Going Beyond “Humanlike”: Using LLMs to advance psycholinguistic theories

Ethan Gotlieb Wilcox
Assistant Professor of Computational Linguistics
Department of Linguistics, Georgetown University

Intro: What is an ANN LM?

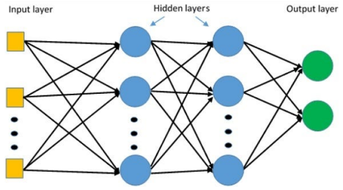
"ANN LLM": Artificial Neural Neural Network based Language Model

"ANNs"

Learning algorithms that combine:

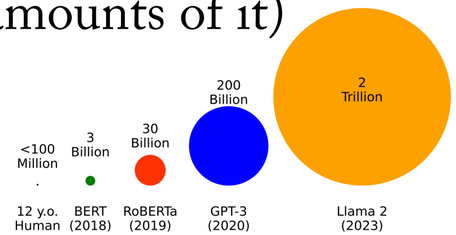
- Continuous representations (i.e., vectors of real numbers)
- Learns non-linear relationships across multiple layers

Visualization of a multi-layer perceptron



- Trained on data (often large amounts of it)

Comparison of training data scales between LMs and people

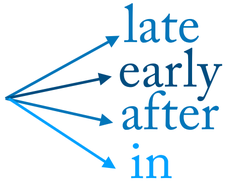


"LMs"

Algorithms that assign a probability to a string of text

- Provide context-sensitive distributions over words

For a long time I went to bed..



ChatGPT Claude Gemini

Form the basis of commercial chatbots

Intro: ANNs vs. Humans

- Artificial Neural Network models process language ***extremely*** differently from people

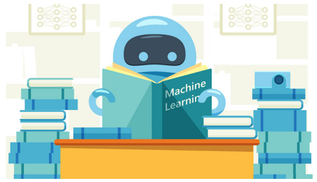
Language processing in ANNs

Not physically Embodied

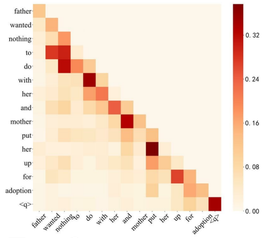
Common Crawl

WIKIPEDIA

Often trained on text data alone



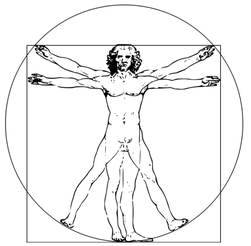
Not trained in social, interactive contexts



Perfect memory over a limited context

Processing runs on a computer

Language processing in people



Embodied

Computations run on "wetware"

Grounded

Interactive / Social

Bob threw the trash…

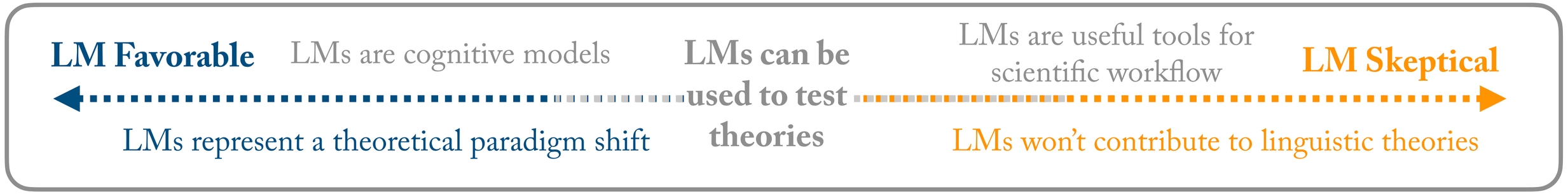
Rely on fallible memory representations

Minimize cross-entropy loss

Communicative Intent

Intro: Role of LLMs in Language Science?

1. Given the vast differences between the two, can ANN LMs help us to understand *human* language processing?
2. If so, what is the appropriate role for ANN LLMs to play in our scientific research?



- Treat a trained LM as instantiating a linguistic theory (Baroni, 2021; Piantadosi, 2023)

  "[ANN's] parameters come to embody a theory of language […] The exact same logic of tuning parameters to formalize and then compare theories is found in other sciences, like modeling hurricanes or pandemics…" (Piantadosi, 2023)

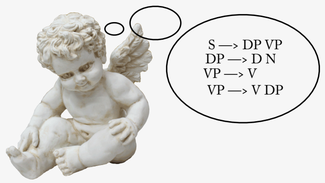
- Difference between probability and grammaticality (Chomsky 1957)
- Models are incapable of understanding meaning (Bender and Koller, 2020)
- A big matrix of numbers can't advance our theories of language

Middle Ground: Language models can be used to empirically test and refine theories of language learning and processing

Intro: Using LLMs to Test Theories

Role for ANN LMs: Language models can be used to empirically test and refine linguistic theories

Theories that make direct predictions about learning outcomes (of LMs)



- What can be learned *in principle* using LLMs (Wilcox et al., 2023; *Linguistic Inquiry*)
- What do LLMs learn when trained on a human-size amount of data? (BabyLM Challenge)

Why Info-theory Approaches?

- Precise models that make **broad-coverage, testable** predictions
- Relatively neutral with respect to linguistic formalisms
- View them as **complementary** (not contradictory) with symbolic approaches

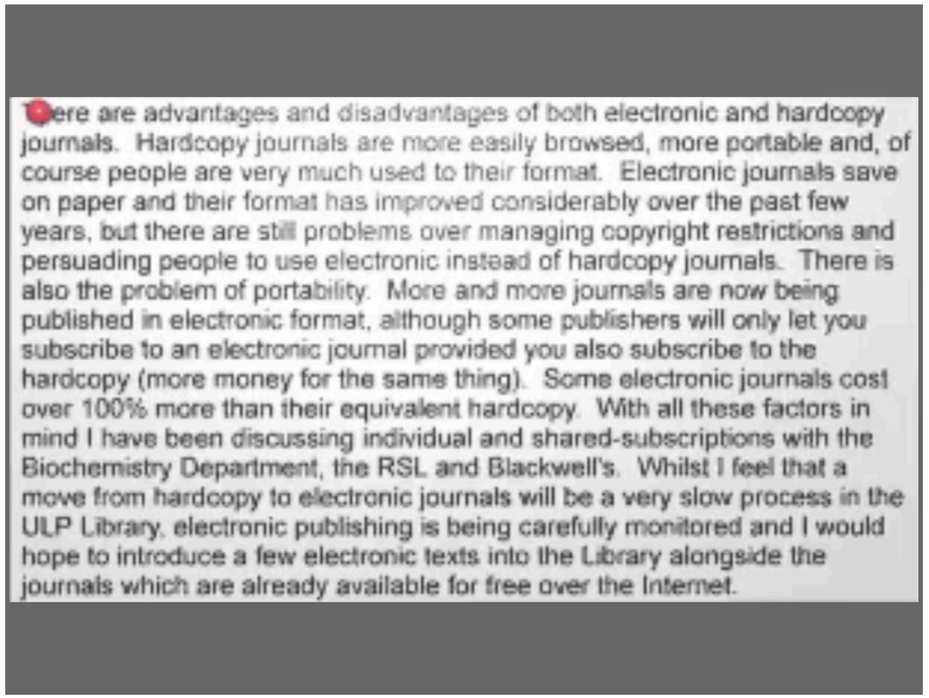
Theories that link human behaviors to the statistical properties of words

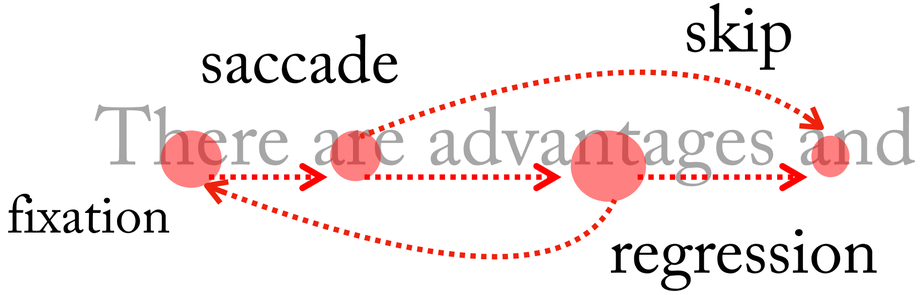
Information-Based Theories of Language Processing

- Hypothesize that languages are designed as a code that maximizes communication relative to certain constraints
- Use the branch of mathematics that describes optimal codes (*Information Theory;* Shannon 1948) with respect to human languages

Phenomena: Real-time Lang. Processing

- **Today:** explore these questions by looking at naturalistic reading behavior
- Recording and analyzing reading gives us a window into what abstract representations drive language processing in real-time





Observe similar behaviors via other measurement tools

- Self-paced reading (SPR) (Just et al., 1982)
- Mouse Tracking for Reading (Wilcox et al., 2023)
- Maze word-by-word judgment task (Forster et al., 2009)

- **Can we use principles of information theory to explain real-time processing behaviors?**
- **Can we use estimates from LLMs as links between theory and data?**

Talk Outline

- **Intro: A role for LLMs**
- **Part 1: Incremental Processing Times — Surprisal Theory**

  Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger Levy "Testing the Predictions of Surprisal Theory in 11 Languages" *TACL, 2023*

  Ethan Gotlieb Wilcox, Pranali Vani and Roger Levy "A Targeted Assessment of Incremental Processing in Neural Language Models and Humans" *ACL, 2021*
- **Part 2: Regressive Saccades — Reactivation vs. Reanalysis**

  Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister and Ryan Cotterell, "An Information-Theoretic Explanation for Regressions during Reading" *Cognition, 2024*
- **Conclusion and Discussion**

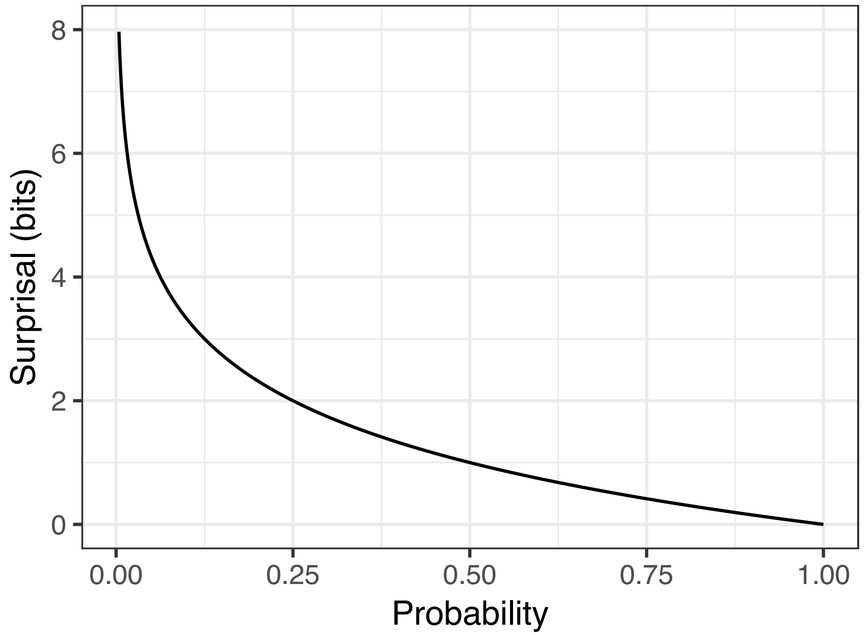
Study 1: Surprisal Theory

Surprisal Theory: Processing difficulty is proportional to word predictability (Hale, 2001;

$$\text{Reading time}\,(w_i) = f\,(\text{surprisal}\,(w_i|\mathbf{w}_{<i}))$$

$$\text{Surprisal}(w_i) \equiv \log \frac{1}{P(w_i|\text{CONTEXT})}$$

$$\approx -\log P(w_i|w_{1\ldots i-1})$$



Role for Computational modeling:

- Remember: LMs give us incremental probability distributions
- If surprisal theory is correct then…
  - A well-tuned Language Model should be predictive of processing times
  - The relationship between processing times and surprisal values should be linear

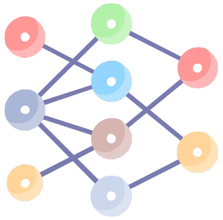
Study 1: Approach

- Test surprisal theory by asking how well LMs predict human language processing behavior during **naturalistic reading**

---- ------ -- **the** ----

Word-by-word processing measures

???



Word-by-word surprisal values

Word frequency

Word length

- My own (and others') previous work has found evidence supporting surprisal theory using this methodology (Wilcox et al., 2020; Smith & Levy, 2013; Goodkind and Bicknell 2018; Meister et al., 2021; Kuribayashi et al., 2021)
- **Shortcoming:** Vast majority of these studies are in English

Testing the Predictions of Surprisal Theory in 11 Languages

Ethan G. Wilcox[1] **Tiago Pimentel**[2] **Clara Meister**[1] **Ryan Cotterell**[1] **Roger P. Levy**[3]
[1]ETH Zürich, Switzerland [2]University of Cambridge, UK [3]MIT, USA
ethan.wilcox@inf.ethz.ch tp472@cam.ac.uk clara.meister@inf.ethz.ch
ryan.cotterell@inf.ethz.ch rplevy@mit.edu

Study 1: Methods

Technical Details

Linear regressor

Target regression = Reading_time(w) ~ f_θ( +  + )

ΔLogLik = Log Likelihood(**target**) - Log Likelihood(**baseline**)

Baseline regression = Reading_time(w) ~ f_θ( + )

- Positive ΔLogLik means that adding surprisal helps to predict reading times
- Models with higher ΔLogLik are said to have better "predictive power" or "psychological accuracy" (Goodkind & Bicknell, 2018, Frank & Bod 2011)

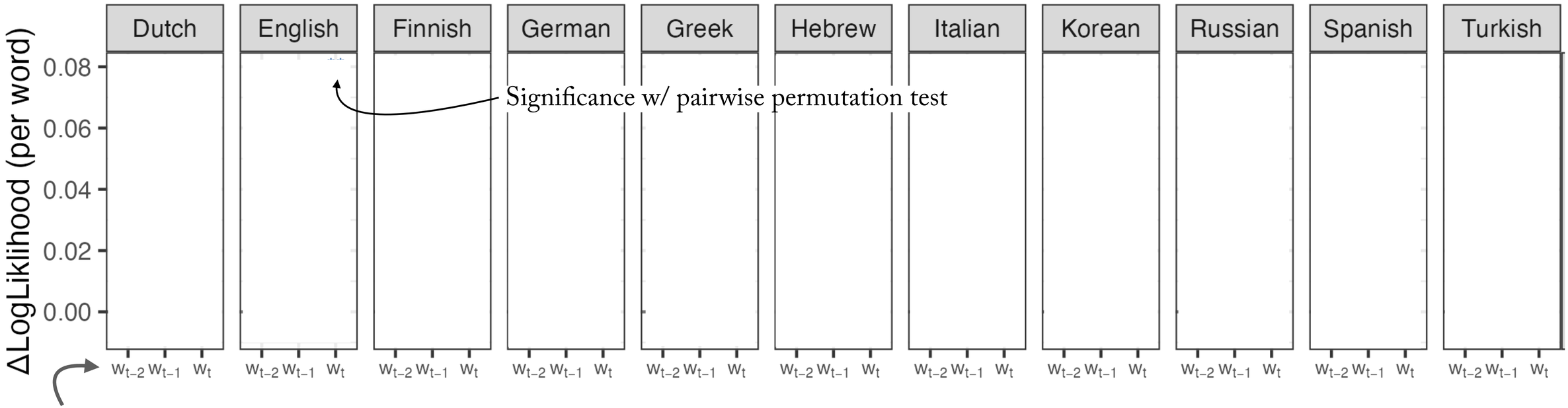
Experimental Setup:

- Measure ΔLogLik on the MECO dataset (Siegelman et al., 2022)
  - Eye tracking data for 11 languages across 5 language families.
  - Simple wikipedia-style articles … content control across languages
- Estimate surprisal using mGPT, a large multilingual transformer based on the GPT2 architecture (Shliazhko et al., 2022)

Study 1: Results (DLL across langs)

- A well-tuned Language Model should be predictive of processing times
- The relationship between processing times and surprisal values should be linear

Do we find significantly positive ΔLogLik across languages?



Isolate the effect of surprisal on current + two previous words by training separate models with surprisal from w_t, w_{t-1}, w_{t-2}

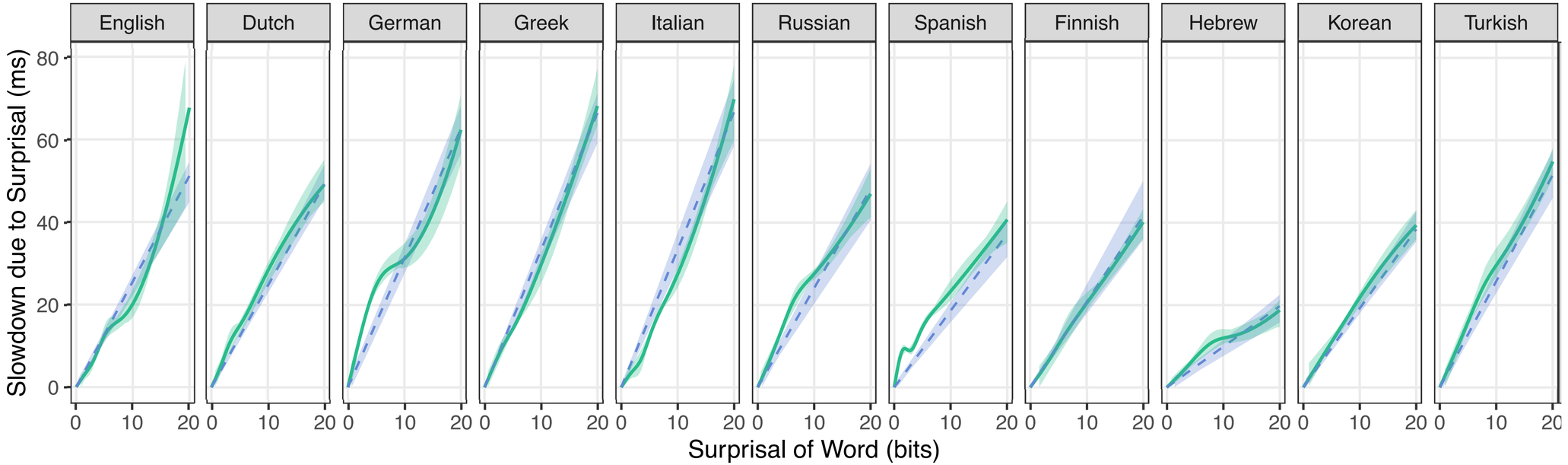
- Takeaway: Surprisal is predictive of reading times across languages
- Effect on current word consistent with little spillover found in eye-tracking

Study 1: Results (Linearity)

- A well-tuned Language Model should be predictive of processing times
- The relationship between processing times and surprisal values should be linear

What is the functional relationship between surprisal & RTs?

Two models: Predict RTs from surprisal with a **linear** and **non-linear** model



- Non-linear models essential recover a linear surprisal/reading time relationship
- Takeaway: Evidence for linear relationship between surprisal across languages

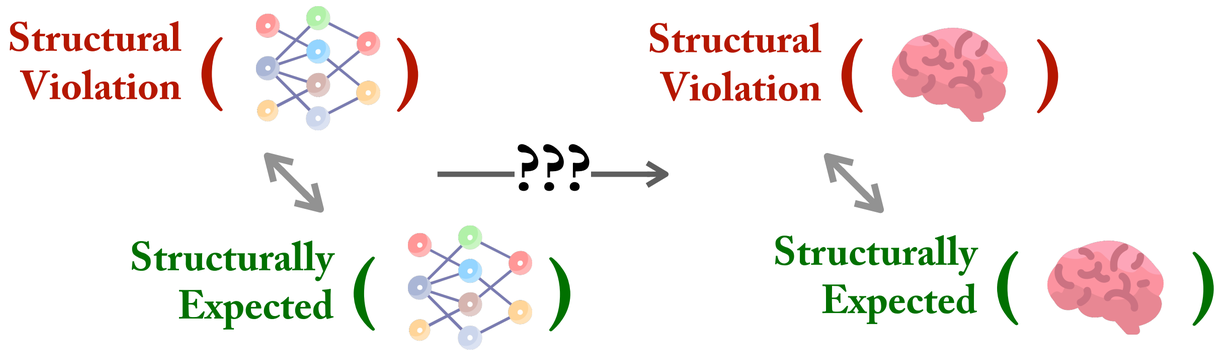
Study 2: Limits of Surprisal Theory

Study 2: Return to Surprisal Theory to show a limitation of LLMs

- Test surprisal theory in controlled experimental setting; subset of the test suites from syntaxgym.org (Gauthier et al., 2020)
- Come in structurally violating (≈ungrammatical) and structurally predictable (≈grammatical) variants

Subject/Verb Number Agreement

* The keys to the cabinet **is**…
The keys to the cabinet **are**…



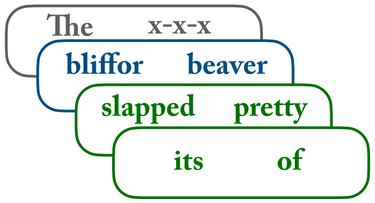


Illustration of the Maze Task

- **Maze Task:** Participants read a sentence incrementally by selecting the grammatical continuation at each point in the sentence. (Forster et al., 2009)
- Novel variant based off the web-deployable variant from Boyce et al. (2020)

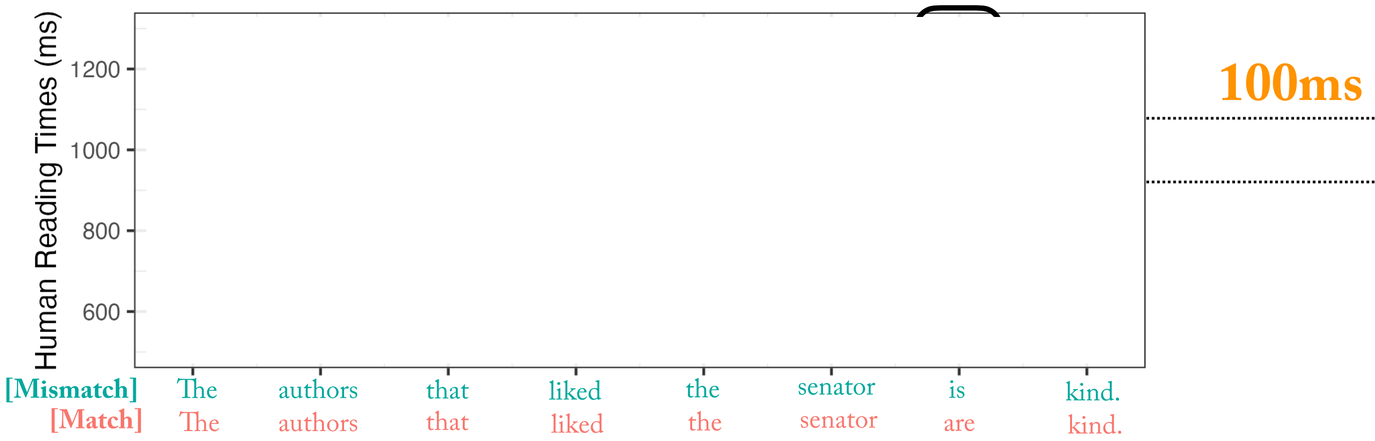
Ask me about Maze / methods in the Q&A!

Study 2: Results

- Start with subject/verb number agreement, then zoom out to all test suites.

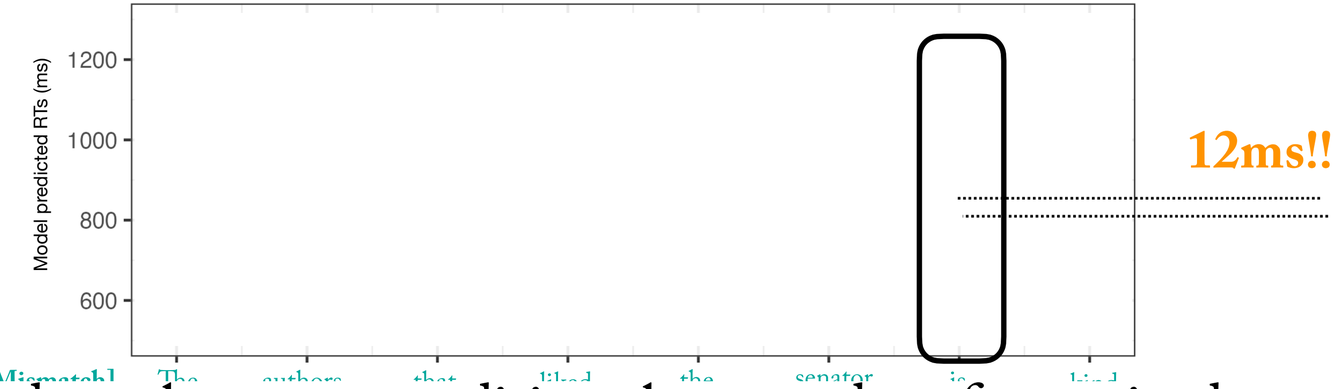
Human Results: Word by word-reading times for the Maze task

- Longer reading times in mismatch (expectation-violation) condition in critical region



Model results: Predict expected reading times with linear regression models using *surprisal*, *length* and log(*frequency*) as predictors

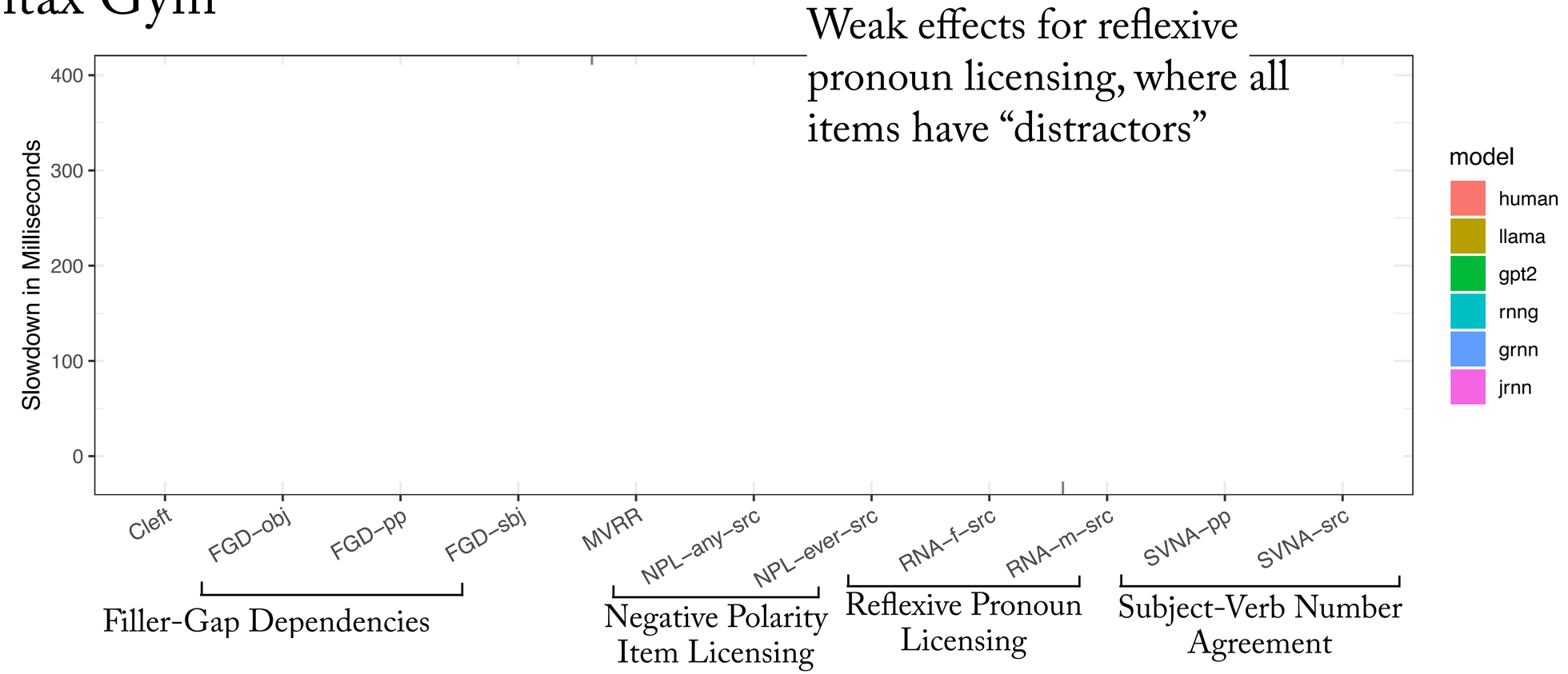
- Predictions show relatively good fit to the human data
- But what about differences between conditions in critical regions?



- Takeaway: Model underpredicts human slowdown between conditions by an order of magnitude

Study 2: More results

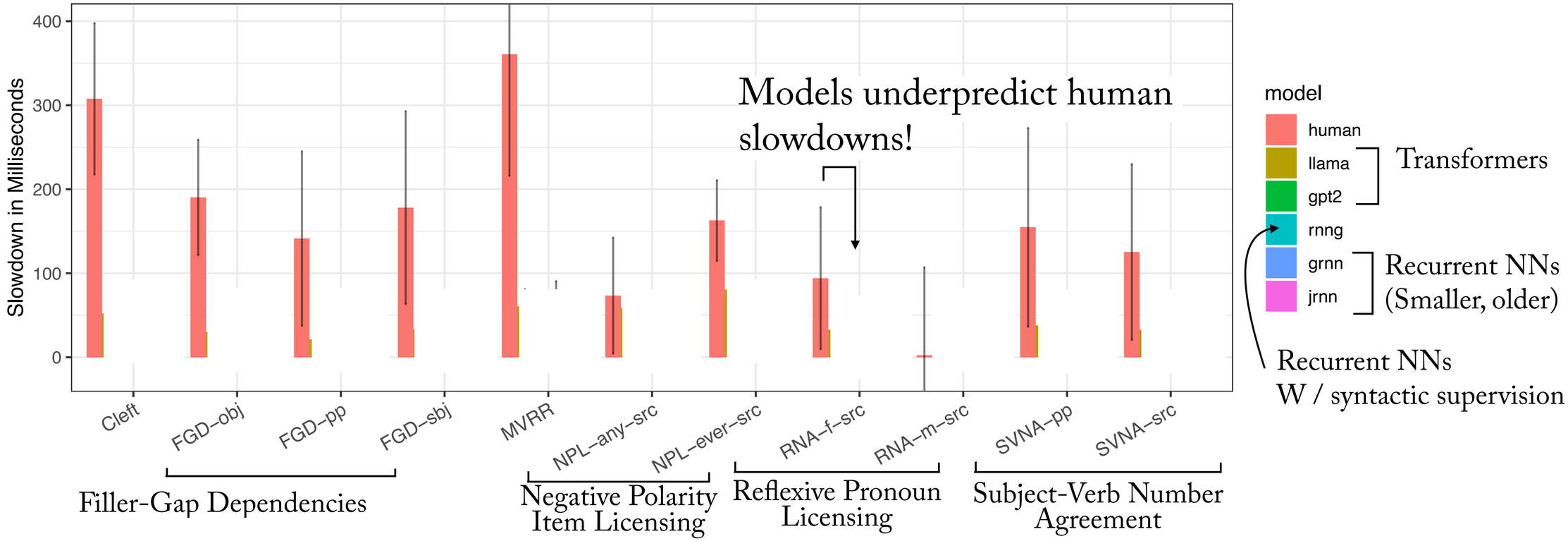
- Compare human slowdowns vs. predicted model slowdowns for many test suites from Syntax Gym



- Above zero = slowdown when reading ungrammatical sentences
- Takeaway: People show robust slowdowns in ungrammatical sentences

Study 2: More results

- Compare human slowdowns vs. predicted model slowdowns for many test suites from Syntax Gym



- Takeaway: LLMs are poor models for human reading behavior of non-typical strings

Part 1: Discussion

- Use LLMs to validate surprisal theory in 11 languages, across 5 language families
- Show that LLM probability estimates are poorly calibrated to reading times when encountring structural violations
- Dilemma: Is it a problem with the *models* or the *theory*?

Surprisal Theory is correct, but human probability distributions don't resemble those of LLMs

Surprisal Theory must be augmented to explain misallignment to probability

- Either way, accurate multilingual estimates from LLMs helped move theoretical debates forward

Talk Outline

- **Intro: A role for LLMs**
- **Part 1: Incremental Processing Times — Surprisal Theory**

  Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger Levy "Testing the Predictions of Surprisal Theory in 11 Languages" *TACL, 2023*

  Ethan Gotlieb Wilcox, Pranali Vani and Roger Levy "A Targeted Assessment of Incremental Processing in Neural Language Models and Humans" *ACL, 2021*

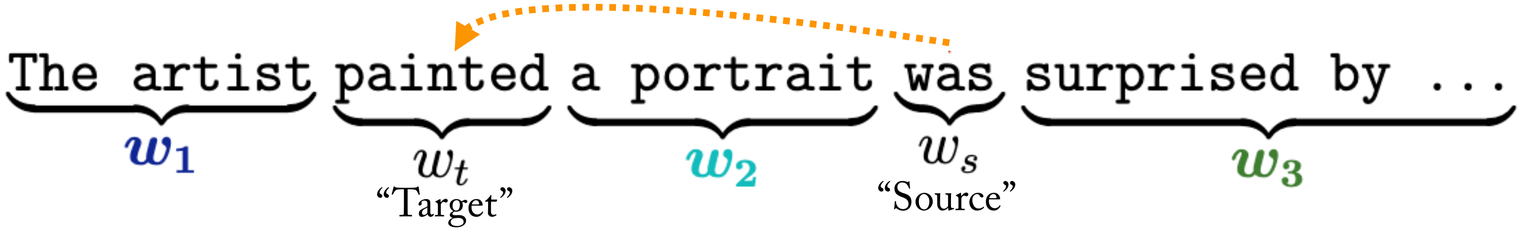
- **Part 2: Regressive Saccades — Reactivation vs. Reanalysis**

  Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister and Ryan Cotterell, "An Information-Theoretic Explanation for Regressions during Reading" *Cognition, 2024*

- **Conclusion and Discussion**

Two Theories for Regressions

- When reading, Saccades are typically **progressive** (i.e. left-to-right in English) but about 10%-20% are **regressive**, going against the general flow (Rayner, 1998); targeted regressive saccades are not well understood!



Reanalysis Hypothesis (Frazier and Rayner, 1982; Bicknell and Levy, 2010, 2011)

- Targeted regressions are used to reanalyze material about which the reader has falling confidence
- Targeted regressions occur between words that are not associated

Reactivation Hypothesis (Kennedy and Murray, 1987; Lopopolo et al., 2019)

- Targeted regressions are used to confirm an interpretive choice (e.g., word identity)
- Targeted regressions occur between words that are associated with each other

Takeaway: Theories make different predictions about the relationship between the source and target of a regression.

Operationalizing our two Theories

Reanalysis Hypothesis: targeted regressions occur between words that are not associated

Reactivation Hypothesis: regressions occur between words that are associated with each other

- Measure association with contextual **Pointwise Mutual Information** (PMI)

$$\text{PMI}(w_s, w_t \mid w_{\neq s,t}) = \log_2 P(W_s = w_s \mid w_t, w_{\neq s,t}) - \log_2 P(W_s = w_s \mid w_{\neq s,t})$$

The words in the sentence that are not w_s or w_t

- PMI measures how much *more* surprising w_s is, given w_t than if w_t was not known.
- Positive PMI (PPMI) values imply association; Negative PMI (NPMI) values imply anti-association
- PMI is a well-used measure of word association in NLP (Jurafsky and Martin, 2000)

- **Extension:** Assume that readers sometimes have imperfect memories about previous words
- Regress based on the **expected PMI** of w_t under uncertainty

$$E[\text{PMI}(w_s; W_t)] = \sum_{w \in V} p(w \mid w_s, w_{\neq s,t}) \, \text{PMI}(w; w_s \mid w_{\neq s,t})$$

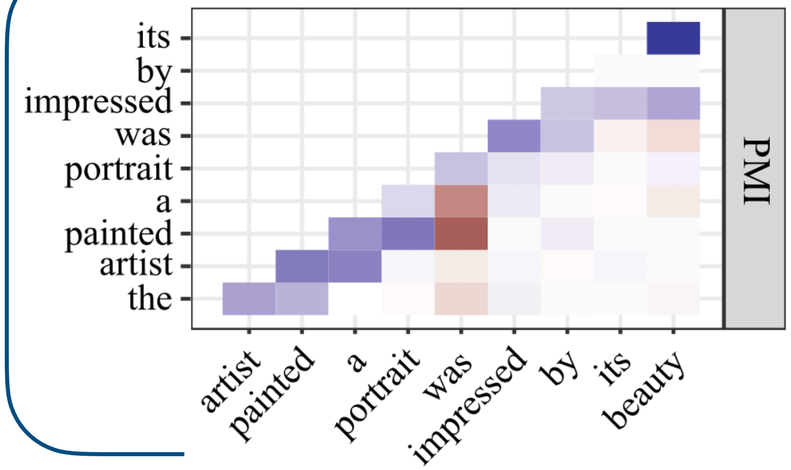
Estimating PMI with LLMs

Estimating Contextual PMI: We estimate conditional PMI values from monolingual Masked Language Models, all in the BERT family (Hoover et al., 2021)

$$\text{PMI}(w_s, w_t \mid w_{\neq s,t}) = \log_2 P(W_s = w_s \mid w_t, w_{\neq s,t}) - \log_2 P(W_s = w_s \mid w_{\neq s,t})$$

The artist **painted** a portrait was happy
The artist painted a portrait was [MASK]
$-\log_2 P(\textbf{happy})$

The artist painted a portrait was happy
The artist [MASK] a portrait was [MASK]
$-\log_2 P(\textbf{happy})$



Example: Contextual PMI estimates derived from BERT

`"The artist painted a portrait was impressed by its beauty"`

- **Blue** is Positive PMI (PPMI); **Red** is Negative PMI (NPMI)
- Positive PMI on the off-diagonal (i.e., between adjacent words)
- Negative PMI between source and disambiguator of the garden path

Study 3: Experimental Methods

- Measure ΔLogLik for predicting counts of regressions between words

Baseline: #Regressions(w_s;w_t) ~ f_θ(surp(w_s) + freq(w_s) + len(w_s) + … for w_t … + dist(w_s;w_t))

Target: #Regressions(w_s;w_t) ~ f_θ(**PMI(w_s, w_t | $w_{\neq s,t}$)** + surp(w_s) + freq(w_s) + len(w_s) + … for w_t … + dist(w_s;w_t))

f_θ is a zero-inflated Poisson regression model, due to the large number of zeros in the dataset

- **Dependent variable:** Count of regressions between each within-sentence word (summed over all participants in a dataset)
- **Datasets:** MECO (Siegelman et al., 2022), Provo (Luke and Christianson, 2018) and Dundee (Kennedy et al., 2003) UCL (Frank et al., 2013)

Also use:

Positive PMI (PPMI) = max(PMI(w_s, w_t | $w_{\neq s,t}$), 0)	**Reactivation**
Expected PPMI (E[PPMI])	
Negative PMI (NPMI) = min(PMI(w_s, w_t | $w_{\neq s,t}$), 0)	**Reanalysis**
Expected NPMI (E[NPMI])	

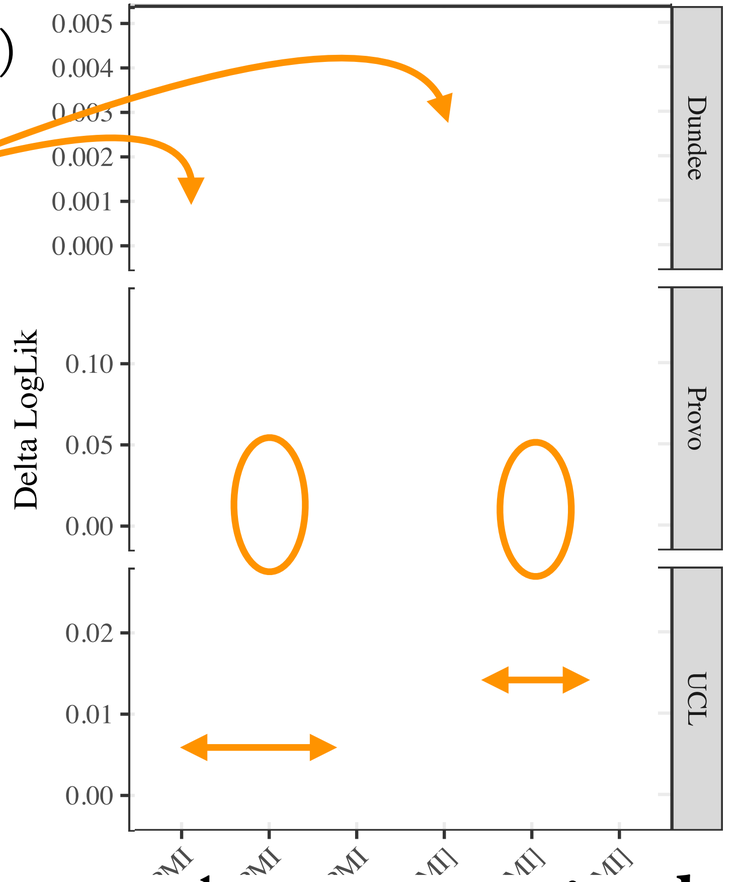
1. **Does PMI between words predict regressions in naturalistic reading?**
2. **What is a better predictor: Positive values of PMI (reactivation) or negative values of PMI (reanalysis)?**

Results: PMI and Regressions in English

- **What is a better predictor: Positive values of PMI (reactivation) or negative values of PMI (reanalysis)?**

Four Big Trends (hold across corpora)

1. Positive values of ΔLogLik for PPMI & E[PPMI]
2. Inconsistent effect of NPMI and E[NPMI]
3. Regressions are better predicted by expectations
4. PPMI/E[PPMI] models are as good as ensemble models



Additional Analyses

- **Concern:** Maybe regressions are inaccurate?
- Redo analysis with PMI-*window* around the target — **Similar results**

- **Baseline Analysis:** Fit regression with just PMI($w_s, w_t \mid w_{\neq s,t}$)

Positive effect of PMI

- As PMI between words increases, the number of regressions does too

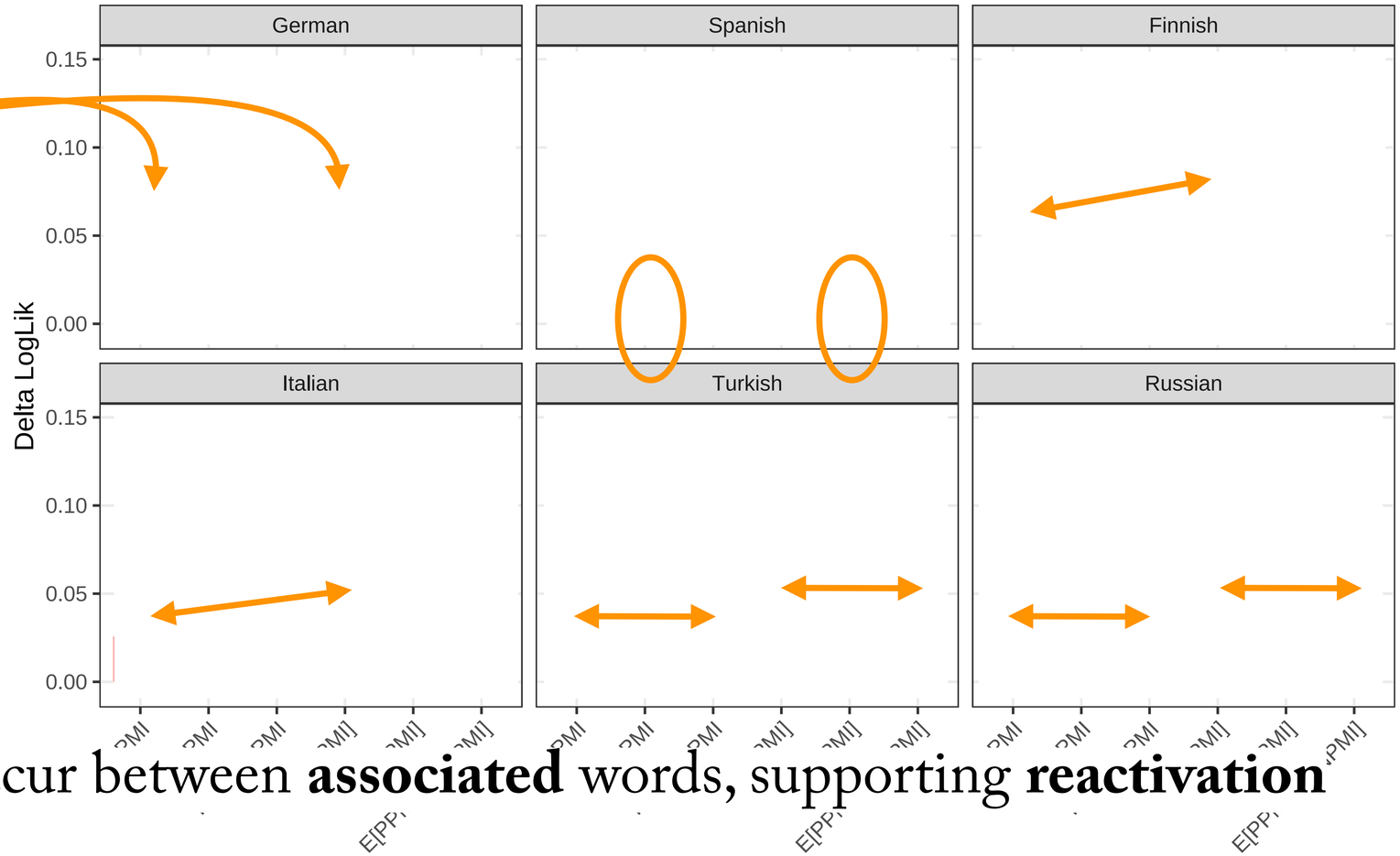
- **Takeaway:** Targeted regressions occur between **associated** words, supporting **reactivation**
- Does this pattern hold in other languages?

Results: PMI and Regressions

- **Do the results in English hold up in other languages?**
- **Test on subset of MECO, with languages from three different language families**

Four Big Trends (hold across corpora)

1. Positive values of ΔLogLik for PPMI & E[PPMI]
2. Inconsistent effect of NPMI and E[NPMI]
3. Regressions are better predicted by expectations
4. PPMI/E[PPMI] models are as good as ensemble models



- **Takeaway:** Targeted regressions occur between **associated** words, supporting **reactivation**

Study 3: Discussion

- **Theoretical Contributions:** Provided an information-theoretic interpretation of the reanalysis and reactivation theory for regressions, based on Pointwise Mutual Information
- **Results:** Supported the reactivation hypothesis:
  - Positive coefficient of PMI: As PMI between two words increases, so do the number of regressions between them
  - Adding PPMI into regressions leads to higher ΔLogLik than NPMI

- **What about Reanalysis?**
  - We know from studies on garden paths that regressions are not *obligatory*
  - People may regress between disassociated words… but not frequently enough to be picked up in these corpora analyses

First comprehensive cross-linguistic analysis of regressions I am aware of in the literature!

Talk Outline

- **Intro: A role for LLMs**
- **Part 1: Incremental Processing Times — Surprisal Theory**

  Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger Levy "Testing the Predictions of Surprisal Theory in 11 Languages" *TACL, 2023*

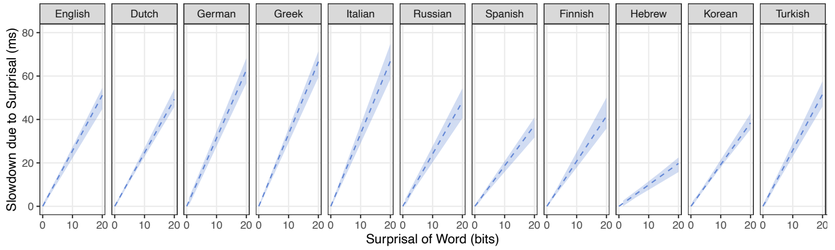
  Ethan Gotlieb Wilcox, Pranali Vani and Roger Levy "A Targeted Assessment of Incremental Processing in Neural Language Models and Humans" *ACL, 2021*
- **Part 2: Regressive Saccades — Reactivation vs. Reanalysis**

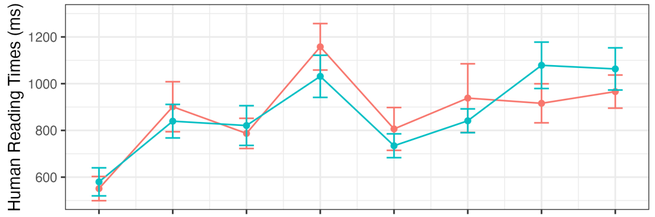
  Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister and Ryan Cotterell, "An Information-Theoretic Explanation for Regressions during Reading" *Cognition, 2024*
- **Conclusion and Discussion**

Discussion

Goal #1: Convince you that LLMs can contribute to linguistics… by empirically testing theories of language processing

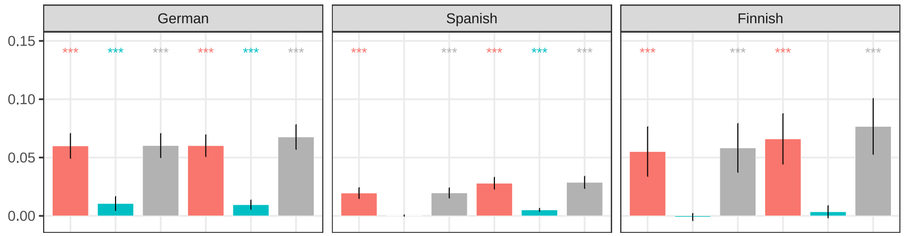
- **Experiment 1:** Used LLMs to support **surprisal theory** in multiple languages





- **Experiment 2:** Used LLMs to predict reading times for ungrammatical sentence regions

- **Experiment 3:** Used LLMs to support the reactivation hypothesis for regressions



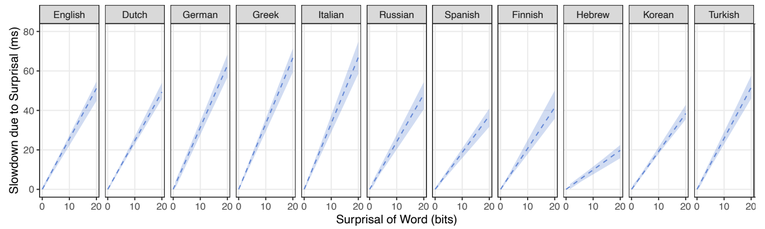
Discussion

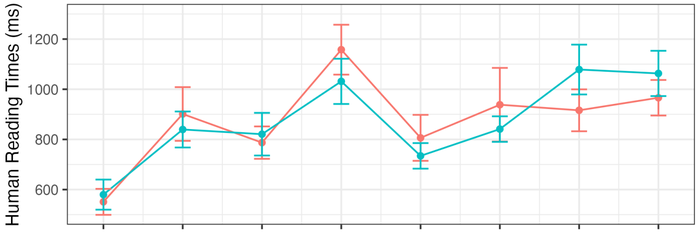
- Productive but *modest* role for LMs
  - LMs are primarily statistical estimators; tried and useful tool!
  - But… LMs allow for **powerful, multilingual, and flexible** estimation procedures
  - Quantify statistical relationships between words that were previously difficult (e.g., PMI)
- How *large* do the LMs need to be?
  - Estimation is effective when using moderate-sized LMs (120M parameters; ~2B training tokens)

Discussion

Goal #2: Case studies for how information-based theories can provide broad coverage explanations for language processing phenomena (across languages!)

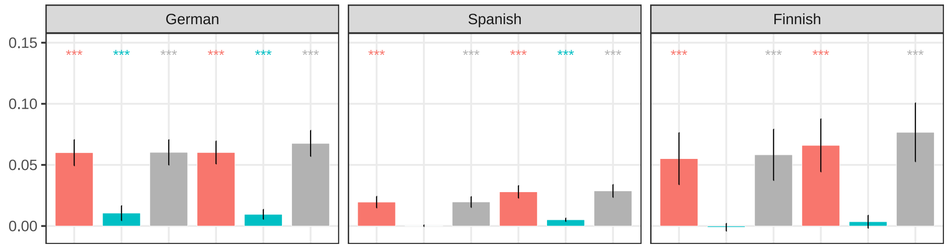
- **Experiment 1:** Reading times for a word are a linear function of its information content





- **Experiment 2:** Explored limits of information-based approaches: Abnormal processing is *not* calibrated to information content!

- **Experiment 3:** Regressions occur between words with high informational association



Discussion

- **How will LLMs change linguistics / linguistic theory?**
  - Scientists take advantage of affordances in the research environment
  - Sometimes methods drive theoretical focus
- Prediction: LLMs drive interest in theories that make predictions about word-level statistical measures
- LLMs may not "prove" or "disprove" a theoretical perspective, but may change what theories can be productively and easily tested

Thank You

Pranali Vani

MIT

Roger Levy

MIT

Ryan Cotterell

ETH Zürich

Clara Meister

ETH Zürich

Tiago Pimentel

ETH Zürich