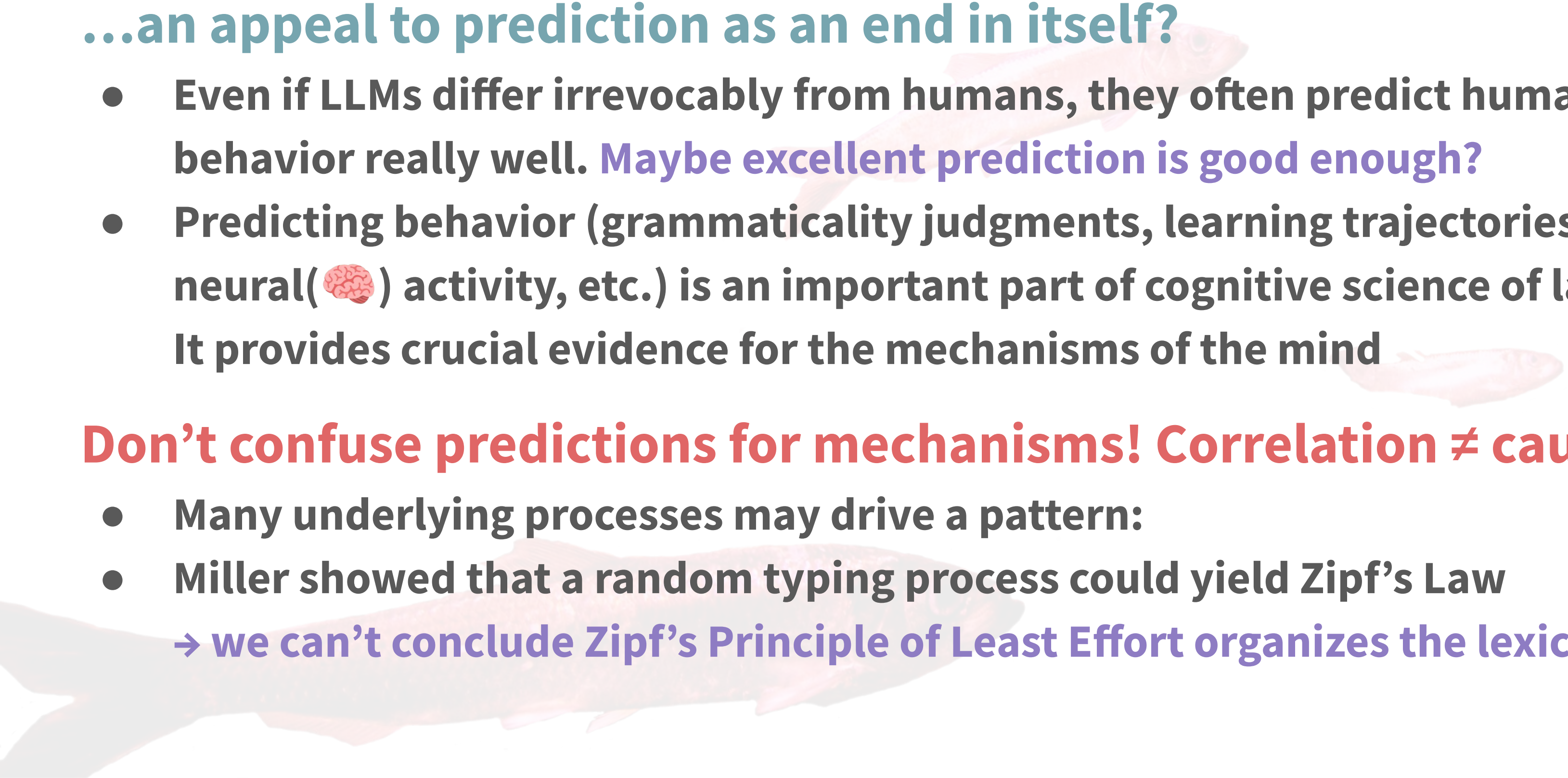
But what about…

…an appeal to prediction as an end in itself?

- Even if LLMs differ irrevocably from humans, they often predict human behavior really well. Maybe excellent prediction is good enough?
- Predicting behavior (grammaticality judgments, learning trajectories, neural(🧠) activity, etc.) is an important part of cognitive science of language. It provides crucial evidence for the mechanisms of the mind

Don’t confuse predictions for mechanisms! Correlation ≠ causation

- Many underlying processes may drive a pattern:
- Miller showed that a random typing process could yield Zipf’s Law
  → we can’t conclude Zipf’s Principle of Least Effort organizes the lexicon

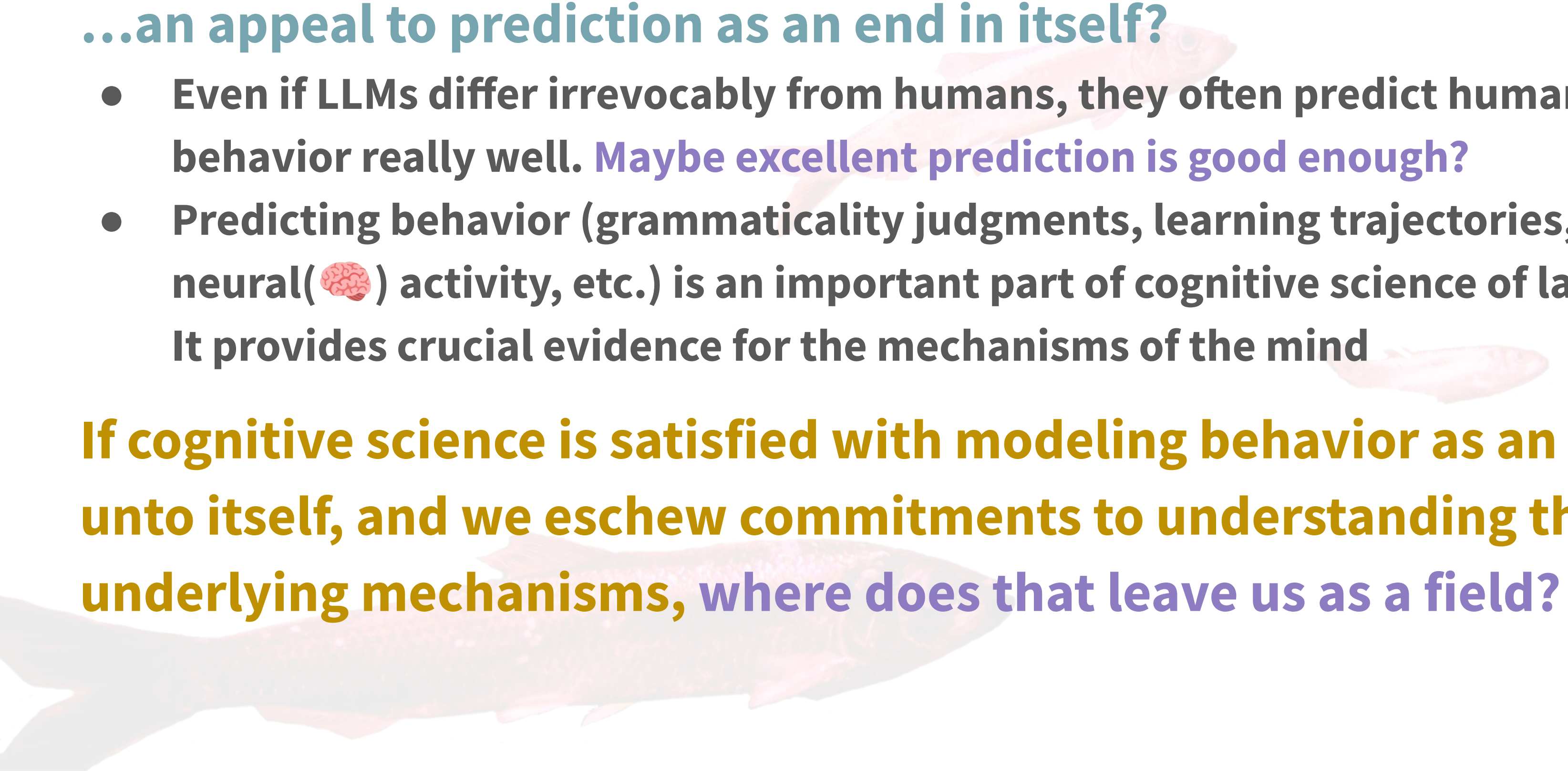


But what about…

…an appeal to prediction as an end in itself?

- Even if LLMs differ irrevocably from humans, they often predict human behavior really well. Maybe excellent prediction is good enough?
- Predicting behavior (grammaticality judgments, learning trajectories, neural(🧠) activity, etc.) is an important part of cognitive science of language. It provides crucial evidence for the mechanisms of the mind

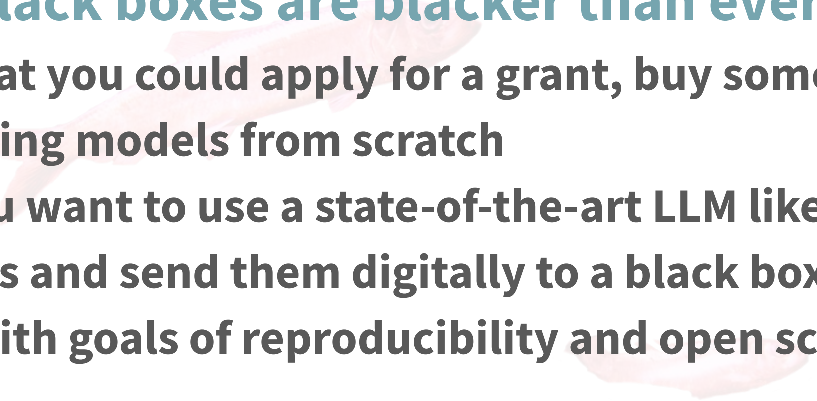
If cognitive science is satisfied with modeling behavior as an end unto itself, and we eschew commitments to understanding the underlying mechanisms, where does that leave us as a field?



The Middlemen

LLMs in the narrow sense - Black boxes are blacker than ever

- It was only a few years ago, that you could apply for a grant, buy some GPUs, and train the latest deep learning models from scratch
- Now they’re proprietary. If you want to use a state-of-the-art LLM like GPT-4, you need to formulate prompts and send them digitally to a black box
- This state of affairs conflicts with goals of reproducibility and open science

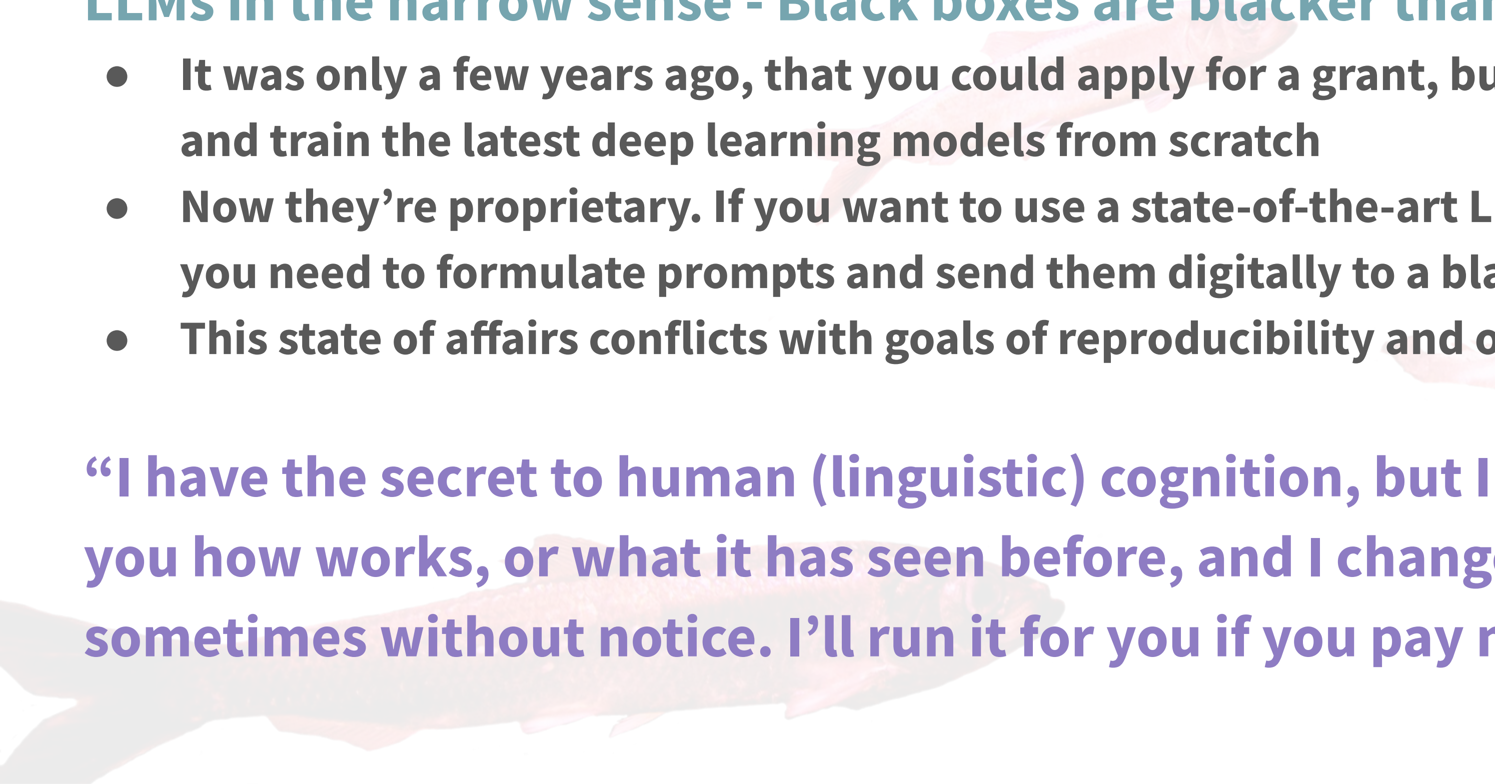


The Middlemen

LLMs in the narrow sense - Black boxes are blacker than ever

- It was only a few years ago, that you could apply for a grant, buy some GPUs, and train the latest deep learning models from scratch
- Now they're proprietary. If you want to use a state-of-the-art LLM like GPT-4, you need to formulate prompts and send them digitally to a black box
- This state of affairs conflicts with goals of reproducibility and open science

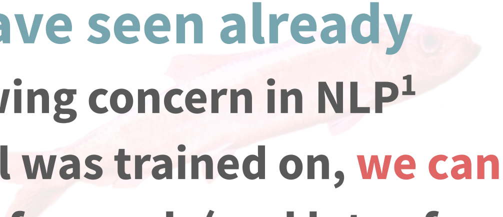
"I have the secret to human (linguistic) cognition, but I won't tell you how works, or what it has seen before, and I change it sometimes without notice. I'll run it for you if you pay me."



The Middlemen

We don’t know what LLMs have seen already

- “Data contamination” - A growing concern in NLP[1]
- If we don’t know what a model was trained on, we can’t avoid test-on-train
- As models suck up most of the free web (and lots of not-free things)[2] this becomes more and more of a problem



[1] Magar & Schwartz (2022), Aiyappa et al. (2023), Sainz et al. (2023), Ballocou et al. (2024),

[2] https://arstechnica.com/tech-policy/2025/02/meta-torrented-over-81-7tb-of-pirated-books-to-train-ai-authors-say/

The Middlemen

What even are the architectures?

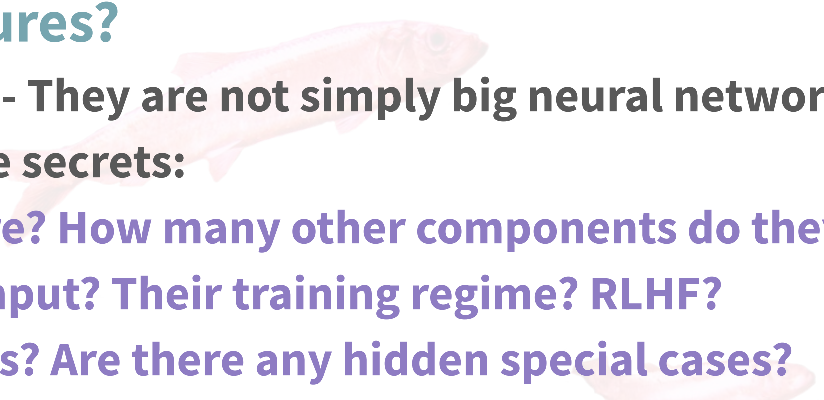
- LLMs are very complex things - They are not simply big neural networks
- Their inner workings are trade secrets:

  What is their basic architecture? How many other components do they have?

  How are they trained? Their input? Their training regime? RLHF?

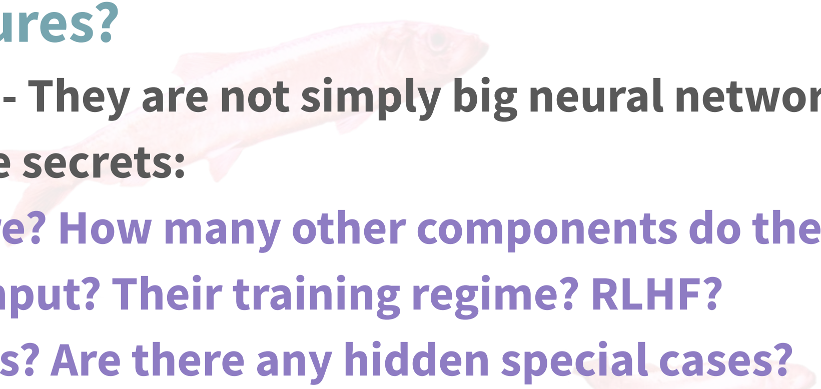
  What are their system prompts? Are there any hidden special cases?

  Is work ever off-loaded to a traditional calculator or info. retrieval system?



The Middlemen

What even are the architectures?



- LLMs are very complex things - They are not simply big neural networks
- Their inner workings are trade secrets:

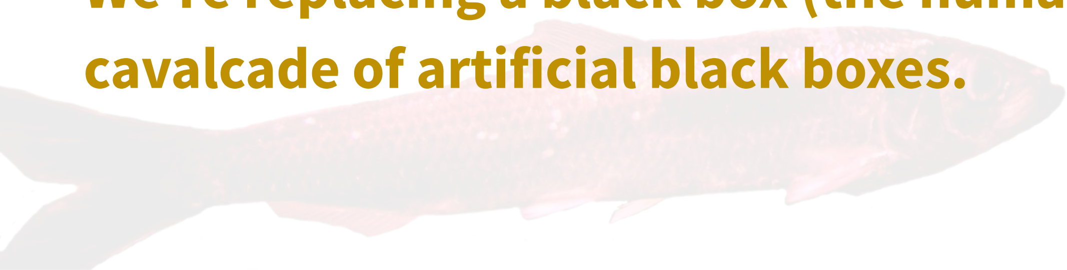
  What is their basic architecture? How many other components do they have?

  How are they trained? Their input? Their training regime? RLHF?

  What are their system prompts? Are there any hidden special cases?

  Is work ever off-loaded to a traditional calculator or info. retrieval system?

We’re replacing a black box (the human mind) with an ever-growing cavalcade of artificial black boxes.



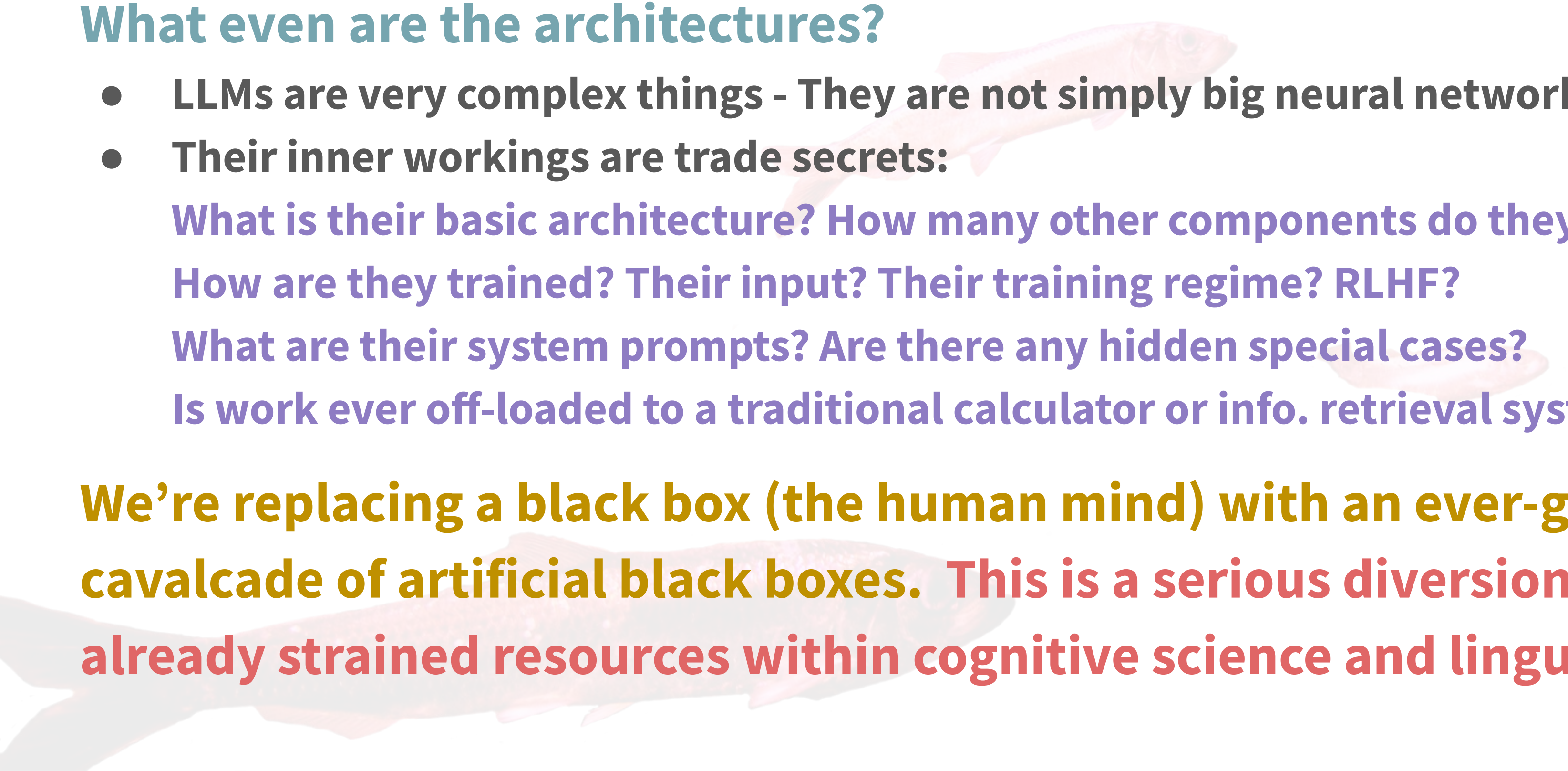
The Middlemen

What even are the architectures?

- LLMs are very complex things - They are not simply big neural networks
- Their inner workings are trade secrets:

  What is their basic architecture? How many other components do they have?
  How are they trained? Their input? Their training regime? RLHF?
  What are their system prompts? Are there any hidden special cases?
  Is work ever off-loaded to a traditional calculator or info. retrieval system?

We’re replacing a black box (the human mind) with an ever-growing cavalcade of artificial black boxes. This is a serious diversion of already strained resources within cognitive science and linguistics.



“Fine, then Use an Open Model”

“The Opening up ChatGPT” leaderboard[1]

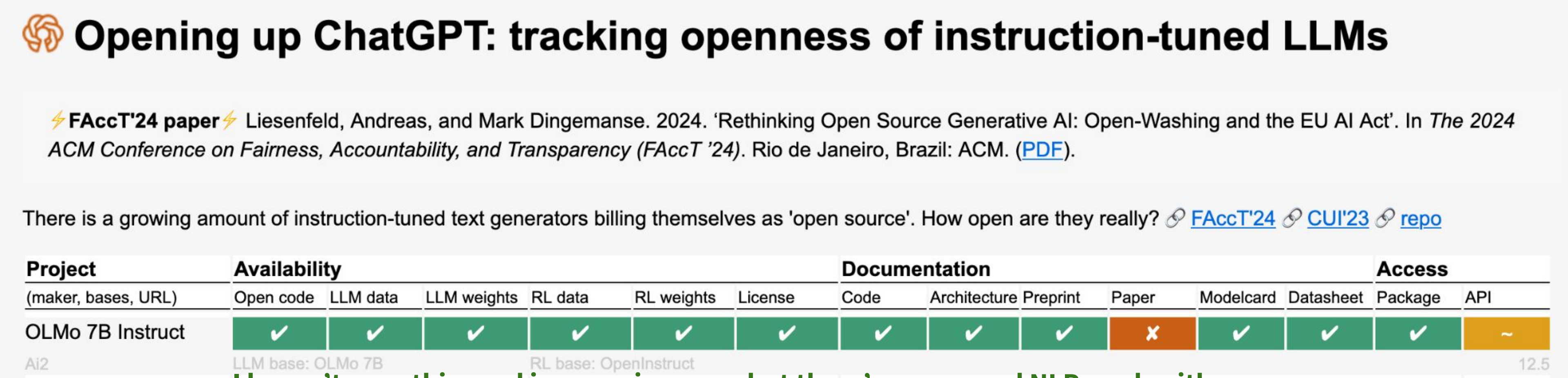
https://opening-up-chatgpt.github.io/

- “Open” is increasingly a marketing buzzword with little meaning

→ Liesenfeld et al. (2023, 2024) call this “open washing”

- Their leaderboard tracks how “open” popular modern LLMs actually are

[1] Liesenfeld et al (2023), Liesenfeld & Dingemanse (2024)



Opening up ChatGPT: tracking openness of instruction-tuned LLMs

⚡**FAccT'24 paper**⚡ Liesenfeld, Andreas, and Mark Dingemanse. 2024. 'Rethinking Open Source Generative AI: Open-Washing and the EU AI Act'. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Rio de Janeiro, Brazil: ACM. (PDF).

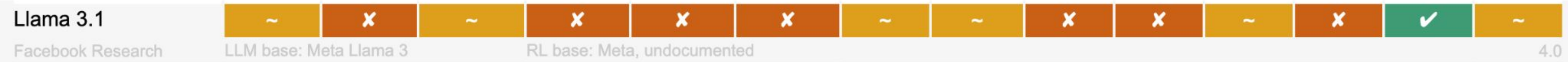
There is a growing amount of instruction-tuned text generators billing themselves as 'open source'. How open are they really? 🔗 FAccT'24 🔗 CUI'23 🔗 repo

Project	**Availability**						**Documentation**						**Access**	
(maker, bases, URL)	Open code	LLM data	LLM weights	RL data	RL weights	License	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
OLMo 7B Instruct	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	~
Ai2	LLM base: OLMo 7B			RL base: OpenInstruct										12.5

I haven't seen this used in cog sci papers, but there's some good NLP work with intervention-based probing methods (e.g., Wiegreffe et al. 2025)

...

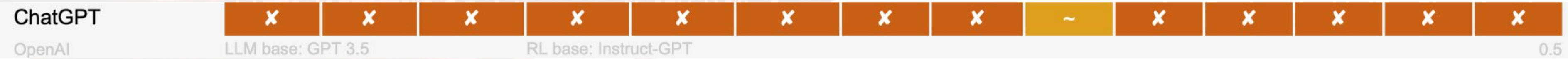
About ⅔ of the way down, we get Meta's Llama, the best known open model...



Llama 3.1	~	✗	~	✗	✗	✗	~	~	✗	✗	~	✗	✓	~
Facebook Research	LLM base: Meta Llama 3			RL base: Meta, undocumented										4.0

...

And OpenAI's ChatGPT (the only GPT on the list as of 2024) comes in dead last...

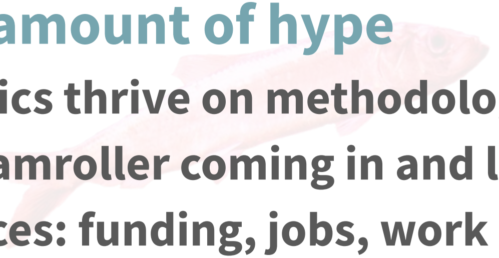


ChatGPT	✗	✗	✗	✗	✗	✗	✗	✗	~	✗	✗	✗	✗	✗
OpenAI	LLM base: GPT 3.5			RL base: Instruct-GPT										0.5

[1] Liesenfeld et al (2023), Liesenfeld & Dingemanse (2024)

A Methodological Steamroller

LLMs are generating a huge amount of hype



- Cognitive science and linguistics thrive on methodological diversity
- But LLMs may represent a steamroller coming in and leveling out this diversity
- A massive diversion of resources: funding, jobs, work hours, publication space

A Methodological Steamroller

LLMs are generating a huge amount of hype

- Cognitive science and linguistics thrive on methodological diversity
- But LLMs may represent a steamroller coming in and leveling out this diversity
- A massive diversion of resources: funding, jobs, work hours, publication space

(sorry about the mismatched axes)

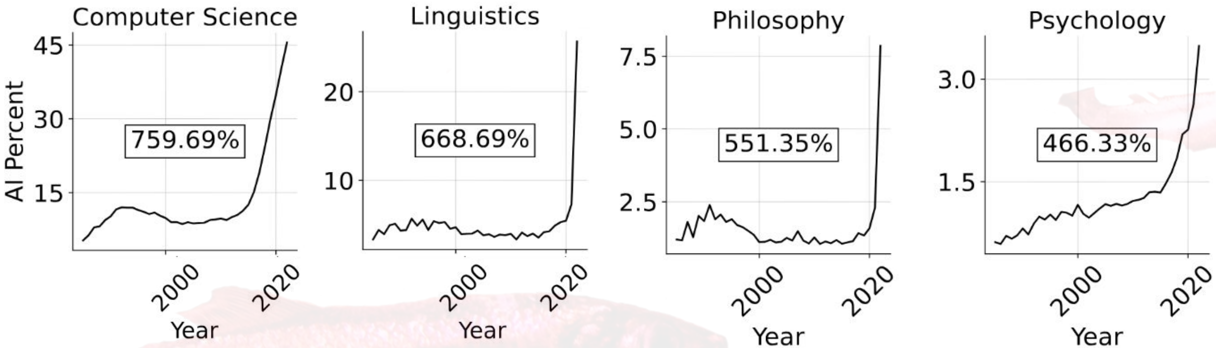


Figure 2: Change in AI engagement percentage from 1985 - 2023 by field. Inserts tally the total change in percentage of AI-engaged publications for each field.

[1] Plots from Duede et al. 2024

Flying High

Today’s Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

My Answer:

LLMs show us that there is more than one way to “know” language

- **Nothing about LLMs is human-like, yet they can do a lot with language**
- **They even surpassed human performance on many tasks (see previously popular benchmarks like GLUE and SuperGLUE[1])**

[1] https://super.gluebenchmark.com/

Today’s Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

My Answer:

LLMs show us that there is more than one way to “know” language

- Nothing about LLMs is human-like, yet they can do a lot with language
- They even surpassed human performance on many tasks
  (see previously popular benchmarks like GLUE and SuperGLUE[1])

I’d like to present a way of thinking about LLMs and language that I believe is more scientifically justified than the one that I have been criticizing.

[1] https://super.gluebenchmark.com/

Today’s Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

Let’s Revisit a Cliché Analogy

- **Human and LLM language capacities are like birds and airplanes in flight**
- **Airplanes did the supposedly impossible, but not by imitating birds**

Today's Prompt:

Let's Revisit a Cliché Analogy

- Human and LLM language capacities are like birds and airplanes in flight
- Airplanes did the supposedly impossible, but not by imitating birds
- Airplanes can do things birds can't

	Exceed Speed of Sound	Reach Outer Space
Airplane	✔ Achieved	✔ Achieved
Bird	✘ Struggling	✘ Still Trying

Today’s Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

Let’s Revisit a Cliché Analogy

- Human and LLM language capacities are like birds and airplanes in flight
- Airplanes did the supposedly impossible, but not by imitating birds
- Airplanes can do things birds can’t and vice-versa

	Exceed Speed of Sound	Reach Outer Space	Use seeds and bugs as fuel
Airplane	✔ Achieved	✔ Achieved	✘ No, Not Yet
Bird	✘ Struggling	✘ Still Trying	✔ Achieved

Today's Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

Let's Revisit a Cliché Analogy

- Human and LLM language capacities are like birds and airplanes in flight
- Airplanes did the supposedly impossible, but not by imitating birds
- Airplanes can do things birds can't and vice-versa

	Exceed Speed of Sound	Reach Outer Space	Use seeds and bugs as fuel	Make more of self 🐣
Airplane	✔ Achieved	✔ Achieved	✘ No, Not Yet	✘ I Hope Not
Bird	✘ Struggling	✘ Still Trying	✔ Achieved	✔ Achieved

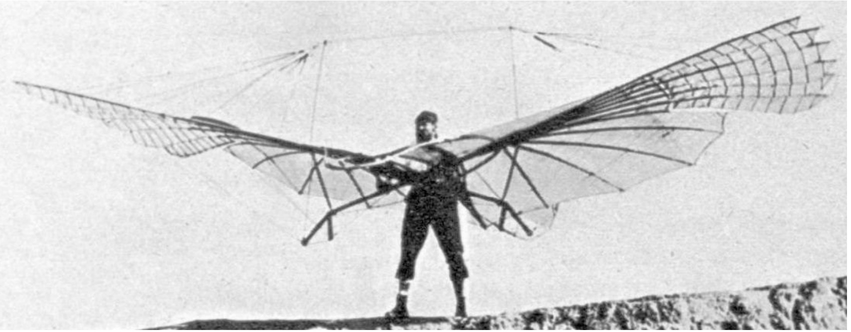
Today’s Prompt:

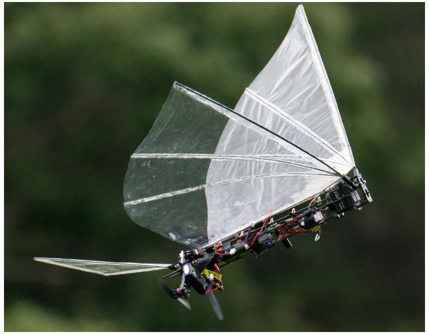
WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

Let’s Revisit a Cliché Analogy

- Human and LLM language capacities are like birds and airplanes in flight
- Airplanes did the supposedly impossible, but not by imitating birds
- Airplanes can do things birds can’t and vice-versa

We’ve rocketed to the moon, but we can hardly build a flapping flying machine





Today’s Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

Let’s Revisit a Cliché Analogy

- Human and LLM language capacities are like birds and airplanes in flight
- Airplanes did the supposedly impossible, but not by imitating birds
- Airplanes can do things birds can’t and vice-versa

We’ve rocketed to the moon, but we can hardly build a flapping flying machine

Still, we don’t even tentatively adopt the hypothesis that birds fly like planes!

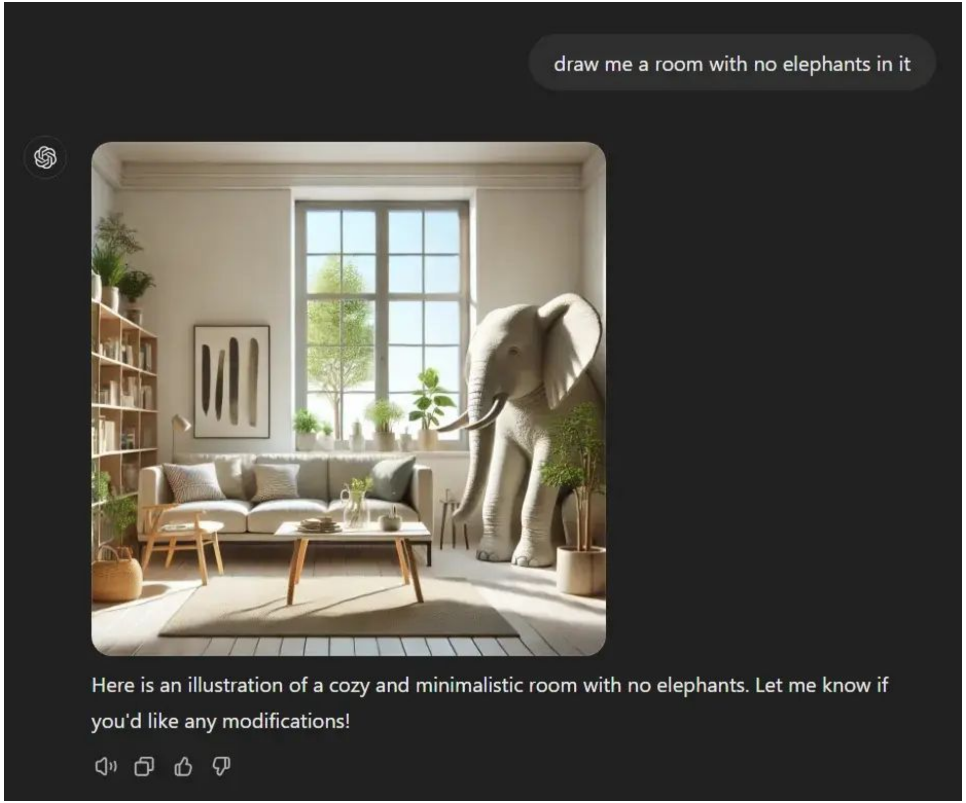
Today's Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

A More Concrete Example

(courtesy of Reddit r/ChatGPT)[1]

	Negation	Draw 🐘
Human	Easy[2]	Hard
ChatGPT	Hard	Easy



draw me a room with no elephants in it

Here is an illustration of a cozy and minimalistic room with no elephants. Let me know if you'd like any modifications!

[1] https://www.reddit.com/r/ChatGPT/comments/1hpte06/gg/

[2] e.g., Gomes et al. (2023)

Let’s not be luddites here

I’ve been critical of proprietary LLMs, but they’re useful

- Particularly when they’re being used as a tool
- Sometimes they’re just software like any other

Microsoft’s Microsoft Word	for word processing
Github’s (Microsoft) Copilot	for programming support
OpenAI’s (~49% Microsoft) GPT	for all kinds of uses

Let’s not be luddites here

I’ve been critical of proprietary LLMs, but they’re useful

- Particularly when they’re being used as a tool
- Sometimes they’re just software like any other

Microsoft’s Microsoft Word	for word processing
Github’s (Microsoft) Copilot	for programming support
OpenAI’s (~49% Microsoft) GPT	for all kinds of uses

But when an LLM is the object of study or source of insight…

- Maybe try out one of the models at the top of the Open leaderboard
- Or reach back just a few years and use a big transformer like BERT

Use them for what they are!

LLMs aren’t models of human cognition, but they’re obviously incredibly powerful tools

- They are the most powerful statistical pattern detectors we have
- They can find subtle patterns in massive amounts of behavioral data

  Many kinds of behavior → text corpora, neural activity…
- I’m going to focus on just one application that has to do with modeling

Population-level behavior in terms of patterns of word use

Distributional Semantics

A population-level description of language use

- LLMs trained on large corpora are particularly well-suited for extracting patterns of word use → distributional semantics
- Reveals organization of the (externalized) lexicon

Distributional Semantics

A population-level description of language use

- LLMs trained on large corpora are particularly well-suited for extracting patterns of word use → distributional semantics
- Reveals organization of the (externalized) lexicon
- Infamously cryptic to interpret, so research on extracting useful info from them is important

  Erk & Apidianaki (2024) - mapping axes to gradable properties

  Chronis, Mahowald, & Erk (2023) - mapping to human-readable properties

Application - Describing Change in Language Use

A contribution to the study of language change

- Once again, this is just a population-level description of what has changed

✘ Cannot reveal the cognitive mechanisms driving language change

✔ But we can uncover patterns that would have otherwise gone unnoticed

Change in the first place

Direction of change

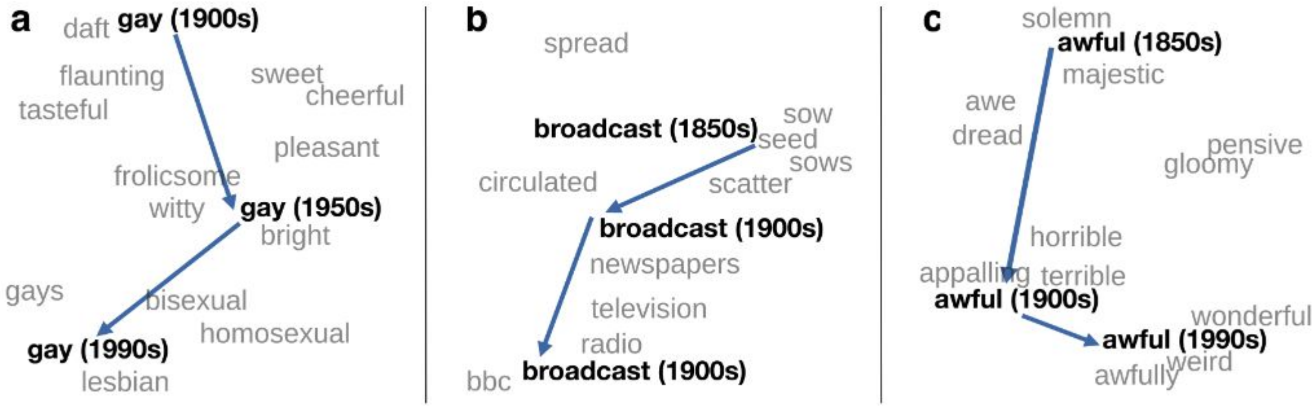
Time course of change…

Application - Describing Change in Language Use

A contribution to the study of language change

- ● Once again, this is just a population-level description of what has changed
- ✘ Cannot reveal the cognitive mechanisms driving language change
- ✔ But we can uncover patterns that would have otherwise gone unnoticed

Hamilton et al. (2016)
word2vec embeddings



Application - Describing Change in Language Use

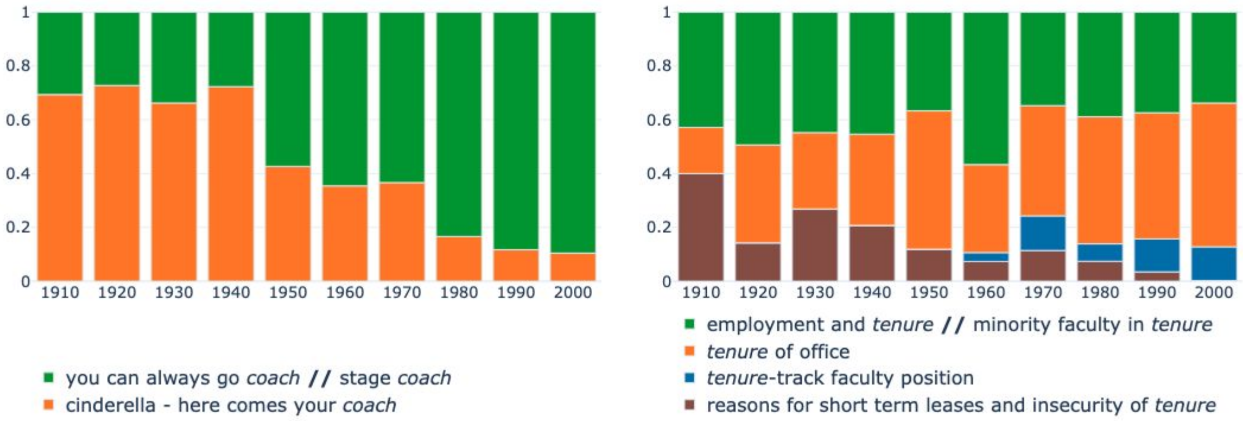
A contribution to the study of language change

- Once again, this is just a population-level description of what has changed

✗ Cannot reveal the cognitive mechanisms driving language change

✓ But we can uncover patterns that would have otherwise gone unnoticed

Giulianelli et al. (2020) contextual embeddings (BERT), data from COHA



(a) *coach*

(b) *tenure*

Conclusion

Today’s Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

My Answer:

LLMs show us that there is more than one way to “know” language

- **LLMs don’t seem to “do” language like humans**
- **This limits what they can directly tell about human language cognition**
- **Yet they are still powerful pattern finders and users of language data**

Today’s Prompt:

We can’t necessarily take evaluations at face-value 🦆

- Overuse of NLP-style evaluations obscure cognitively interesting performance

Today’s Prompt:

We can’t necessarily take evaluations at face-value 🦆

- Overuse of NLP-style evaluations obscure cognitively interesting performance

They don’t behave like a human would when you look carefully 🤖

- 40 years of evidence: LLMs don’t acquire morphology like humans

Today's Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

We can't necessarily take evaluations at face-value 🦆

- Overuse of NLP-style evaluations obscure cognitively interesting performance

They don't behave like a human would when you look carefully 🤖

- 40 years of evidence: LLMs don't acquire morphology like humans

The whole premise of evaluating performance is a red herring 🐟

- The idea that if an LLM eventually behaves like a human then it will tell us how linguistic cognition works is fallacious reasoning

Today’s Prompt:

WHAT DO LARGE LANGUAGE MODELS TELL US ABOUT HUMAN LANGUAGE?

We can’t necessarily take evaluations at face-value 🦆

- Overuse of NLP-style evaluations obscure cognitively interesting performance

They don’t behave like a human would when you look carefully 🤖

- 40 years of evidence: LLMs don’t acquire morphology like humans

The whole premise of evaluating performance is a red herring 🐟

- The idea that if an LLM eventually behaves like a human then it will tell us how linguistic cognition works is fallacious reasoning

They’re certainly useful, but so are airplanes… 🛩️

- There may be applications for LLMs in cognitive science, but their nature sharply limits the insights that they can provide

Today's Prompt:

Use them for what they are

[distributional pattern extractors],

and don't try to make them what they aren't

[models of language cognition]

Some recent work making similar points

Cuskley et al. (2023), Katzir (2023), Kodner et al. (2023), Lan et al. (2024), Milway (2023), Payne & Kodner (*under review*), Rawski & Heinz (2019), Vázquez Martínez et al. (2023), Ziv et al. (2025)...

LLMs and Linguistics:

Use them for what they are, and don’t try to make them what they aren’t

My collaborators in works cited (listed alphabetically)

Stony Brook University:
Jeffrey Heinz, Salam Khalifa, Sarah Payne

University of Pennsylvania:
Nitish Gupta (now Google), Annika Heuser, Héctor Vázquez Martínez, Charles Yang

University of Utah:
Caleb Belth M

University of Florida:
Zoey Liu

Stony Brook University

iACS
INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE