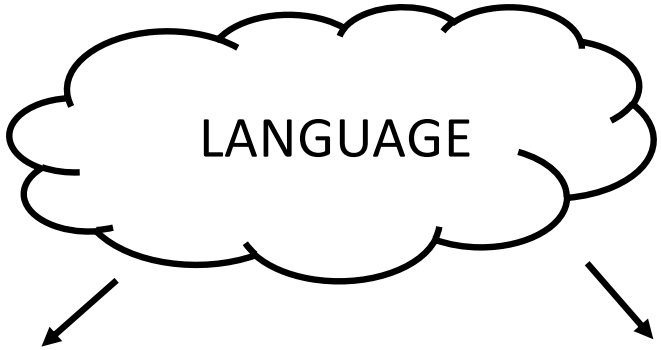
THE UNIVERSITY OF CHICAGO

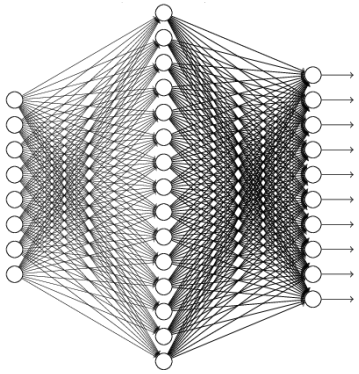
"Understanding" and prediction: Disentangling meaning extraction and predictive processes in humans and AI

Allyson Ettinger

ILFC Seminar

Nov 16, 2022





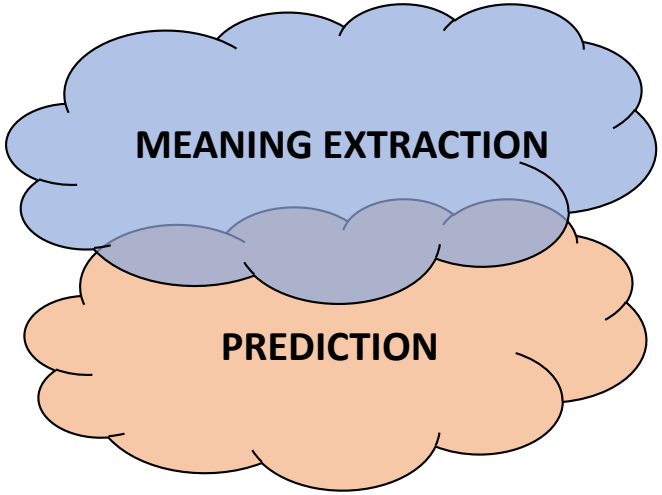
Linguistics/cognitive science

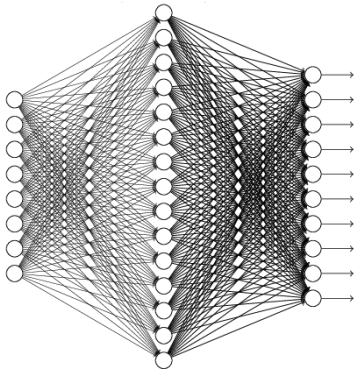
How does language work in humans?

Natural language processing (NLP)

How can we make language work effectively in AI?

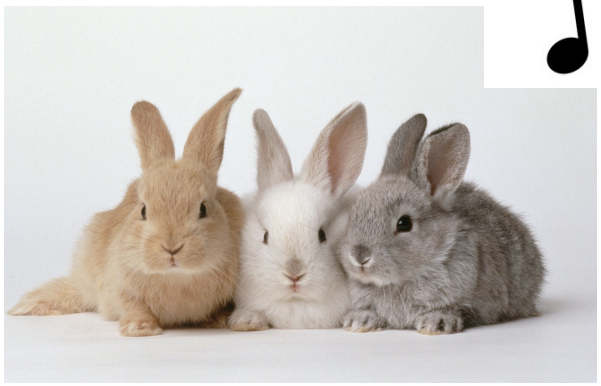
Our research

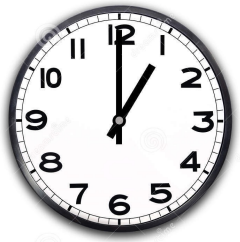




Meaning extraction in humans

Three singing rabbits walked into the local bar last Wednesday afternoon

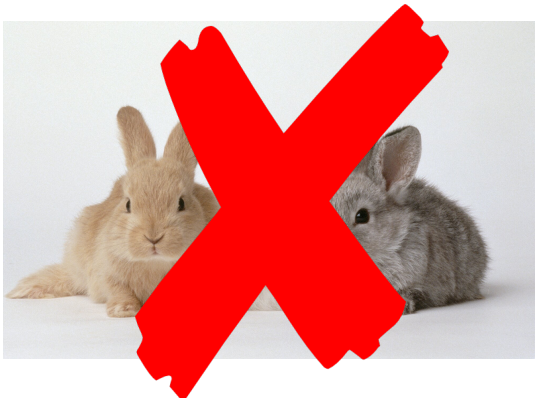




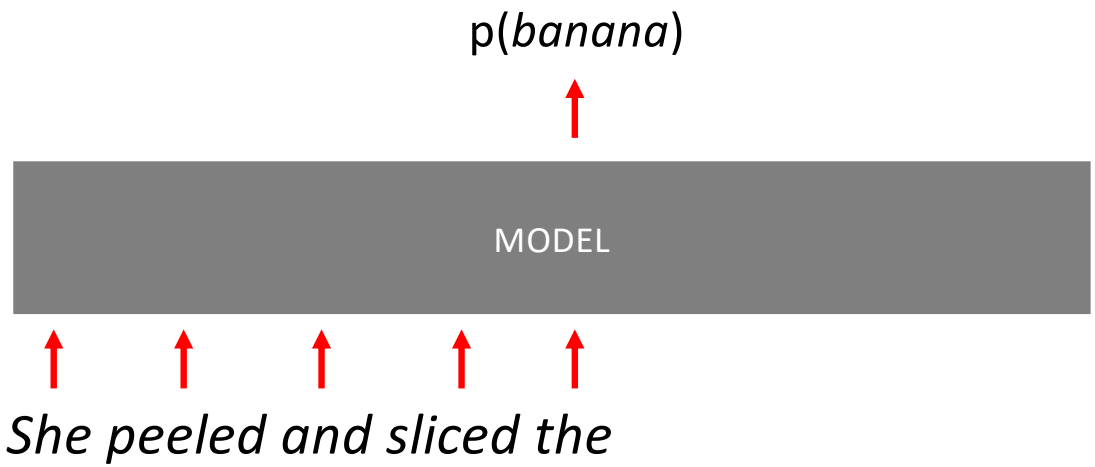
Monday	
Tuesday	
Wednesday	
Thursday	
Friday	
Saturday & Sunday	

Prediction in humans

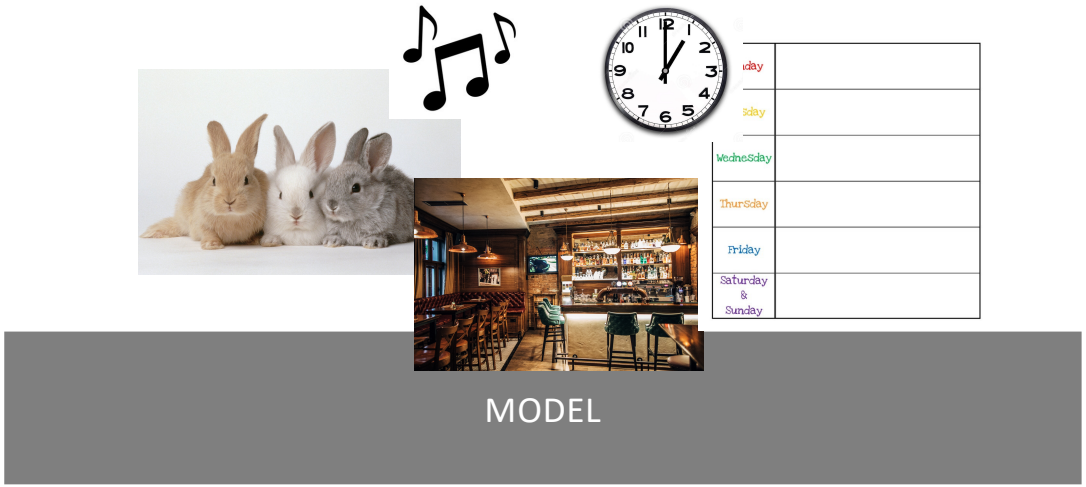
Three singing _____ walked into the local bar last Wednesday afternoon



Prediction in NLP

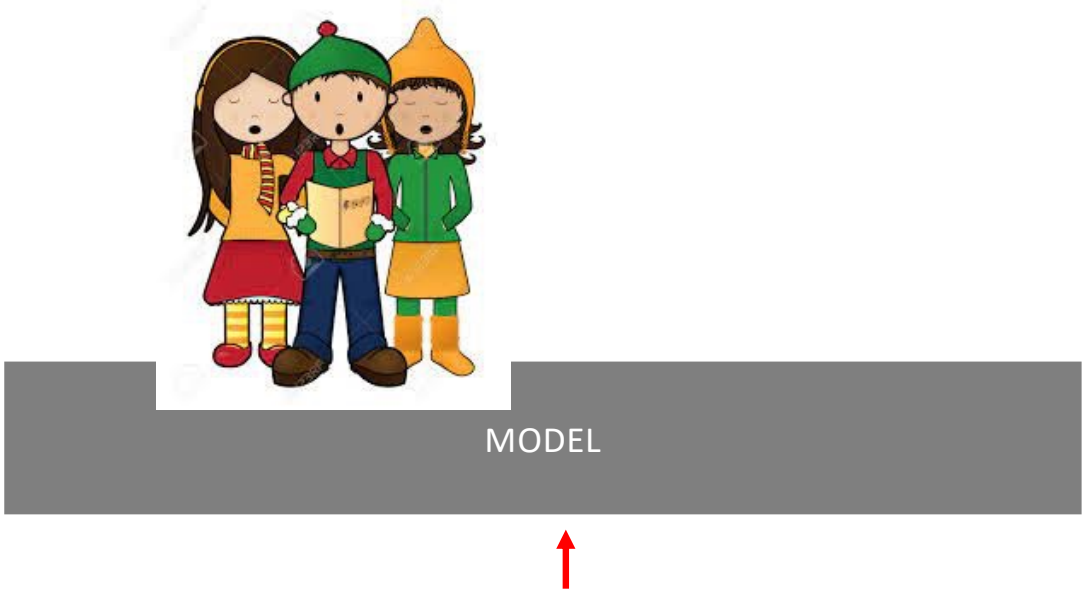


Meaning extraction in NLP



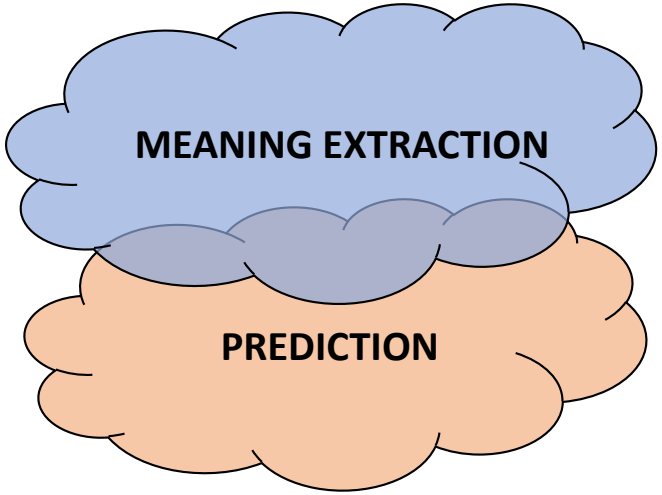
Three singing rabbits walked into the local bar last Wednesday afternoon

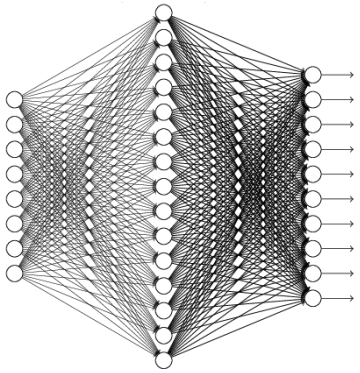
Meaning extraction in NLP



Three singing rabbits walked into the local bar last Wednesday afternoon

Our research





How do meaning and prediction processes interact in the brain during processing?

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

Key questions

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

How do meaning and prediction processes interact in the brain during real-time language processing?

Key questions

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

How do meaning and prediction processes interact in the brain during real-time language processing?

Achieving natural language “understanding”

- Ultimate goal in NLP: “natural language understanding” in AI
- Can define as human-level ability to extract, represent, and deploy information conveyed in input
- In other words: ability to extract and use meaning

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

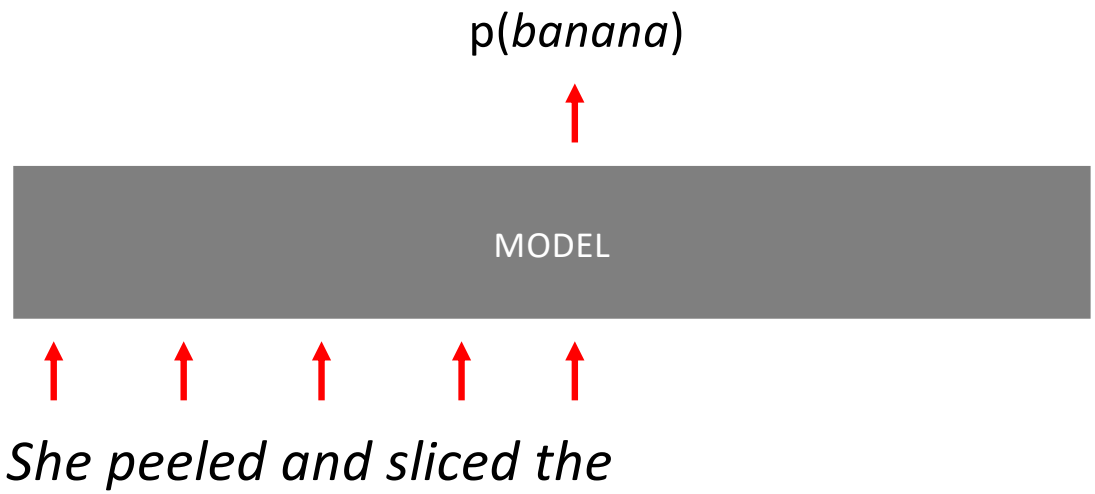
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978

Prediction-based learning

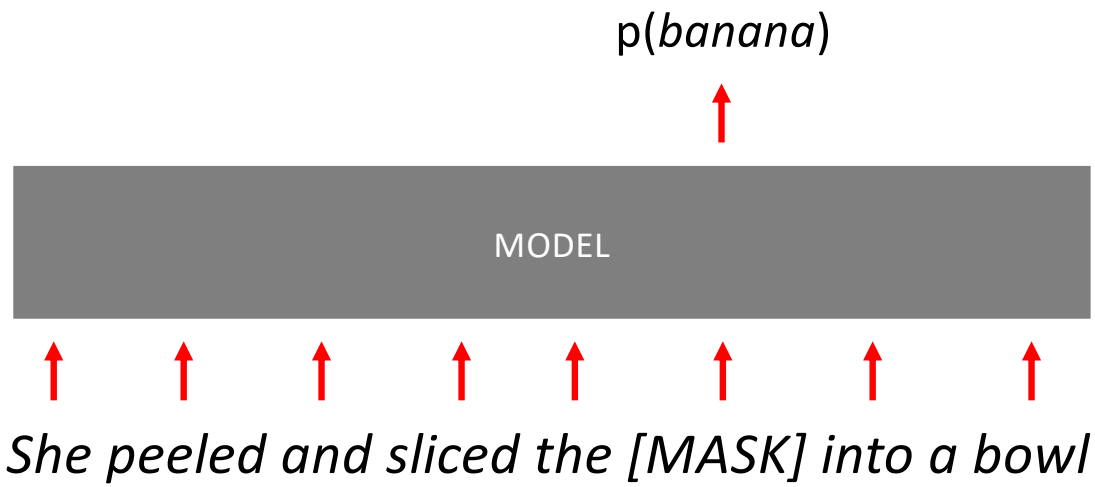
- Recent successes have been driven by pre-trained language models

- These models are trained based on prediction of words in context

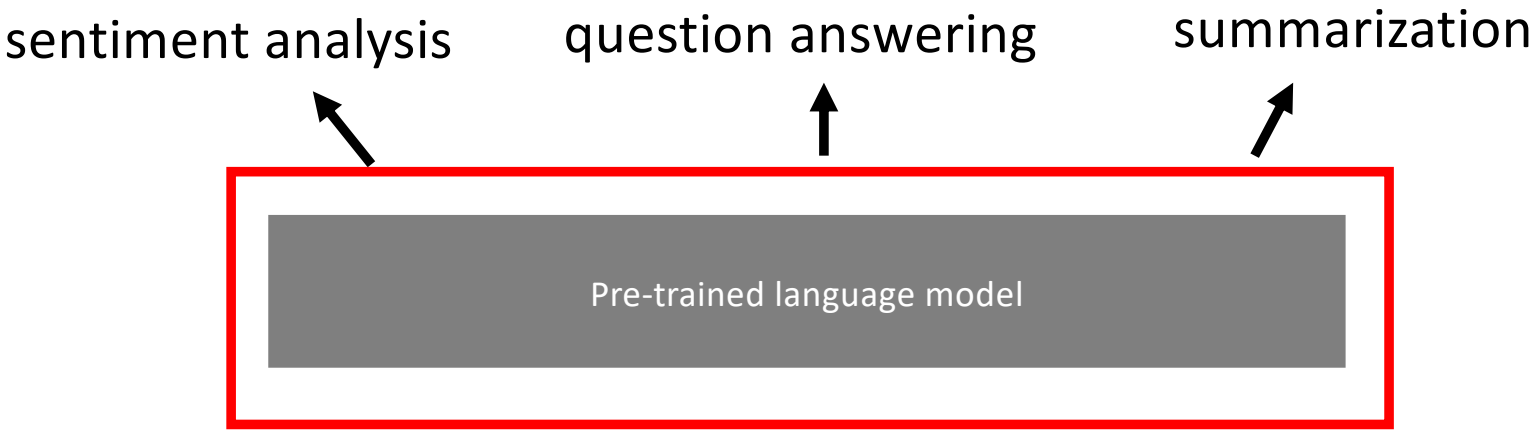
Pre-trained language models



Pre-trained language models



Pre-trained language models



Interpreting these results

- Have pre-trained LMs solved “language understanding”?
- Have these models learned to extract meaning from input in a robust manner, matching humans?

Taking a closer look

- Although NLP benchmarks aim to assess “understanding”, evidence indicates that performance is inflated by shallower factors
- Models can exploit cues and heuristics that don’t rely on understanding of meaning
- To know what these models are actually capturing, we need more effective tests to assess meaning capabilities with greater precision

Our tests

- We introduce controlled tests targeting meaning understanding
- Are pre-trained LMs acquiring a robust ability to encode and use meaning of language inputs?
- Or using shallower predictive heuristics instead?

Key questions

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

How do meaning and prediction processes interact in the brain during real-time language processing?

Key questions

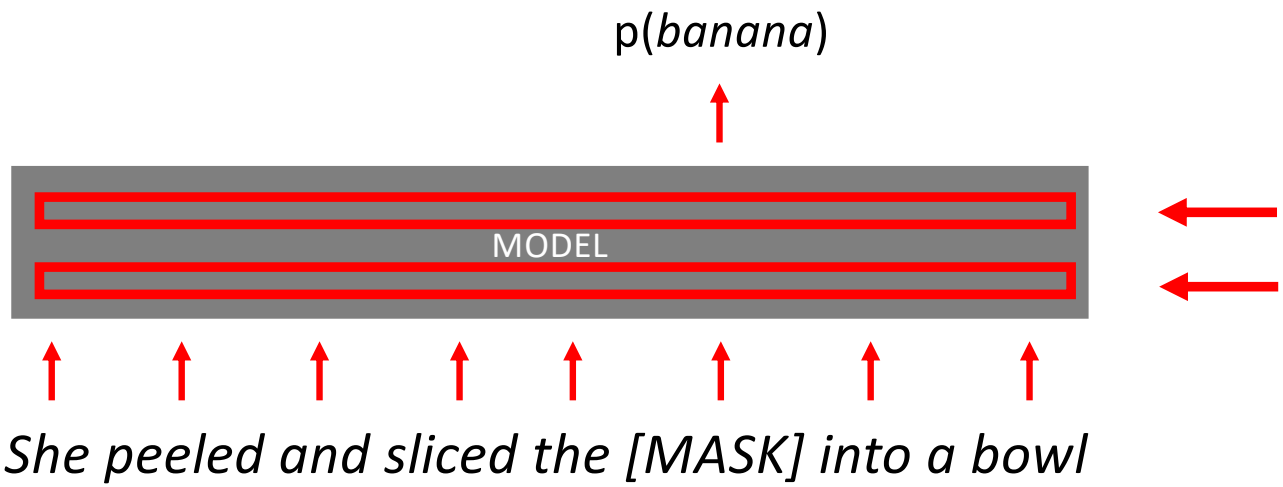
To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

Phrase composition

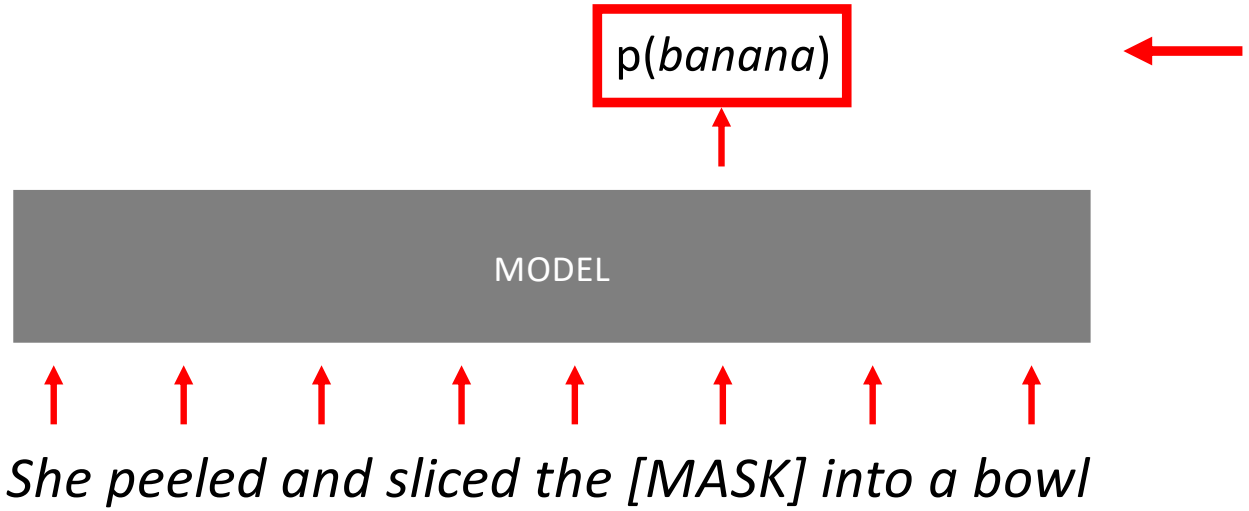
Psycholinguistic diagnostics

Information robustness

Levels of analysis



Levels of analysis



Key questions

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

Phrase composition

Psycholinguistic diagnostics

Information robustness

Key questions

Lang Yu

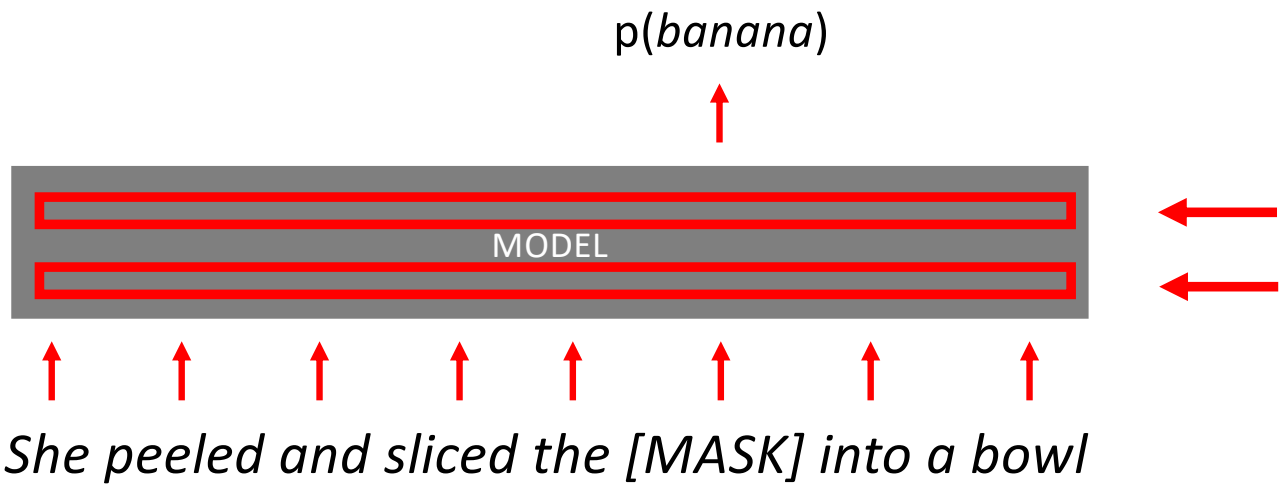
To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

Phrase composition

Psycholinguistic diagnostics

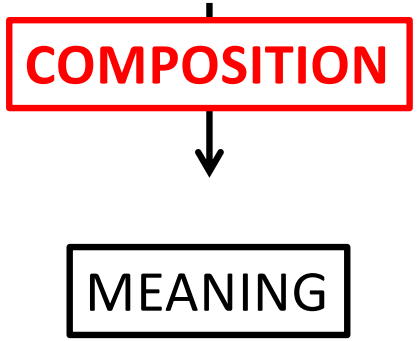
Information robustness

Levels of analysis



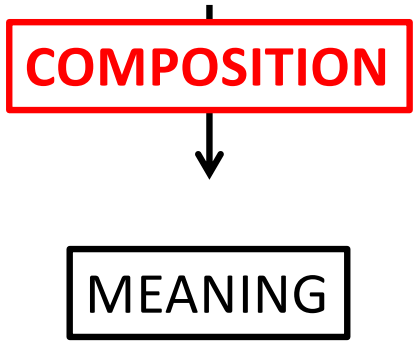
Deriving meaning systematically

The turquoise giraffe recited the sonnet but did not forgive the flight attendant



Deriving meaning systematically

The turquoise giraffe recited the sonnet but did not forgive the flight attendant



Phrase-level composition

old

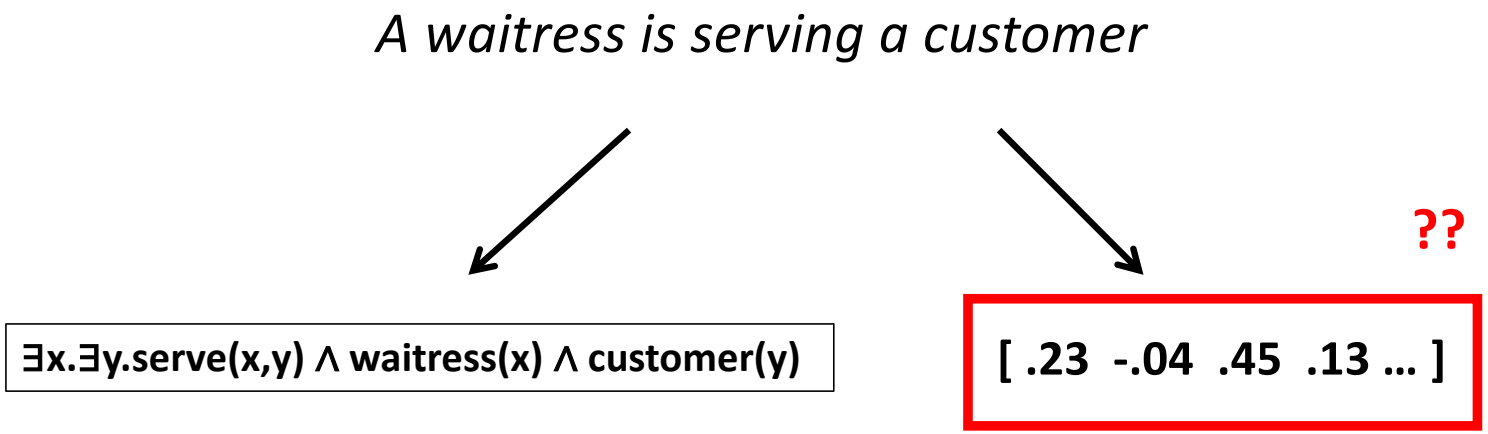
cat

old cat

old

cat

Opaque representations



Testing opaque representations

- Goal: models represent composed phrase meaning as humans do

Two simple metrics

1. Correlate representation similarity with human similarity ratings
2. Test how well model representations enable paraphrase judgments

- In both cases, introduce controlled tests that remove cues based on amount of word overlap

Similarity correlations

ordinary citizen vs *average person*

large country vs *ordinary citizen*

arms control vs *control arms*

human similarity ratings

model representation similarities

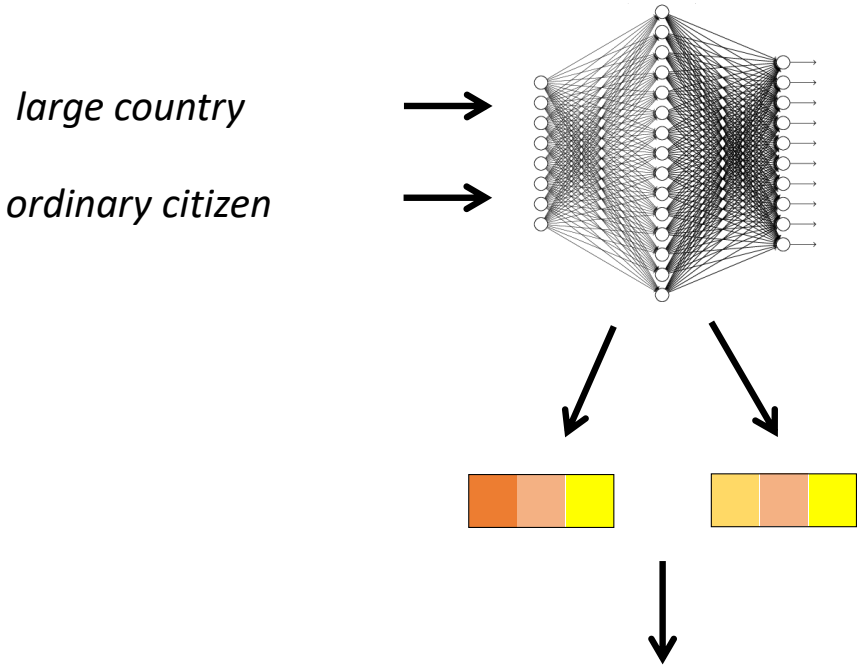
Similarity correlations

Normal Examples	
Source Phrase	Target Phrase & Score
average person	ordinary citizen (0.724)
	person average (0.518)
	country (0.255)
AB-BA Examples	
Source Phrase	Target Phrase & Score
law school	school law (0.382)
adult female	female adult (0.812)
arms control	control arms (0.473)

BiRD dataset (Asaadi et al., 2019)

Paraphrase classification

Do the representations have information to support a paraphrase judgment?



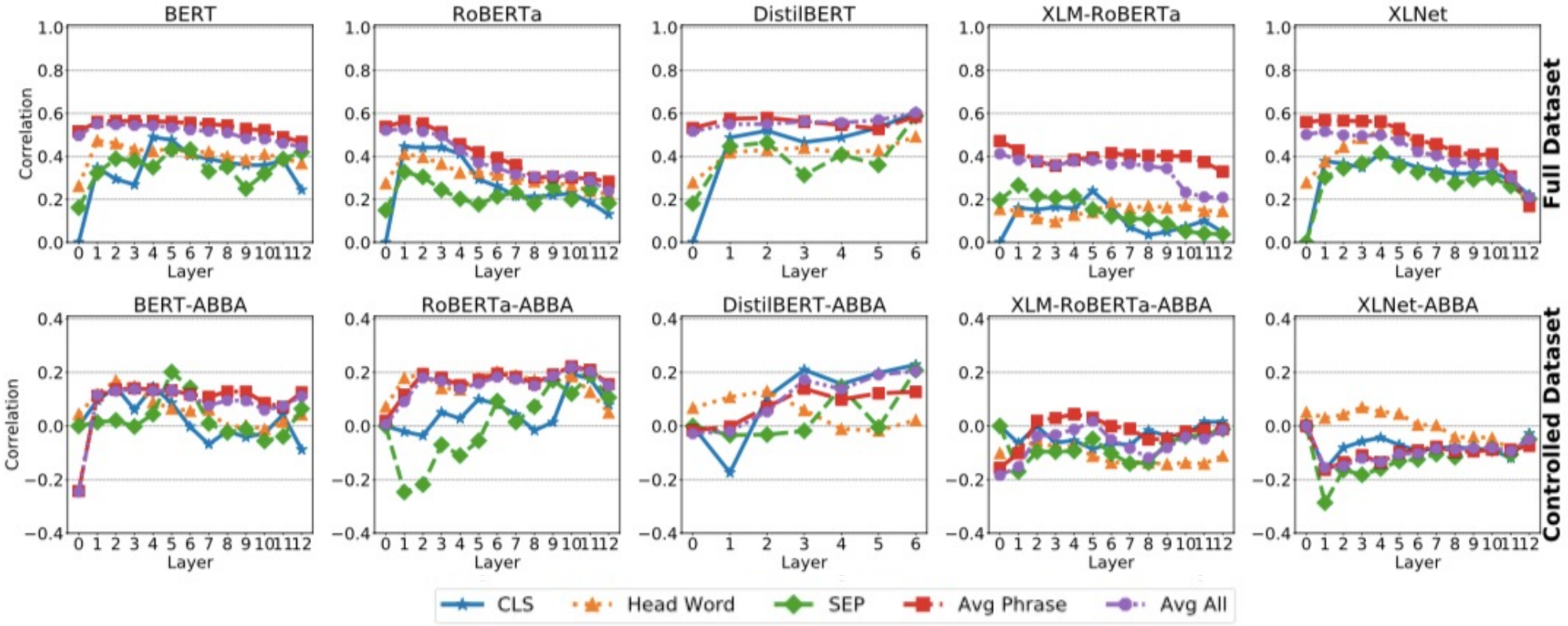
Do the two phrases *x* and *y* have the same meaning?

Paraphrase classification

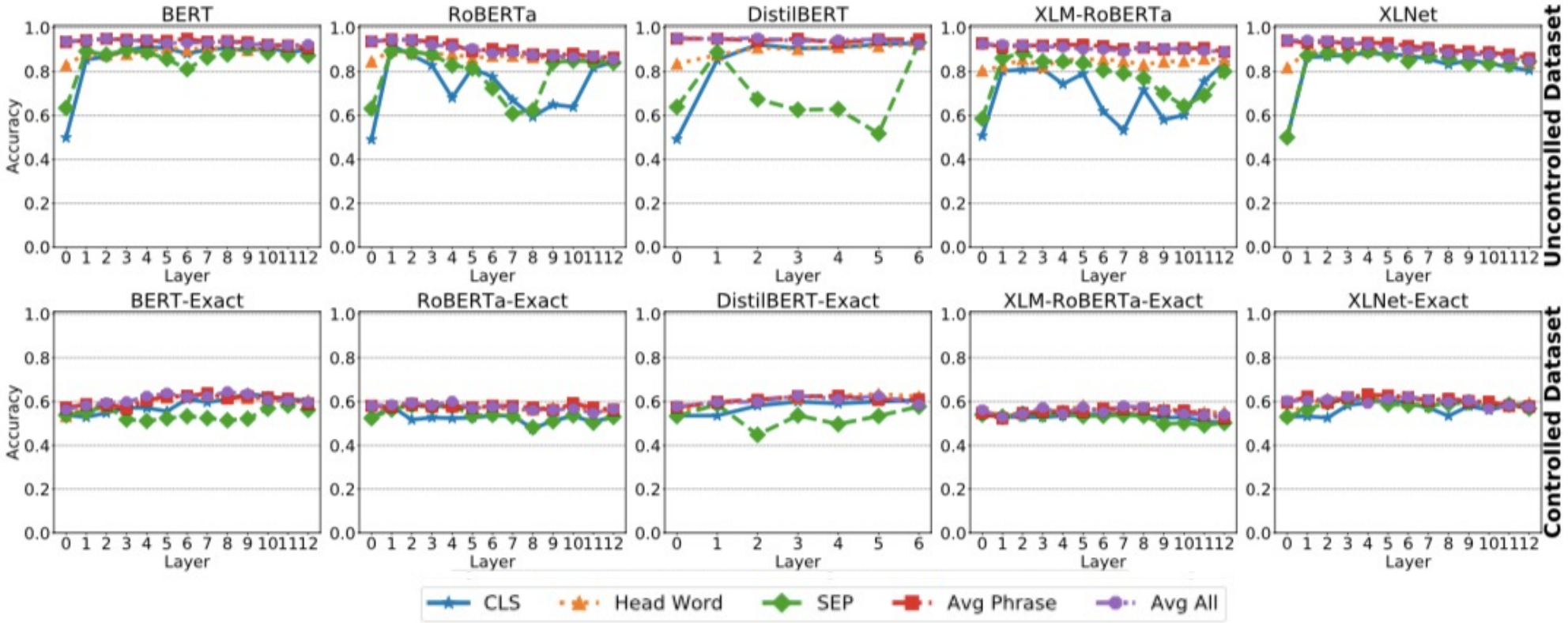
Normal Examples	
Source Phrase	Target Phrase
are crucial	is absolutely vital (pos)
	was a matter of concern (neg)
	is an essential part (pos)
	are exacerbating (neg)
Controlled Examples	
Source Phrase	Target Phrase
communication infrastructure	telecommunications infrastructure (pos)
	data infrastructure (neg)

PPDB 2.0 (Pavlick et al., 2015)

Similarity correlations



Paraphrase classification



Takeaways

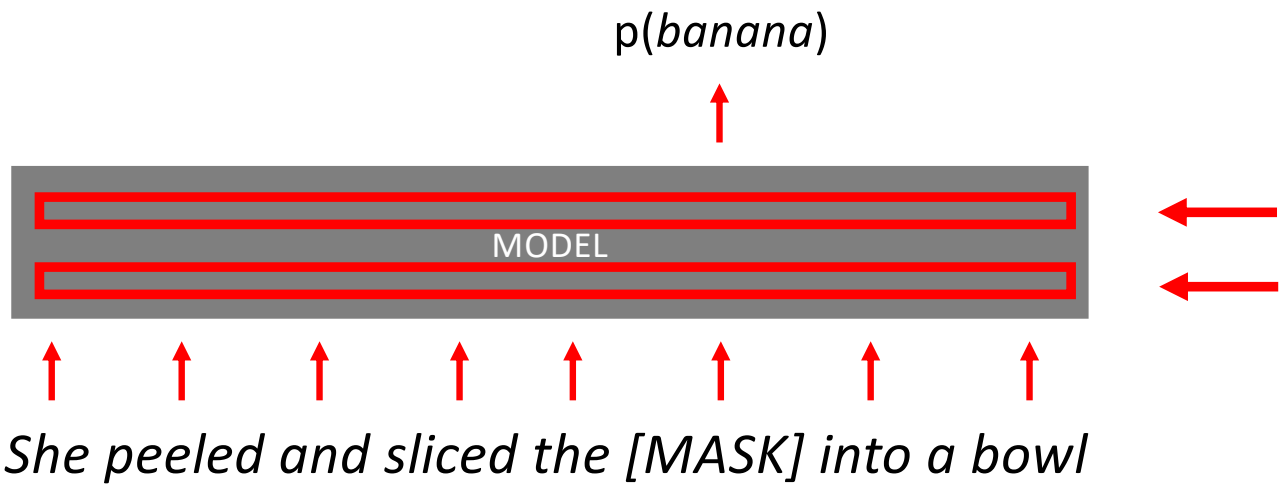
- With controlled test, no indication that representations are encoding compositional phrase meaning matching humans

- Appear mostly to be sensitive to information at lexical level

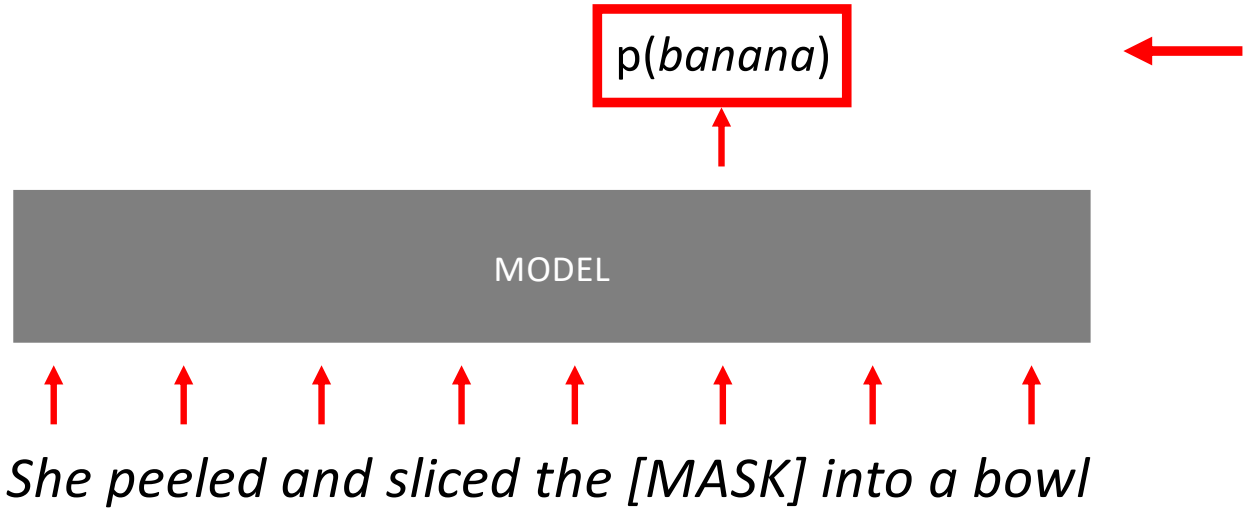
Fair test?

- Maybe models *are* robustly encoding compositional meaning, but we aren't checking the right representations, detection methods

Levels of analysis



Levels of analysis



Studying word prediction in context

- Everything models learn is for the sake of optimizing word/token prediction in context
- If models learn to encode meaning, will have been for prediction
- Maximally fair test for meaning

Key questions

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

Phrase composition

Psycholinguistic diagnostics

Information robustness

Key questions

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

Phrase composition

Psycholinguistic diagnostics

Information robustness

Studying word prediction in context

At the dinner party, I wondered why my mother wasn't eating her soup. Then I noticed that she didn't have a ____

Psycholinguistic tests

- Designed to study human responses to words in context – can test information sensitivity by examining word predictions directly
- Controlled to ask targeted questions about linguistic mechanisms underlying predictive responses
- Can be repurposed as diagnostics for models' linguistic sensitivities

These diagnostics

- Identify cases where

1. Humans show one pattern of conscious word predictions based on cloze task (fill-in-the-blank)
2. N400 (predictive brain response) shows different patterns, seeming to miss information that would inform word expectations

- Allows for targeting this challenging information and testing whether models manage to represent and use it for predictions
- Also allows for comparison against a more heuristic human response

Psycholinguistic diagnostics

- CPRAG-102: commonsense/pragmatic inference
- ROLE-88: event knowledge and semantic roles
- NEG-136: negation

Psycholinguistic diagnostics

- CPRAG-102: commonsense/pragmatic inference
- ROLE-88: event knowledge and semantic roles
- NEG-136: negation

Three measures (per diagnostic)

At the dinner party, I wondered why my mother wasn't eating her soup. Then I noticed that she didn't have a ____

1. Prediction accuracy — is "spoon" in top k model predictions?
2. Sensitivity tests — does model assign higher probability to "spoon" than to "knife"/"bowl"?
3. Qualitative analysis — what do models' top predictions tell us about the information they have access to?

Experiments

Case study: BERT model

Tested two variants

- $\text{BERT}_{\text{BASE}}$
- $\text{BERT}_{\text{LARGE}}$

Experiments

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of [MASK] . [SEP]

Experiments

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of [MASK] . [SEP]

Extract BERT word probabilities on [MASK] token, as in pre-training

Diagnostics and results

- CPRAG-102: commonsense/pragmatic inference
- ROLE-88: event knowledge and semantic roles
- NEG-136: negation

Diagnostics and results

- **CPRAG-102: commonsense/pragmatic inference**
- ROLE-88: event knowledge and semantic roles
- NEG-136: negation

CPRAG-102: commonsense/pragmatic inference

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that ____

- Need commonsense/pragmatic inference to infer situation and relation of first vs second sentence

Original study: Federmeier & Kutas (1999)

Prediction accuracy test

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of [MASK]

football in top k BERT predictions ?

Results: CPRAG accuracy test

	Orig
$\text{BERT}_{\text{BASE}}$ $k = 1$	23.5
$\text{BERT}_{\text{LARGE}}$ $k = 1$	35.3
$\text{BERT}_{\text{BASE}}$ $k = 5$	52.9
$\text{BERT}_{\text{LARGE}}$ $k = 5$	52.9

CPRAG sensitivity test

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of [MASK]

football >

baseball and *monopoly* ?

Results: CPRAG sensitivity test

	Prefer good
$\text{BERT}_{\text{BASE}}$	73.5
$\text{BERT}_{\text{LARGE}}$	79.4

CPRAG qualitative analysis

Context	$\text{BERT}_{\text{LARGE}}$ predictions
Pablo wanted to cut the lumber he had bought to make some shelves. He asked his neighbor if he could borrow her ____	*car, house, room, truck, apartment*
The snow had piled up on the drive so high that they couldn't get the car out. When Albert woke up, his father handed him a ____	*note, letter, gun, blanket, newspaper*
At the zoo, my sister asked if they painted the black and white stripes on the animal. I explained to her that they were natural features of a ____	*cat, person, human, bird, species*

Diagnostics and results

- **CPRAG-102: commonsense/pragmatic inference**
- ROLE-88: event knowledge and semantic roles
- NEG-136: negation

Diagnostics and results

- CPRAG-102: commonsense/pragmatic inference
- ROLE-88: event knowledge and semantic roles
- **NEG-136: negation**

NEG-136: negation

A robin is a ____

A robin is not a ____

Original study: Fischler et al. (1983)

NEG accuracy test

- This test doesn’t make sense in negated contexts, so accuracy is tested only on affirmative contexts

- Accurate predictions here require access to hypernym information

NEG accuracy test

A robin is a [MASK]

bird in top k BERT predictions ?

A robin is [illegible] a [MASK]

Results: NEG accuracy test

	Accuracy
$\text{BERT}_{\text{BASE}}$ $k = 1$	38.9
$\text{BERT}_{\text{LARGE}}$ $k = 1$	44.4
$\text{BERT}_{\text{BASE}}$ $k = 5$	100
$\text{BERT}_{\text{LARGE}}$ $k = 5$	100

NEG sensitivity test

- This is where the test of negation comes in
- Can BERT prefer true continuations to false continuations, with and without negation?

NEG sensitivity test

A robin is a [MASK]

bird > *tree* ?

A robin is not a [MASK]

tree > *bird* ?

Results: NEG sensitivity test

	Affirmative	Negative
$\text{BERT}_{\text{BASE}}$	100	0.0
$\text{BERT}_{\text{LARGE}}$	100	0.0

NEG qualitative analysis

Context	$\text{BERT}_{\text{LARGE}}$ predictions
A robin is a ____	*bird, robin, person, hunter, pigeon*
A daisy is a ____	*daisy, rose, flower, berry, tree*
A hammer is a ____	*hammer, tool, weapon, nail, device*
A hammer is an ____	*object, instrument, axe, implement, explosive*
A robin is not a ____	*robin, bird, penguin, man, fly*
A daisy is not a ____	*daisy, rose, flower, lily, cherry*
A hammer is not a ____	*hammer, weapon, tool, gun, rock*
A hammer is not an ____	*object, instrument, axe, animal, artifact*

Takeaways from BERT case study

- Model predictions do reflect a reasonable amount of linguistic information from context, but also show clear limitations
- Sometimes show signs of the patterns of insensitivity seen in the human N400 response
- In particular, clear lack of sensitivity to contextual impacts of negation

Discussion

- Not ultimately surprising that models trained for word prediction would show this behavior with negation
- “A robin is not __” is not conducive to any precise word prediction
- Makes sense to fall back on most constraining information in the context, and make predictions accordingly
- But this is an example of heuristic predictive behaviors, rather than systematic reliance on meaning

Meaning and prediction

- Let's examine this more systematically
- In previous experiment negation lacked utility for prediction
- If we pit meaning cues that *are* helpful for prediction against tempting but irrelevant superficial cues, which will drive predictions?

Key questions

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

Phrase composition

Psycholinguistic diagnostics

Information robustness

Key questions

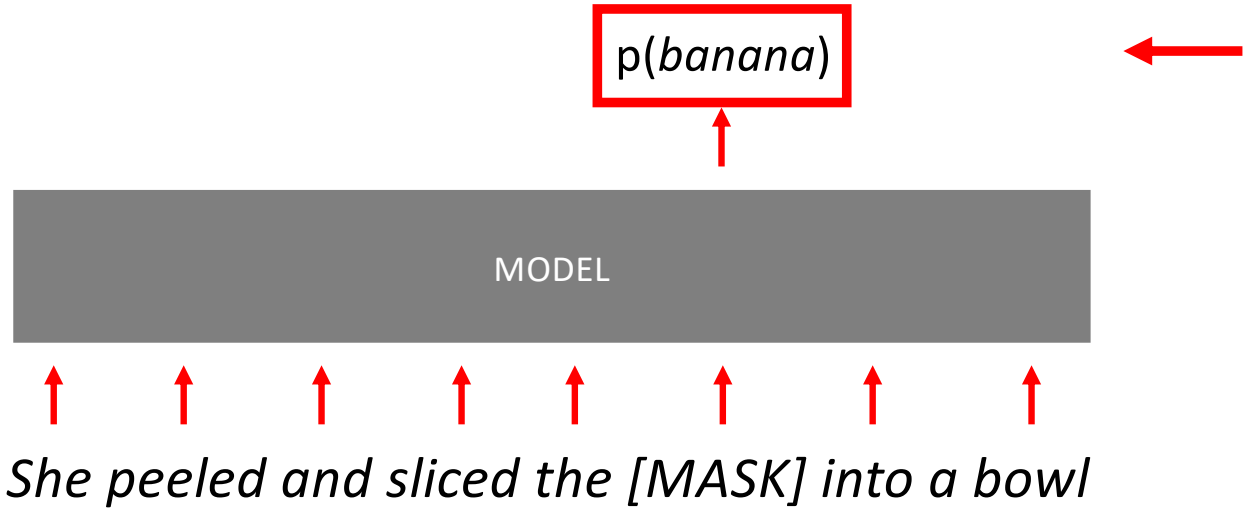
To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

Phrase composition

Psycholinguistic diagnostics

Information robustness

Levels of analysis



Extracting meaning from context

Sebastian lives in France. The capital of Sebastian's country is ___

Extracting meaning from context

Lalchand Pandia

- Start from simple base contexts

Sebastian lives in France. The capital of Sebastian's country is ___

- Correct predictions in base context could have various explanations
- What if "France" isn't the only country mentioned in context?

Sebastian lives in France, Rowan lives in Indonesia, and Daniel lives in Chile. The capital of Sebastian's country is ___

Extracting meaning from context

- Insert and systematically manipulate irrelevant “attractors” in context
- Attractors are words not relevant for the correct prediction, but may influence model outputs if relying on superficial cues

Semantically-related attractors

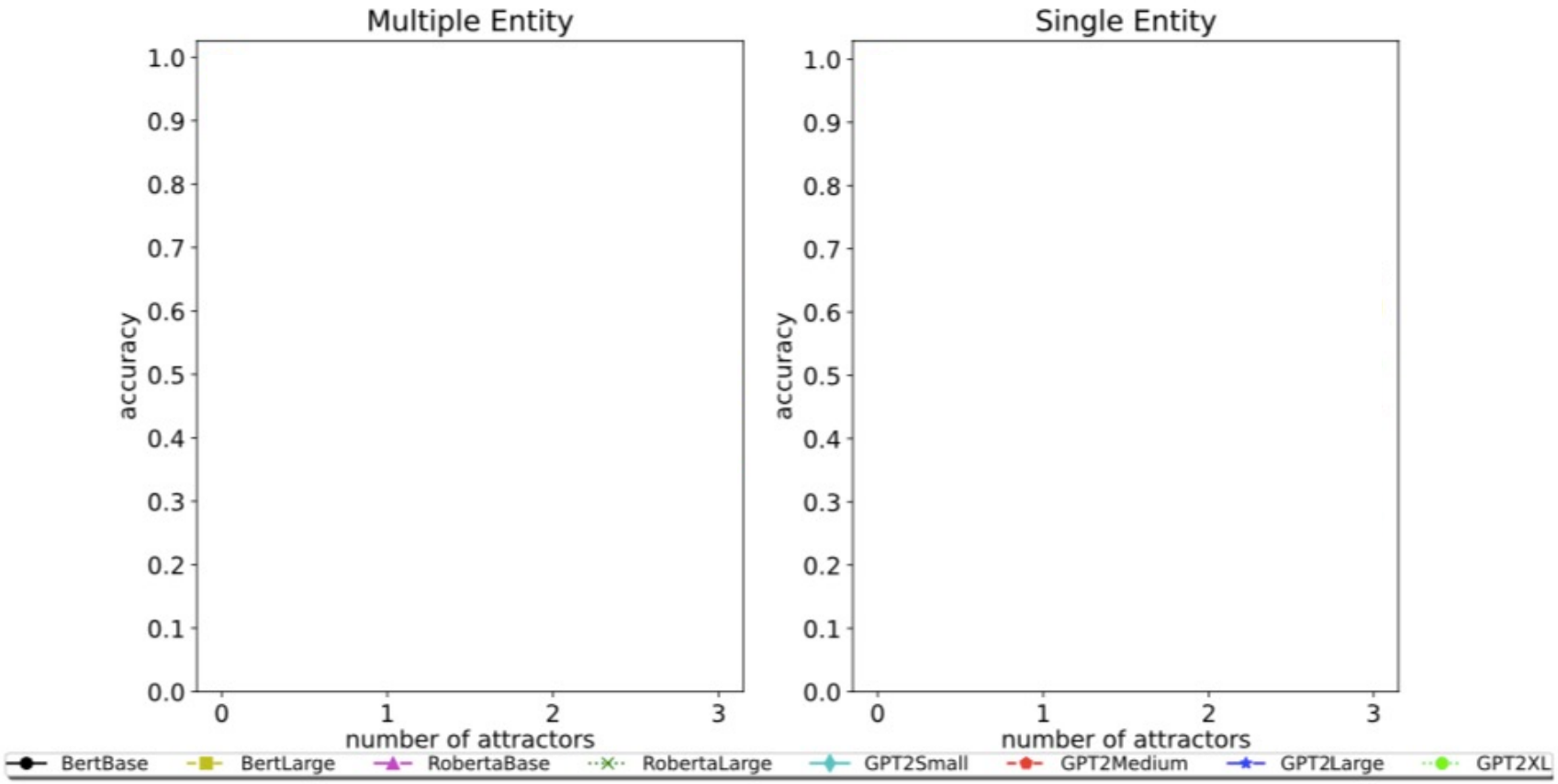
- Multiple-entity:

Sebastian *lives in France,* ***Rowan*** *lives in Indonesia, and* ***Daniel*** *lives in Chile. The capital of* ***Sebastian's*** *country is ___*

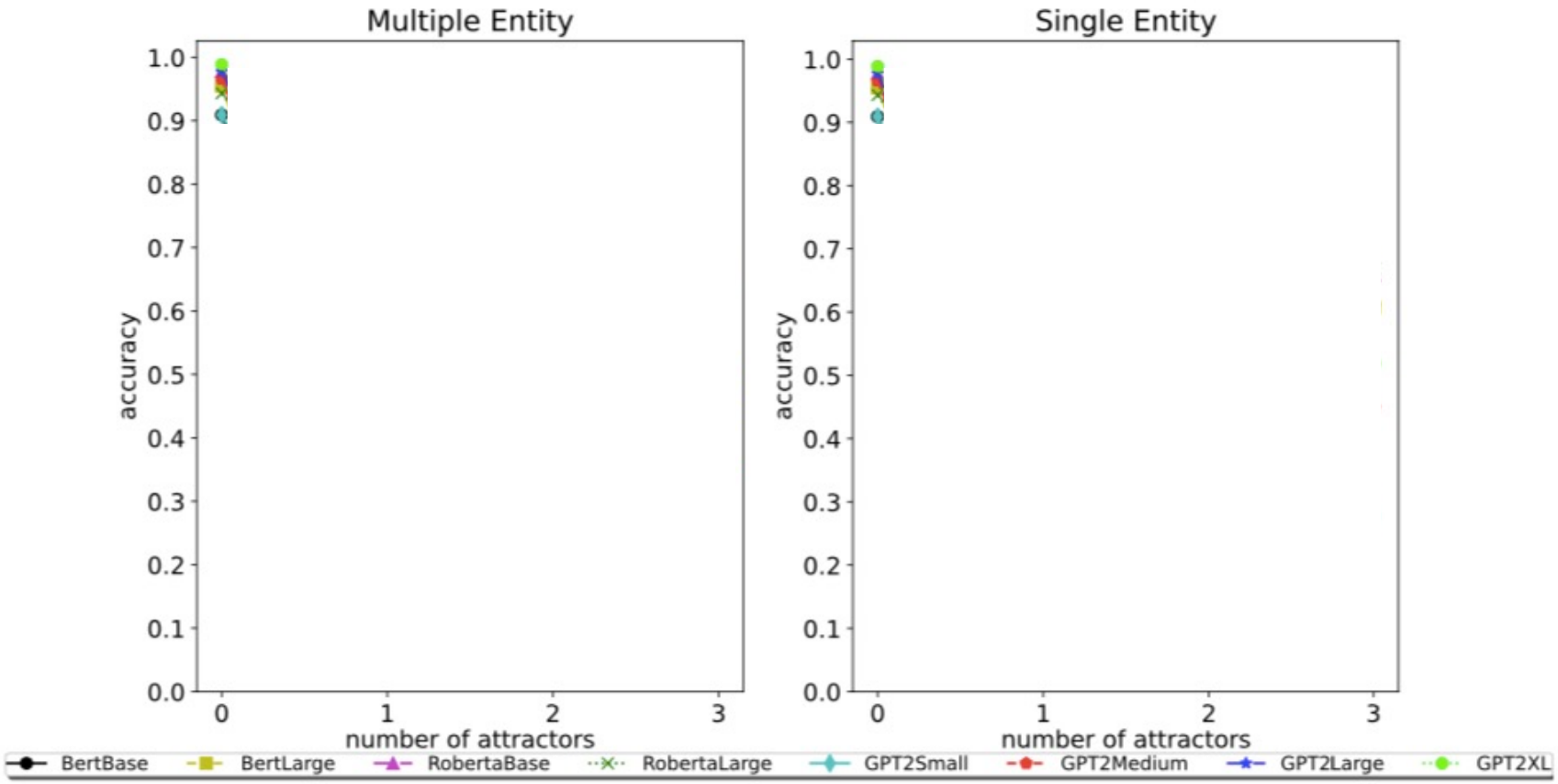
- Single-entity:

Sebastian *lives in France, and has visited Indonesia and Chile. The capital of* ***Sebastian's*** *country is ____*

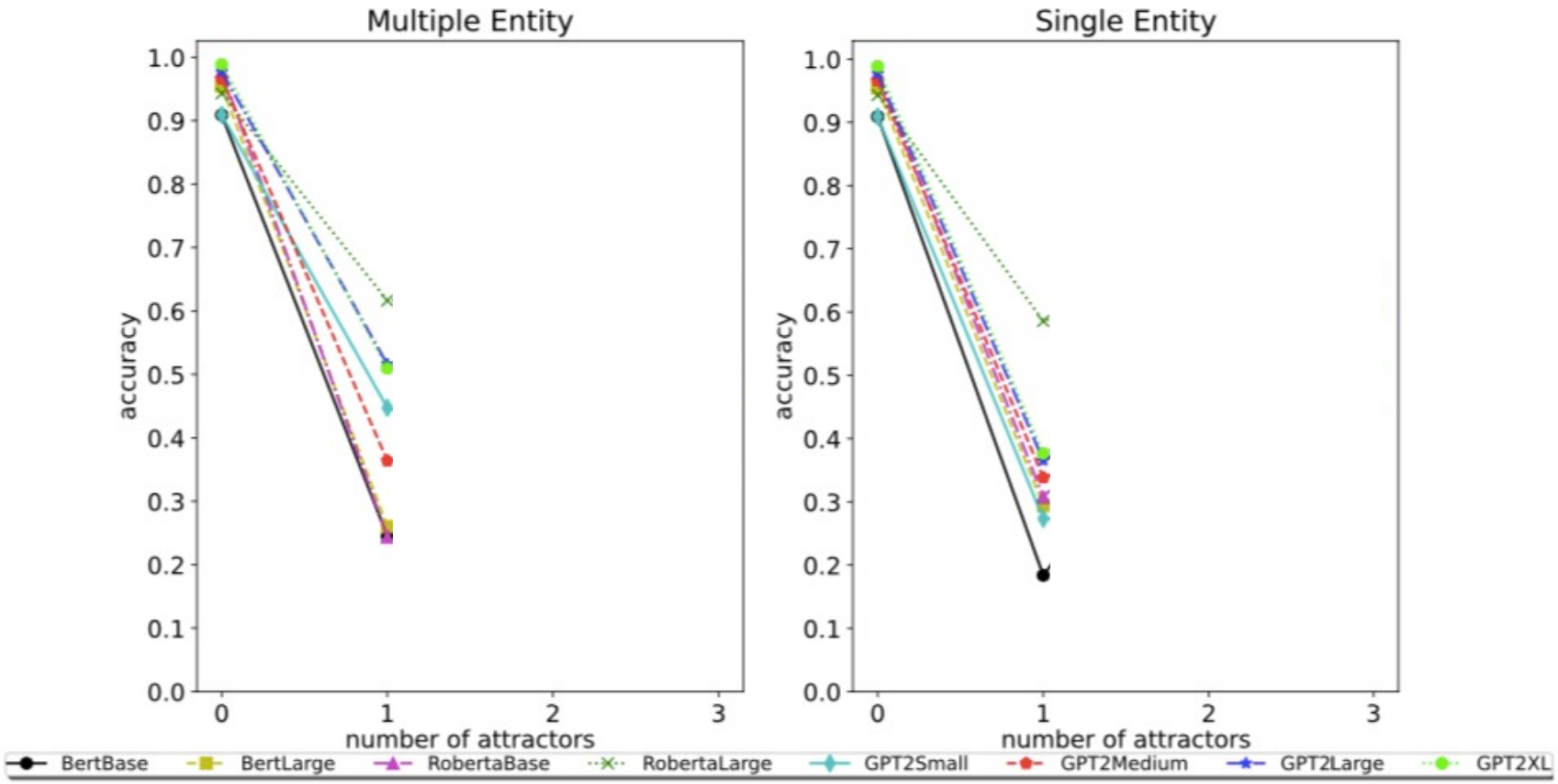
Accuracy (correct target prob > other words in semantic set)



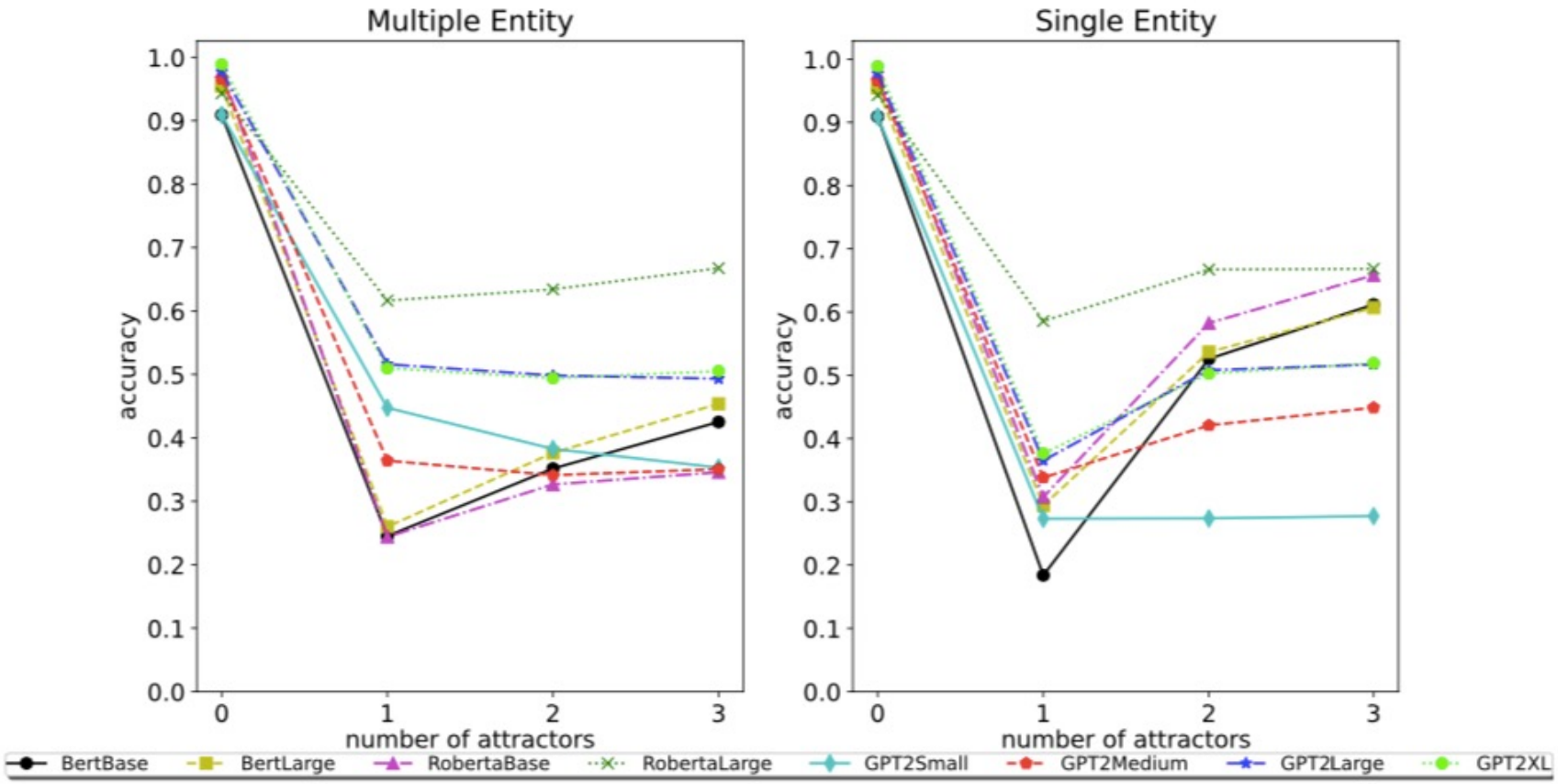
Accuracy (correct target prob > other words in semantic set)



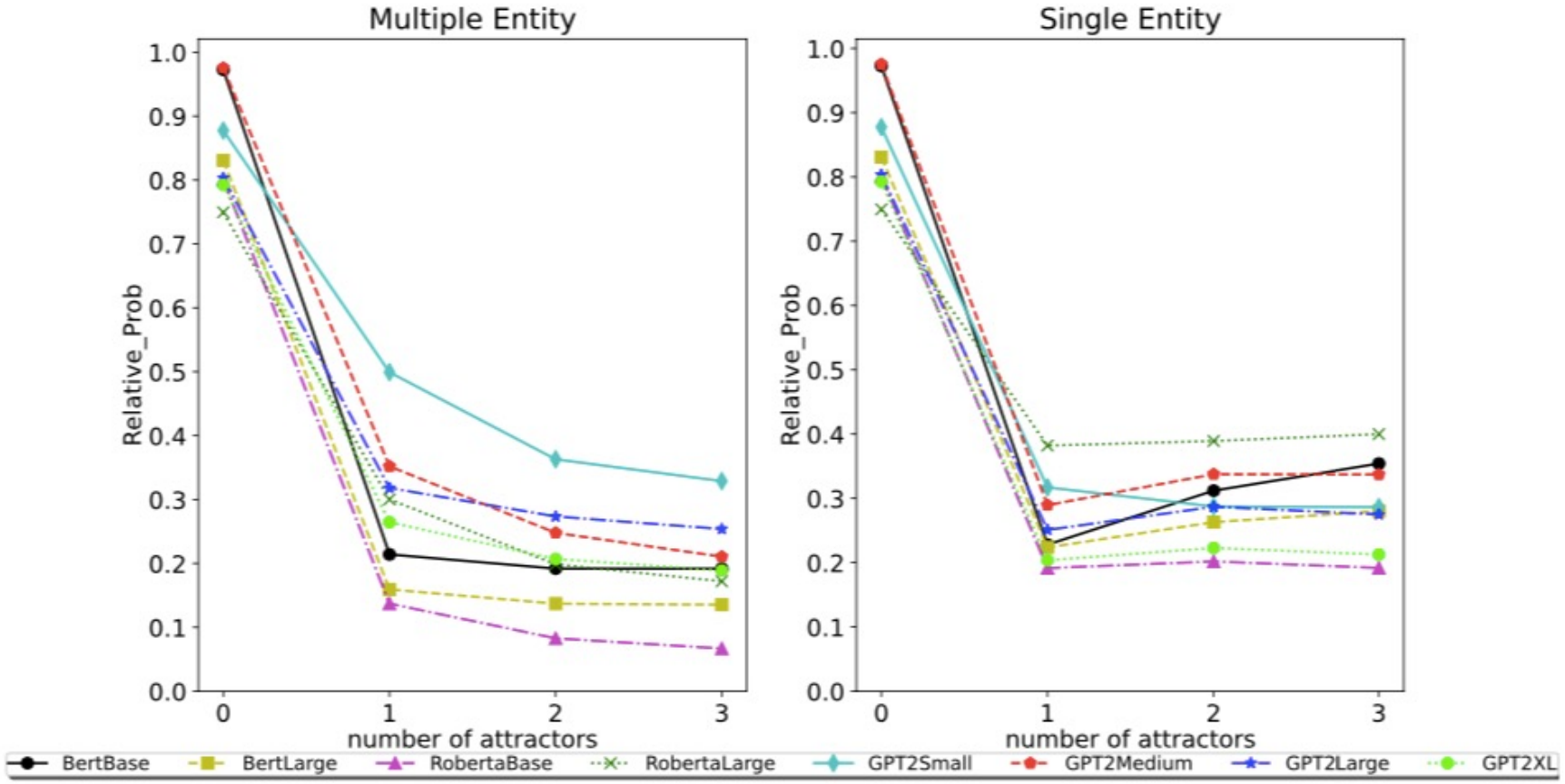
Accuracy (correct target prob > other words in semantic set)



Accuracy (correct target prob > other words in semantic set)



Relative probability (attractor context prob/base context prob)



Semantically-unrelated attractors

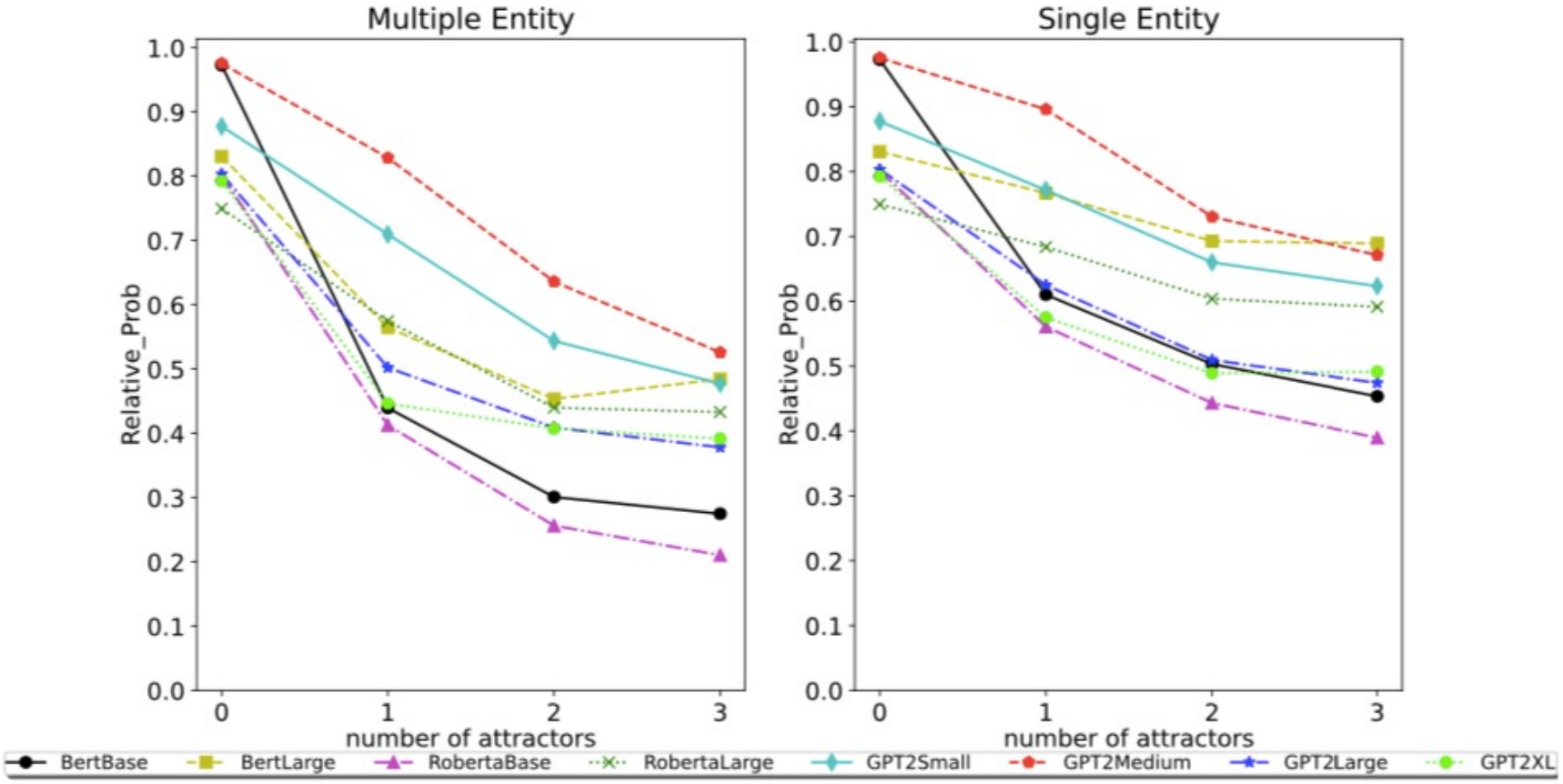
- Multiple-entity:

Sebastian lives in France, Rowan drives a car, and Daniel writes poetry. The capital of Sebastian's country is ___

- Single-entity:

Sebastian lives in France, drives a car, and writes poetry. The capital of Sebastian's country is ____

Relative probability (attractor context prob/base context prob)

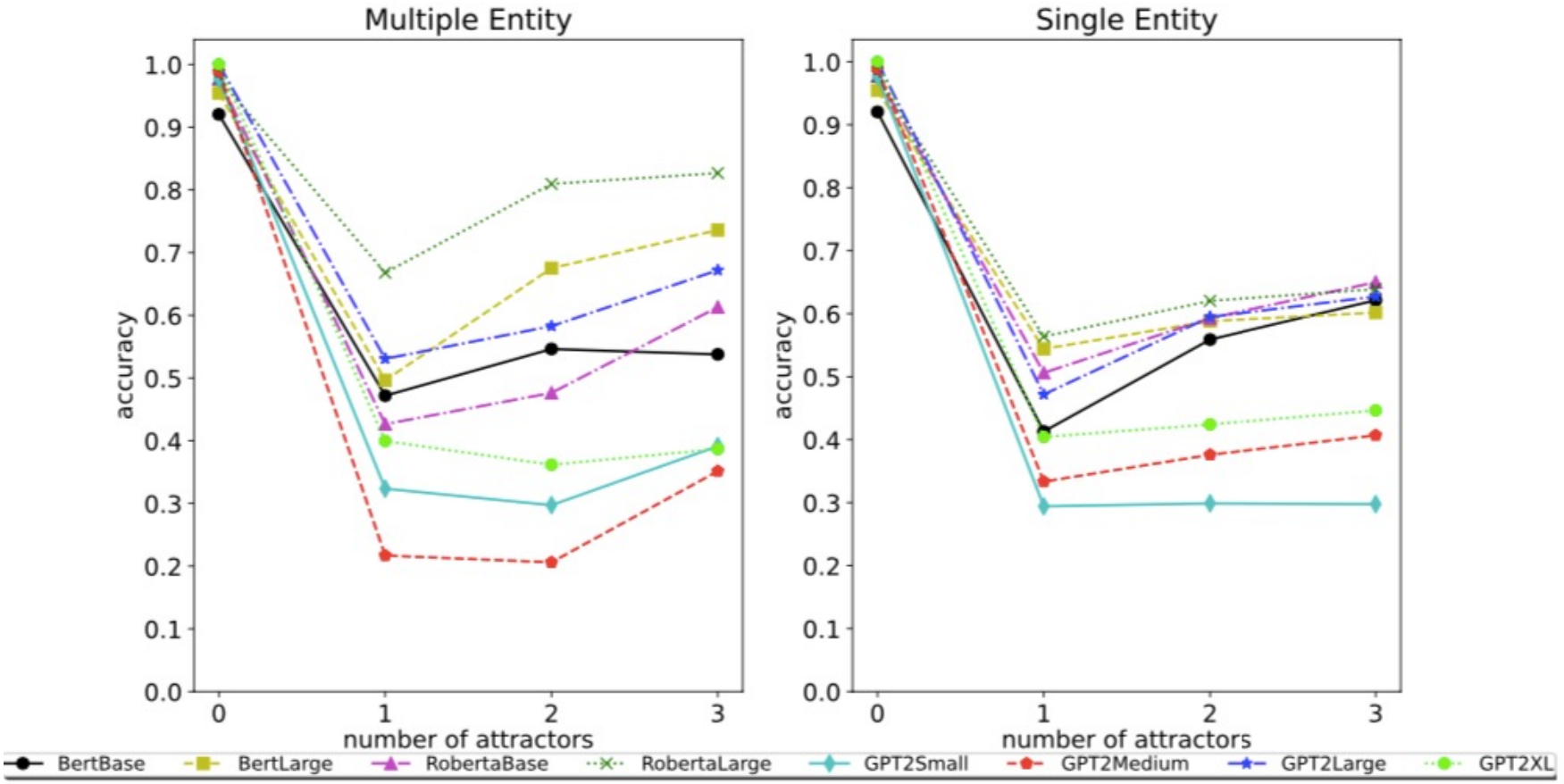


Varying attractor position

- Key fact is most recent:

Daniel *knows that Jack lives in Beijing and he himself* ***lives in Chile****. The capital of Daniel's country is ____*

Accuracy (correct target prob > other words in semantic set)

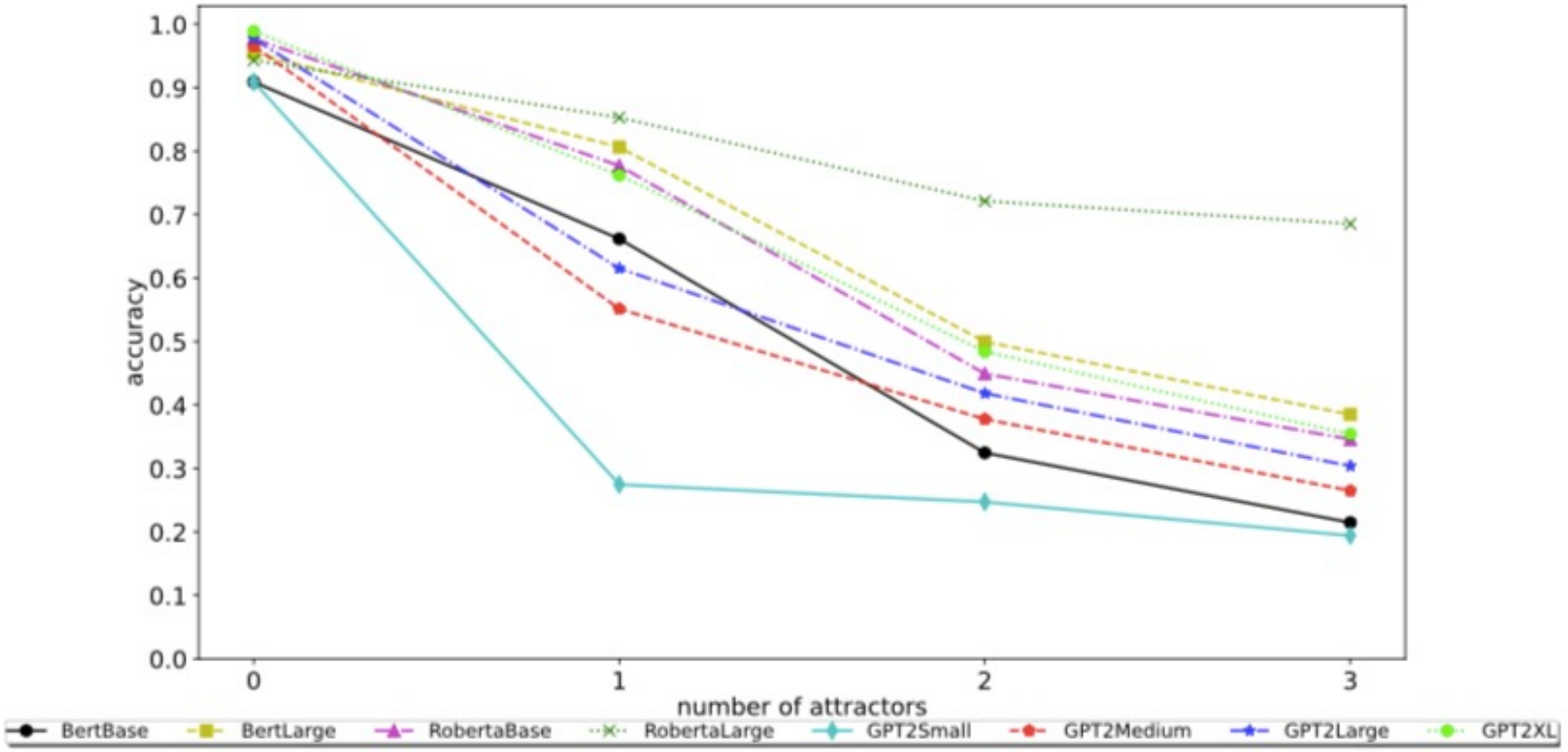


Varying attractor position

- Key entity not first-mentioned entity (attractors often earlier than key entity):

*Sebastian lives in France, and **Rowan** lives in Indonesia. The capital of **Rowan's** country is ____*

Accuracy (correct target prob > other words in semantic set)



Takeaways

- Clear, substantial impact of irrelevant attractors on model predictions
- Attractors have most dramatic impact when
  - Semantically related to words involved in prediction
  - Occurring after first mention of key entity in prediction
- Suggests use of coarse-grained semantic similarity, relative word position heuristics for word prediction

Further support for distractibility

Kanishka Misra

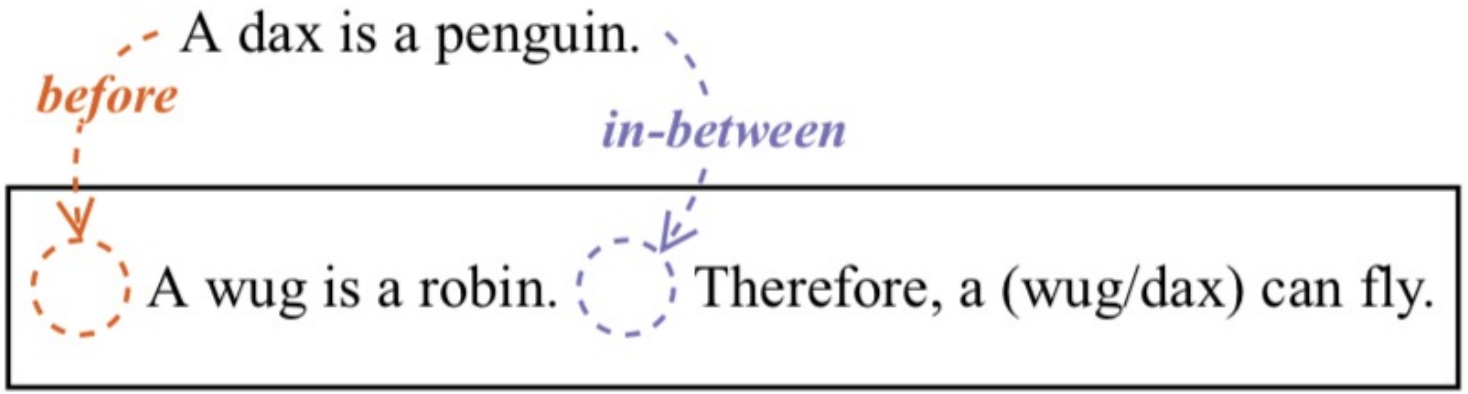
- Testing property knowledge

*A **robin / penguin** can fly.*

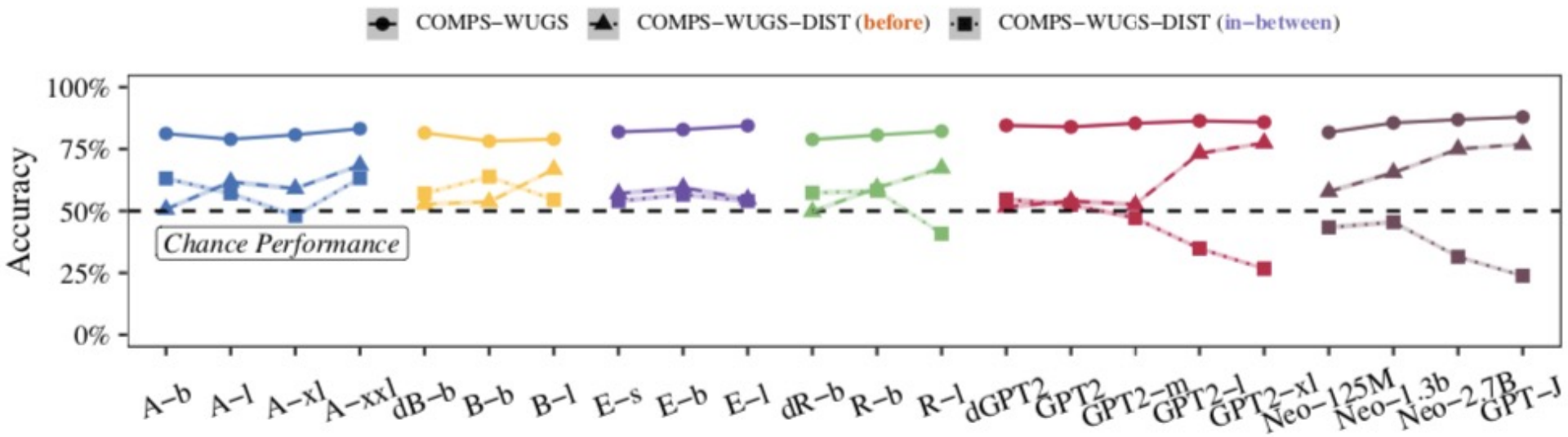
- Testing property inheritance

*A wug is a **robin / penguin**. Therefore, a wug can fly.*

Further support for distractibility



Further support for distractibility



Further support for distractibility

- Irrelevant attractors once again have substantial impact on model performance, bringing accuracy roughly to chance
- Inserting semantically related words has large effect on accuracy while inclusion of intervening words like “wug” does not
- For larger left-to-right models, distractor is particularly damaging when occurring more recently than relevant content
- Suggests semantic similarity and recency heuristics

Discussion

- Again, we can imagine that these are heuristics that in many contexts will be reliable for prediction
- Not surprising for models optimized on word prediction to learn to rely on these heuristics
- But also gives no indication that what models encode is a robust meaning representation that can be queried for prediction purposes

Extracting meaning from context

Sebastian lives in France. The capital of Sebastian's country is ___

Discussion

- Supports and expands on negation result – even when meaning information does have utility for prediction, heuristics are winning

- Patterns observed here do bear resemblance to patterns seen in human real-time processing (e.g., semantic priming, N400)

Key questions

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

Phrase composition

Psycholinguistic diagnostics

Information robustness

Part one takeaways

- When we use controlled tests, we discover key limitations in models' capability to handle meaning systematically and robustly

- Behaviors that are unsurprising given models' training, but that reflect heuristic predictive strategies rather than robust compositional meaning understanding

Cognitive implications

- Should not treat these models as valid cognitive models with respect to processes of compositional meaning
- But may align with more heuristic predictive mechanisms in humans
- See some alignment with N400, semantic priming patterns
- Good reason to suspect that mechanisms in humans designed for prediction would develop similar strategies, statistical sensitivities

NLP implications

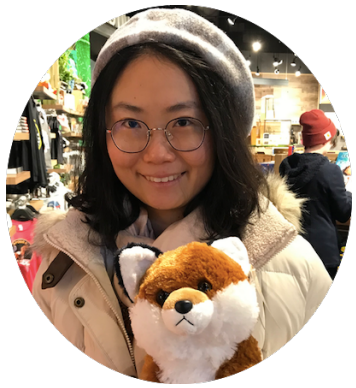
- Appearances of “understanding” should be taken with grain of salt
- Results suggest that even basic aspects of meaning above word level remain unsolved in these NLP models
- Some limitations may be intrinsic to learning based on language modeling (word prediction)

Key questions

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

How do meaning and prediction processes interact in the brain during real-time language processing?

Key questions



Jiaxuan Li

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

How do meaning and prediction processes interact in the brain during real-time language processing?

Language processing in real time

Language processing in real time

Each

Language processing in real time

Each morning

Language processing in real time

Each morning to

Language processing in real time

Each morning to wake

Language processing in real time

Each morning to wake up

Language processing in real time

Each morning to wake up I

Language processing in real time

Each morning to wake up I pour

Language processing in real time

Each morning to wake up I pour myself

Language processing in real time

Each morning to wake up I pour myself a

Language processing in real time

Each morning to wake up I pour myself a steaming

Language processing in real time

Each morning to wake up I pour myself a steaming cup

Language processing in real time

Each morning to wake up I pour myself a steaming cup of

Language processing in real time

Each morning to wake up I pour myself a steaming cup of salsa

Meaning and prediction

- Humans can extract meaning regardless of probability
- But we are also sensitive to probabilistic properties of inputs – likely increases processing efficiency
- How do meaning extraction and predictive processes interact in the brain to achieve rapid, robust language processing?

Semantic anomalies

Each morning to wake up I pour myself a steaming cup of salsa

I like my coffee with cream and rabbits

Measuring brain activity (EEG)

N400 and P600 components

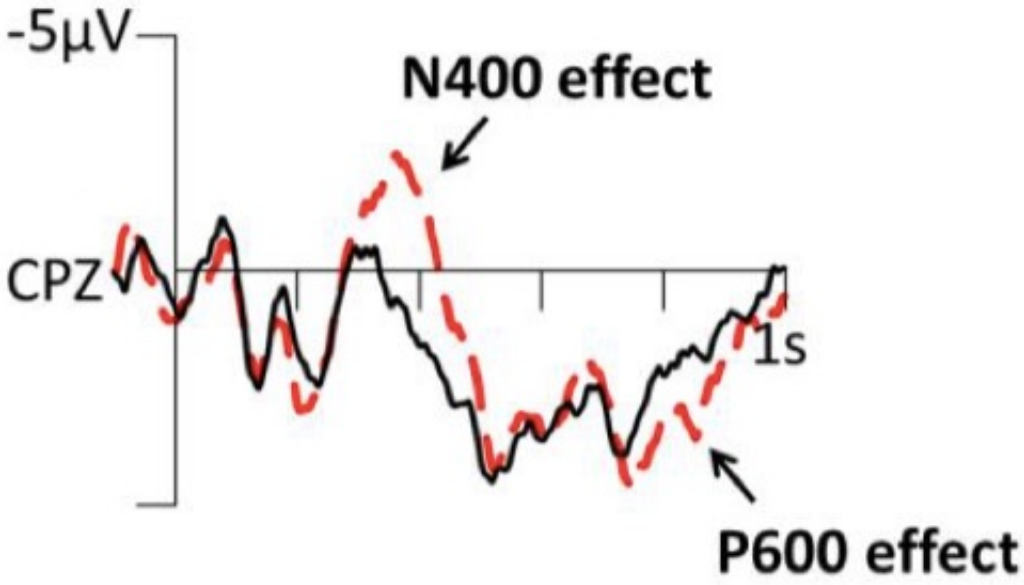
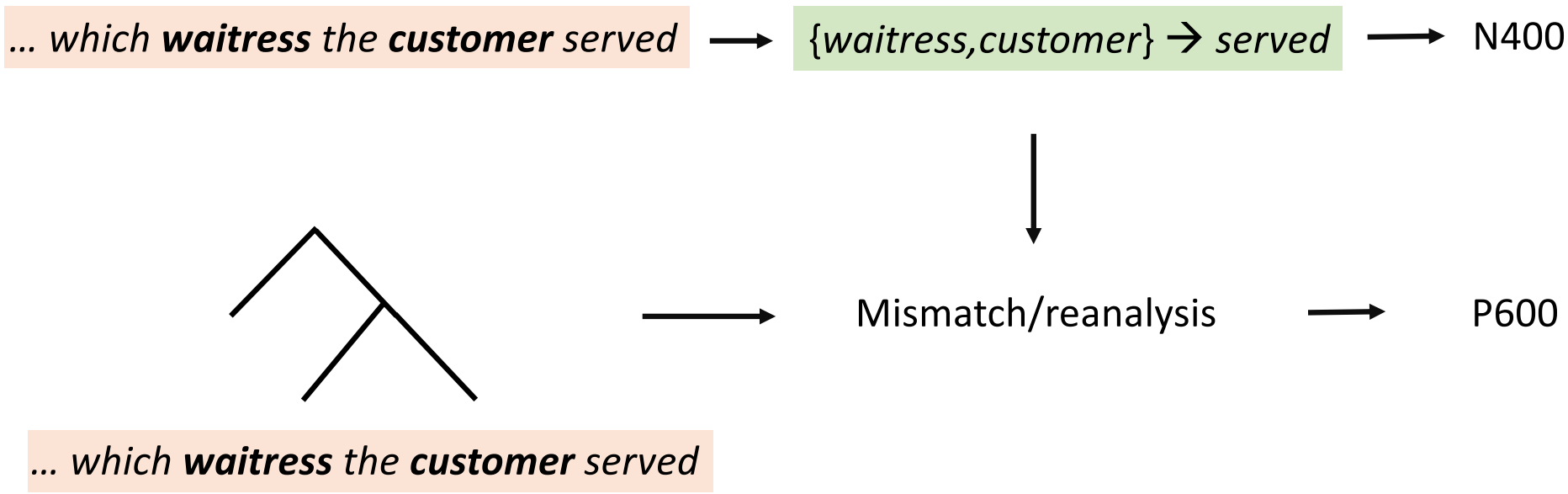


Image: Chow et al. (2015)

Heuristic interpretation theories



Heuristic interpretation theories

- Heuristic interpretation theories can predict some, but not all, of observed responses to semantic anomalies

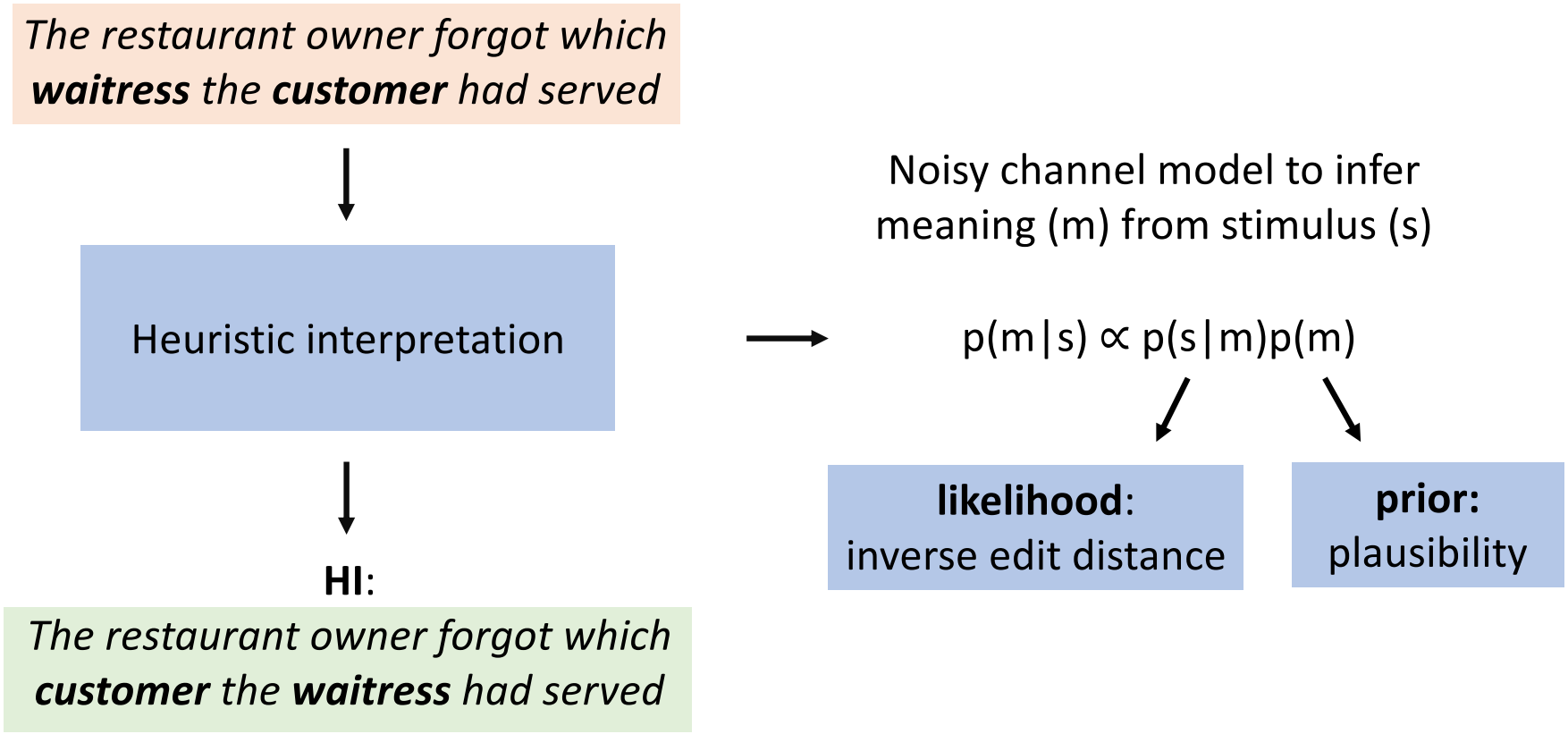
- Struggle with two particular observed patterns
  - Biphasic N400/P600 effects
  - Diverging N400 vs P600 patterns for same phenomenon in different experiments

Heuristic interpretation model

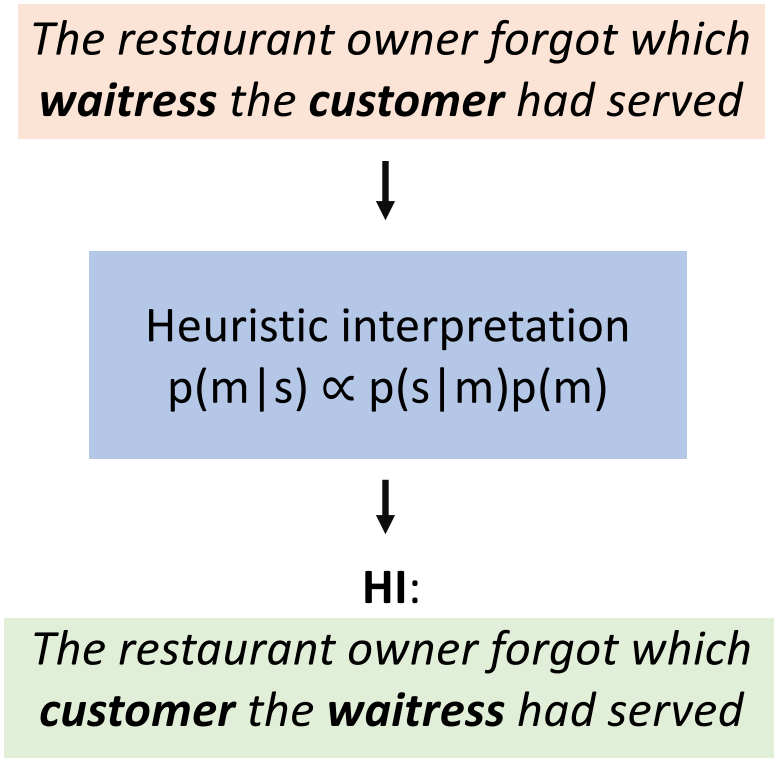
- We formalize the heuristic interpretation mechanism within a computational modeling framework

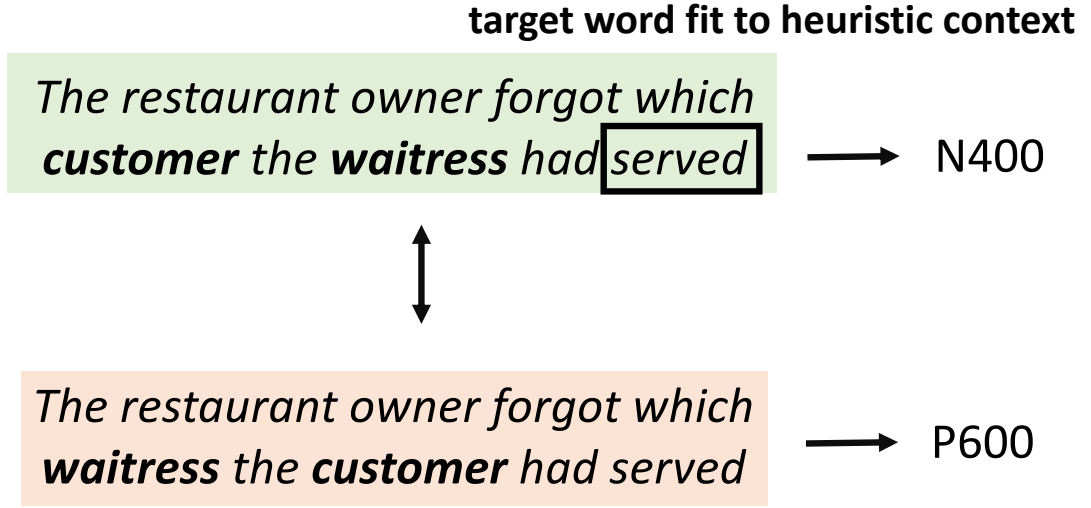
- We incorporate estimates from pre-trained LMs to capture probabilistic sensitivities in this heuristic processing mechanism

Heuristic interpretation model

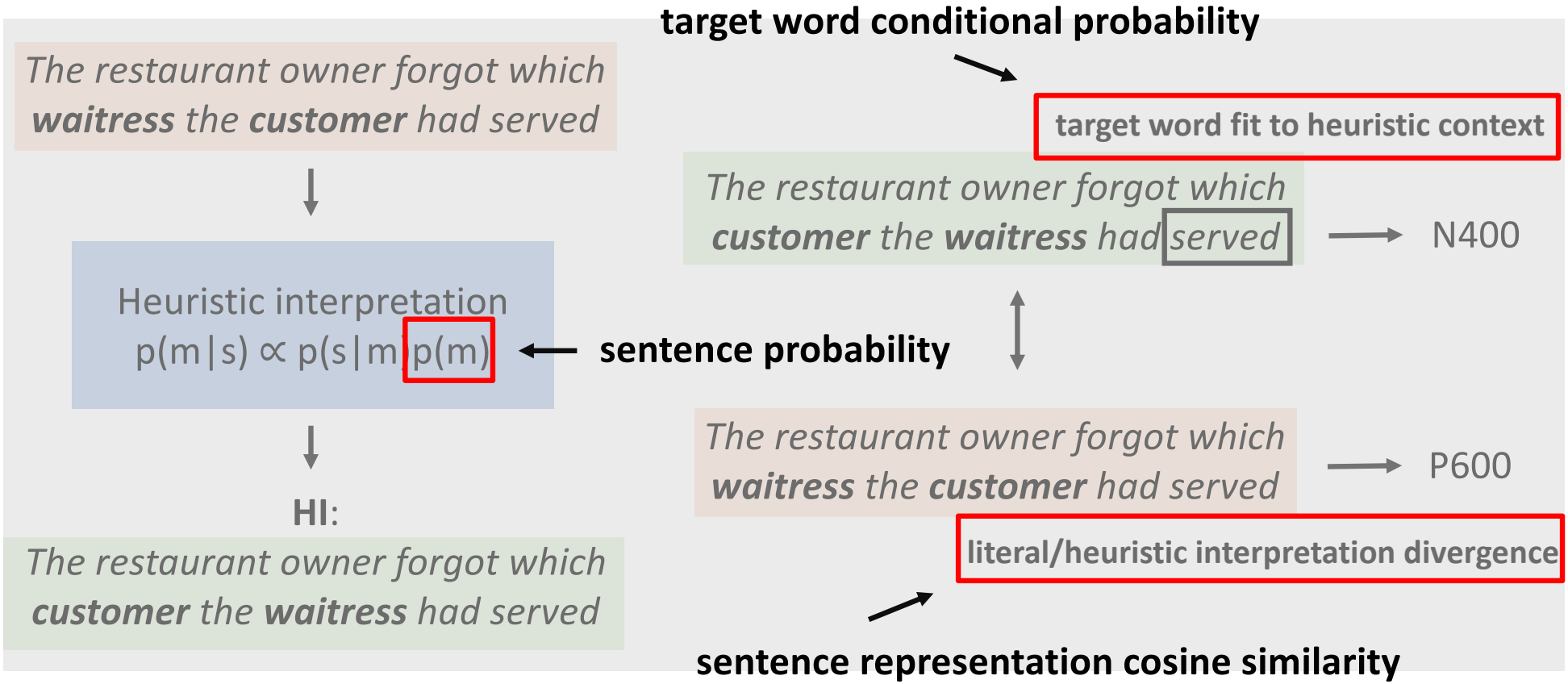


Heuristic interpretation model





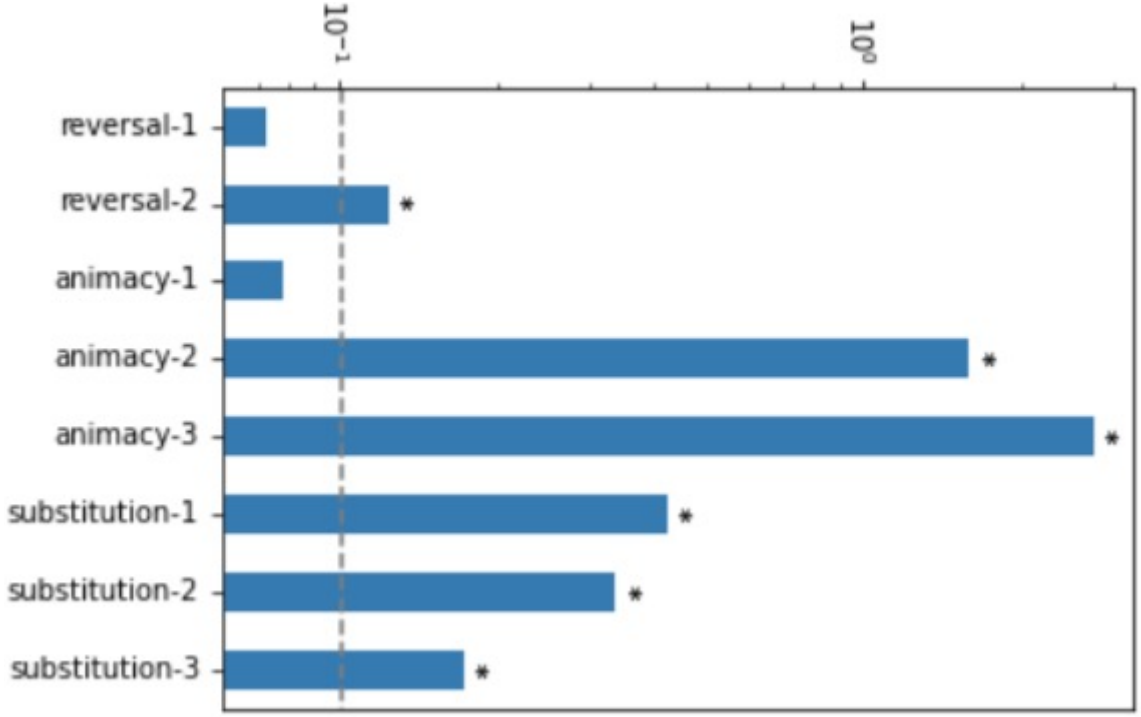
Estimates drawn from pre-trained LMs



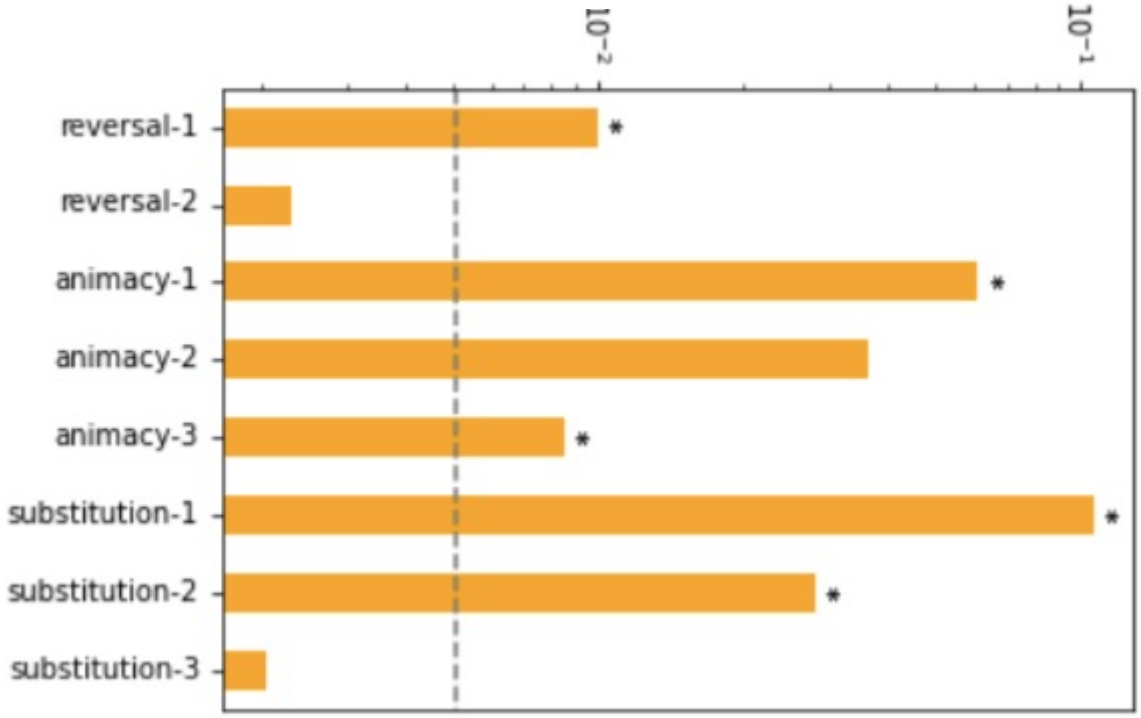
Model simulations and results

- Use model to simulate magnitude of N400 and P600 effects for wide range of target human studies
- Simulated effects closely align with presence/absence of statistically significant effects in humans
- Including challenging results not accounted for by heuristic interpretation theories as originally formulated

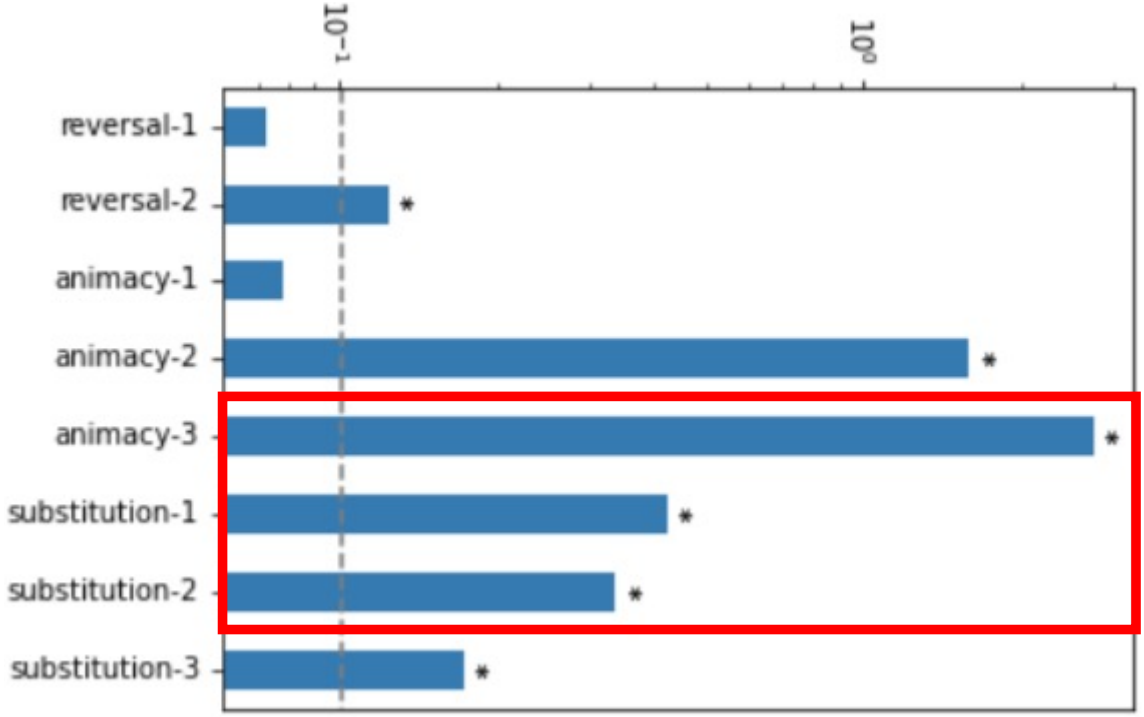
Simulated N400 effect



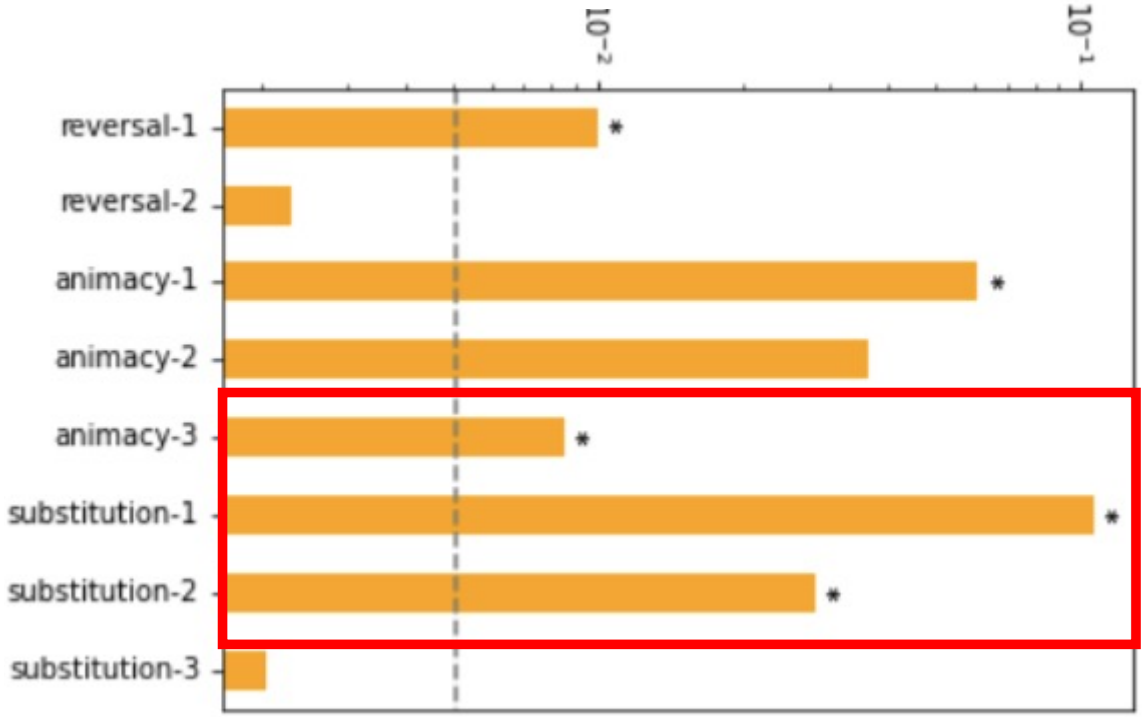
Simulated P600 effect



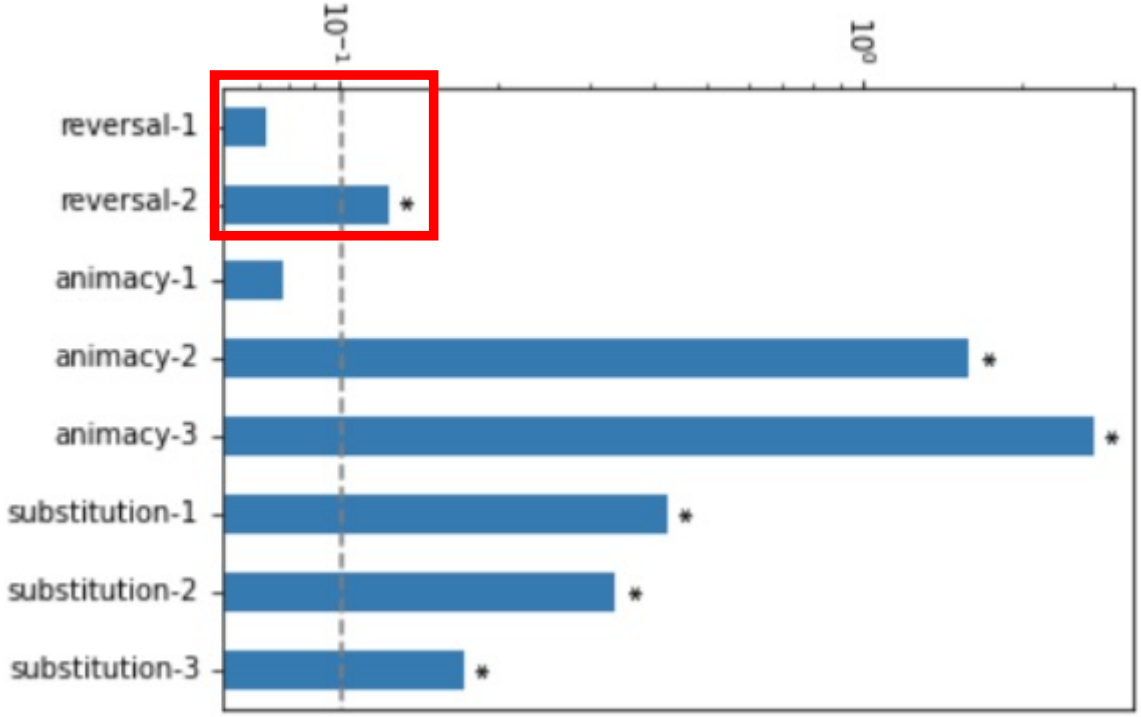
Simulated N400 effect



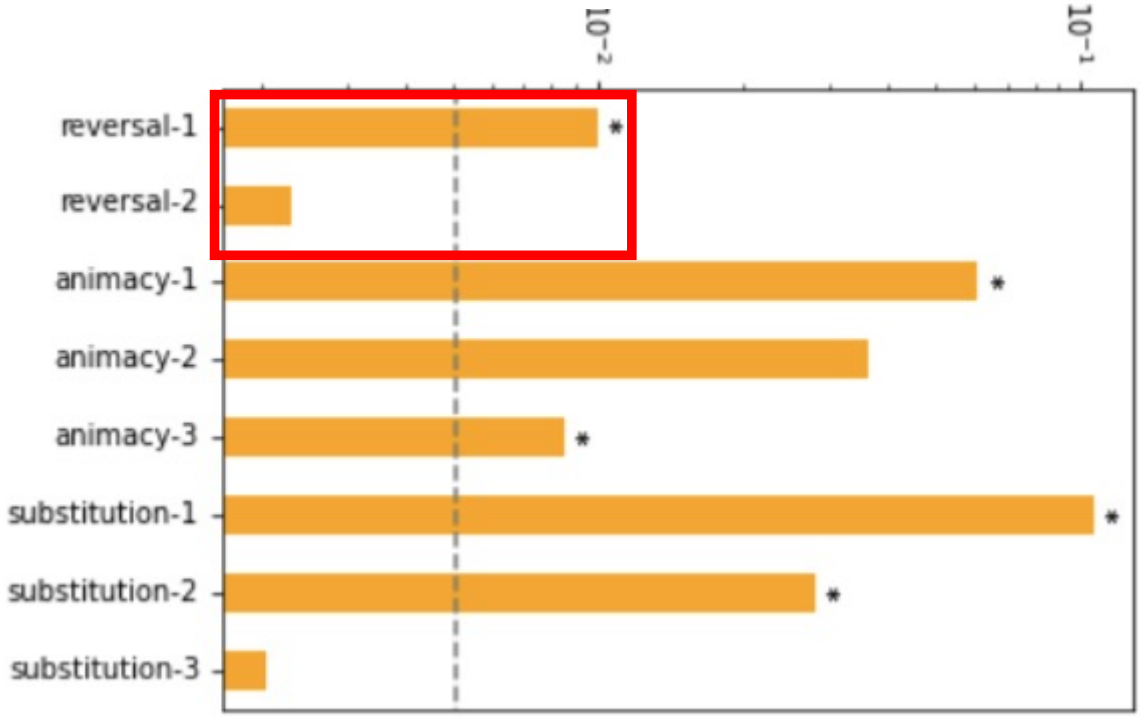
Simulated P600 effect



Simulated N400 effect



Simulated P600 effect



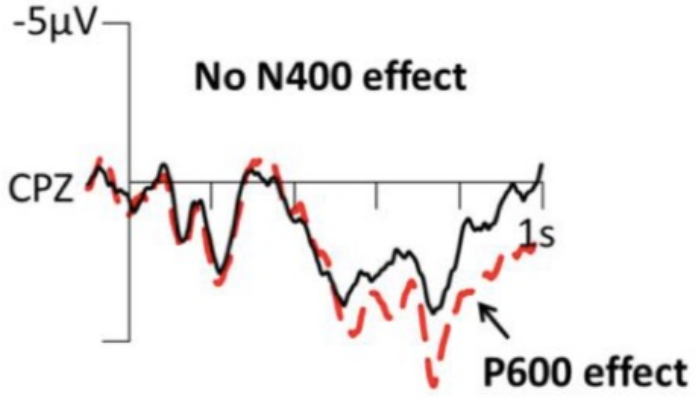
Analysis

- How does our model capture observed effects?

- Focus on two main challenges for existing theories:
  - Diverging results for same manipulation across experiments
  - Biphasic N400/P600 effects

Reversal-1 (Chow et al. 2015)

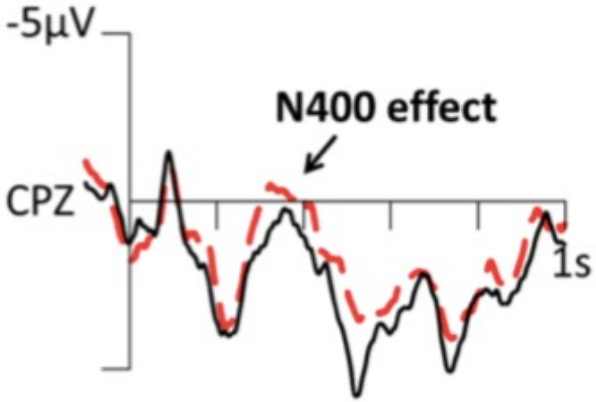
High cloze -- *The restaurant owner forgot which customer the waitress had served*
Low cloze -- *The restaurant owner forgot which waitress the customer had served*



Chow et al. (2015)

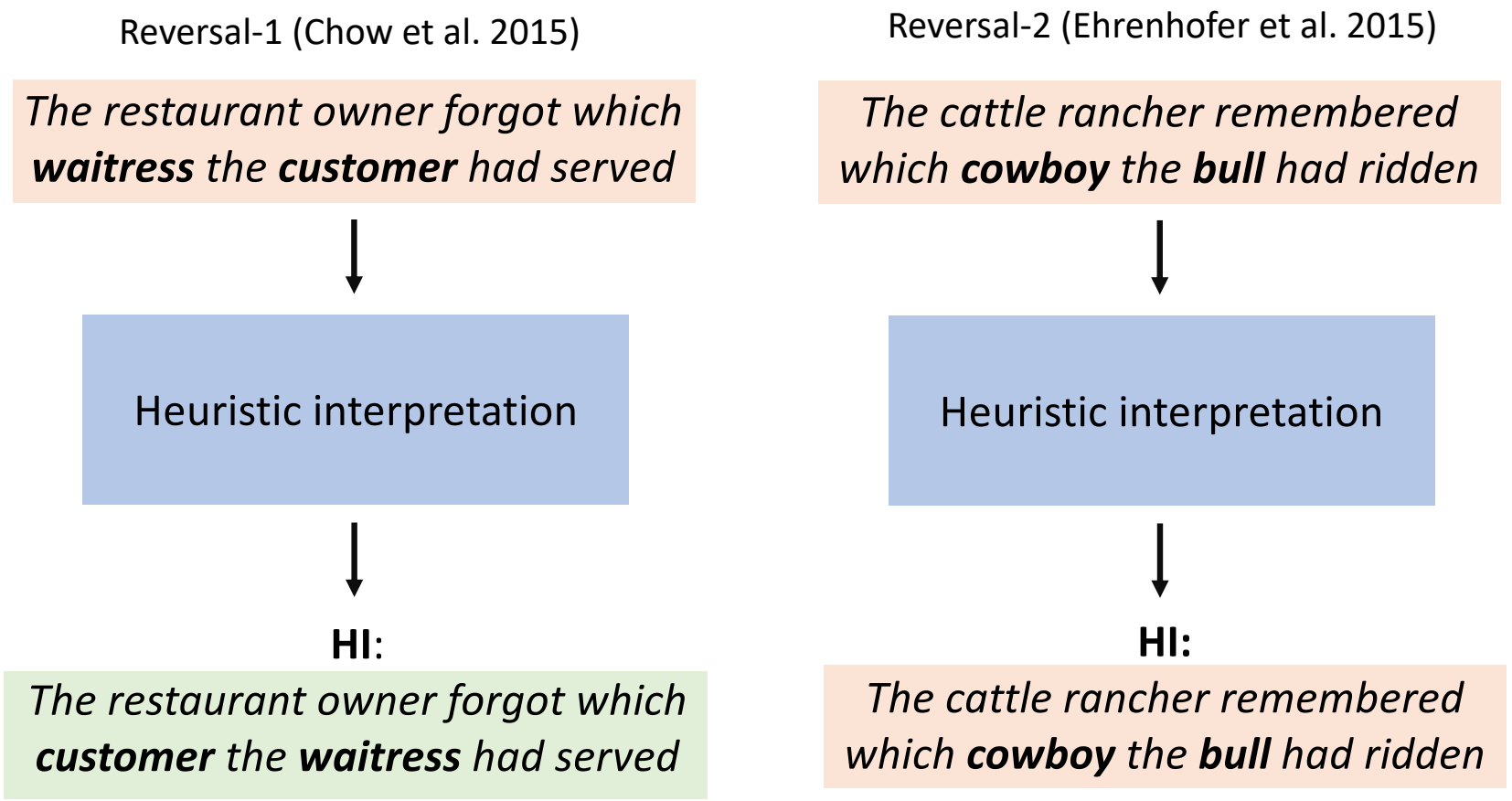
Reversal-2 (Ehrenhofer et al. 2015)

High cloze -- *The cattle rancher remembered which bull the cowboy had ridden*
Low cloze -- *The cattle rancher remembered which cowboy the bull had ridden*

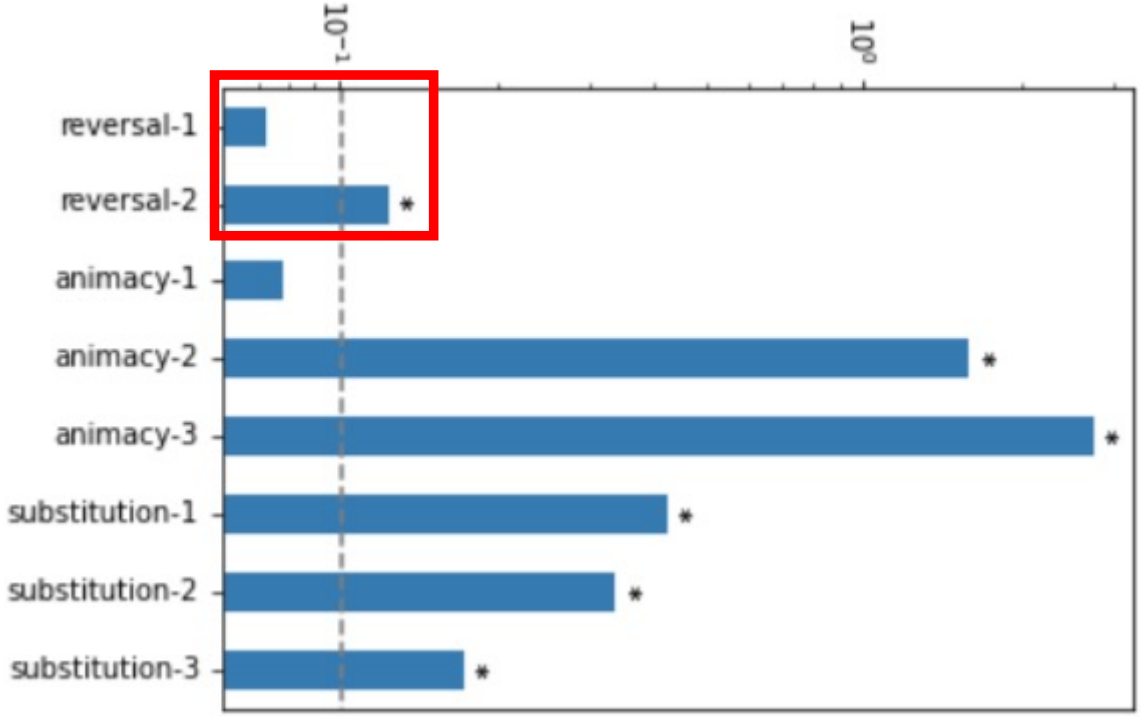


Ehrenhofer et al. (in press) (image Chow et al. 2015)

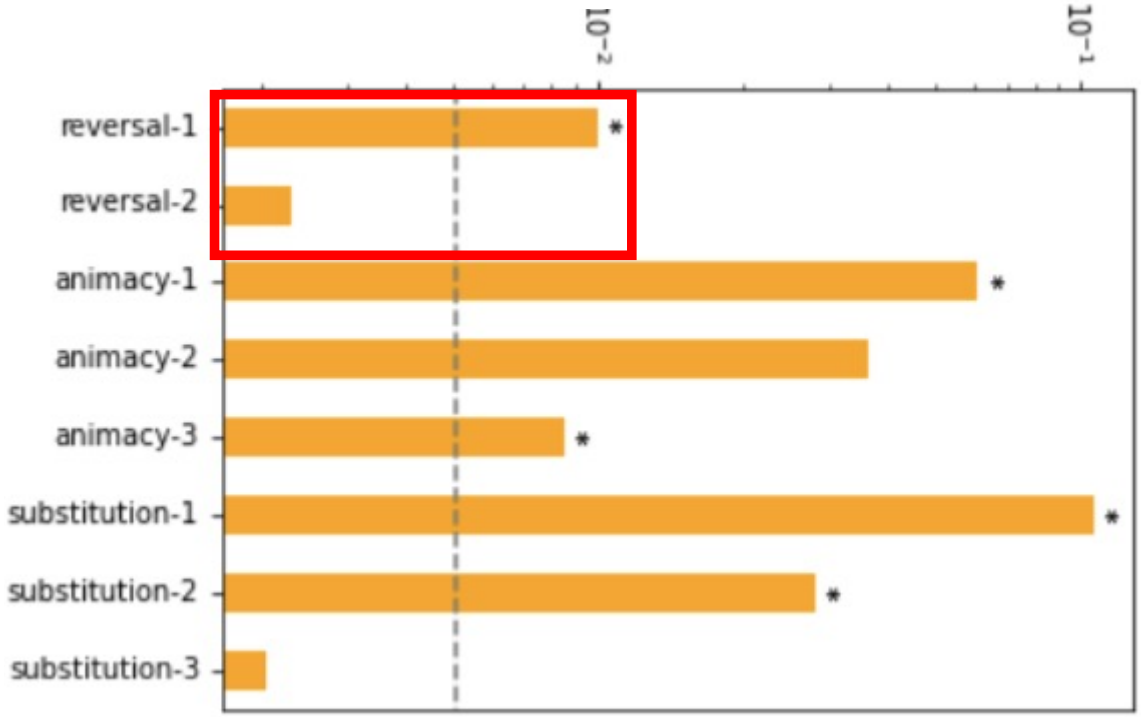
Role reversal divergence



Simulated N400 effect

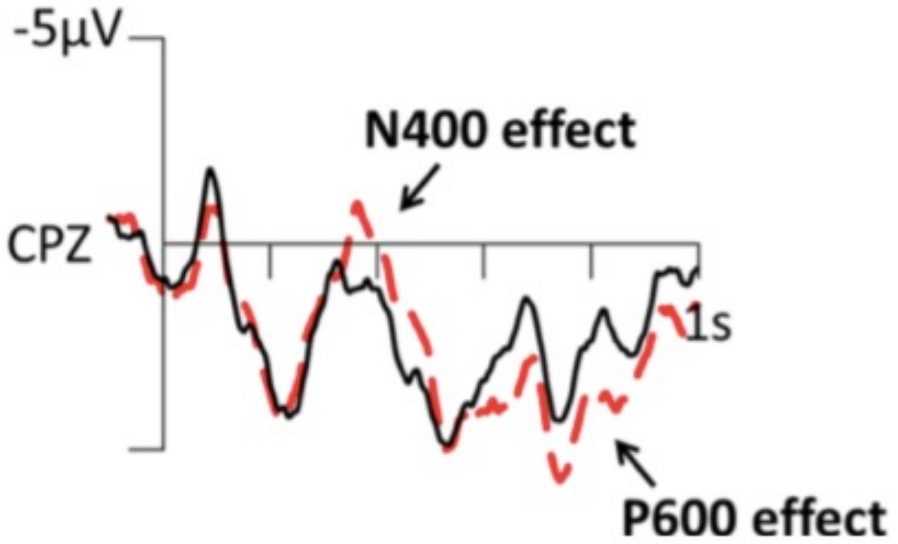


Simulated P600 effect



____ High cloze - *The tenant inquired which neighbor the landlord had evicted …*

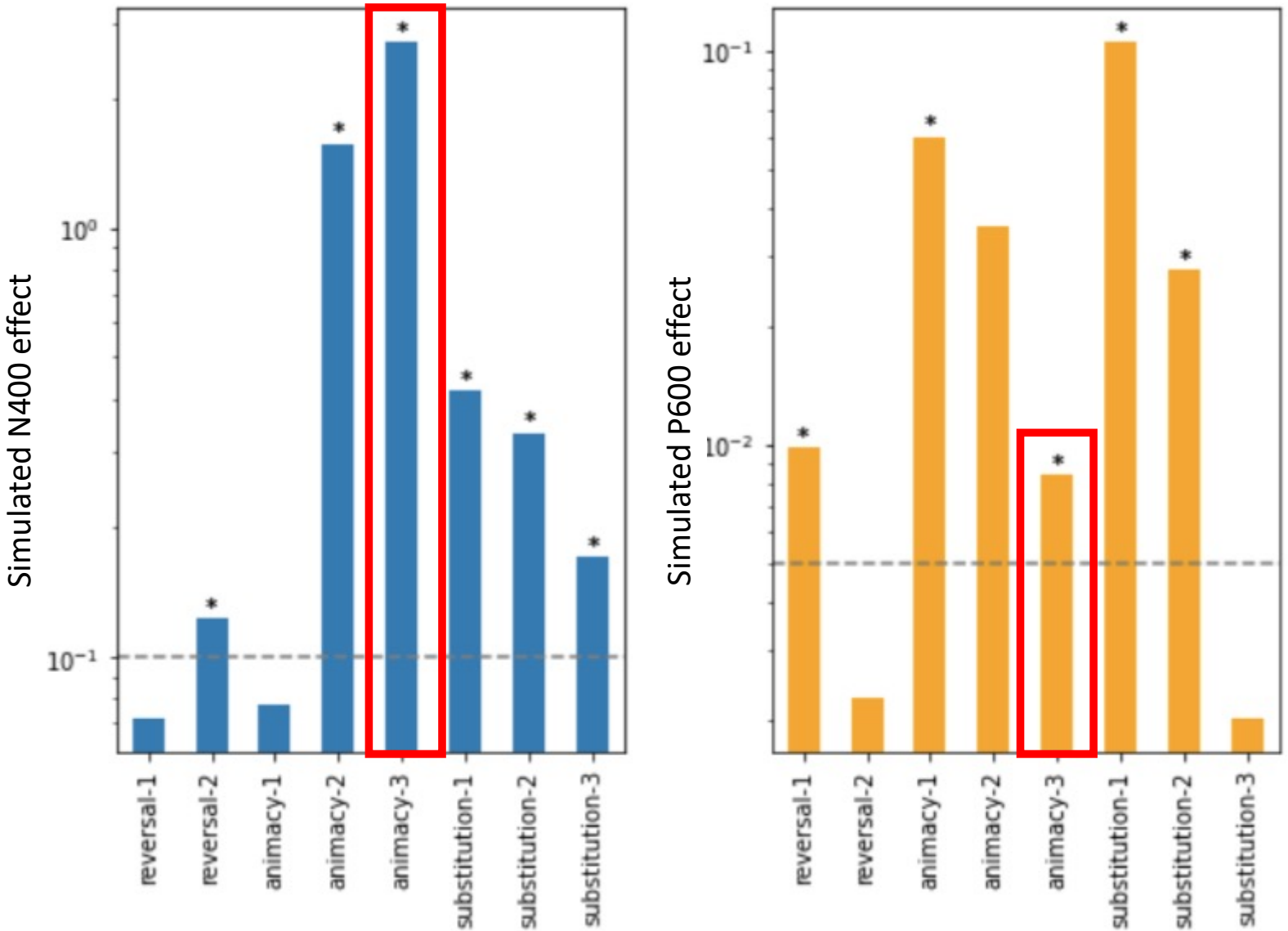
_ _ _ _ Low cloze - *The tenant inquired which exterminator the landlord had evicted…*



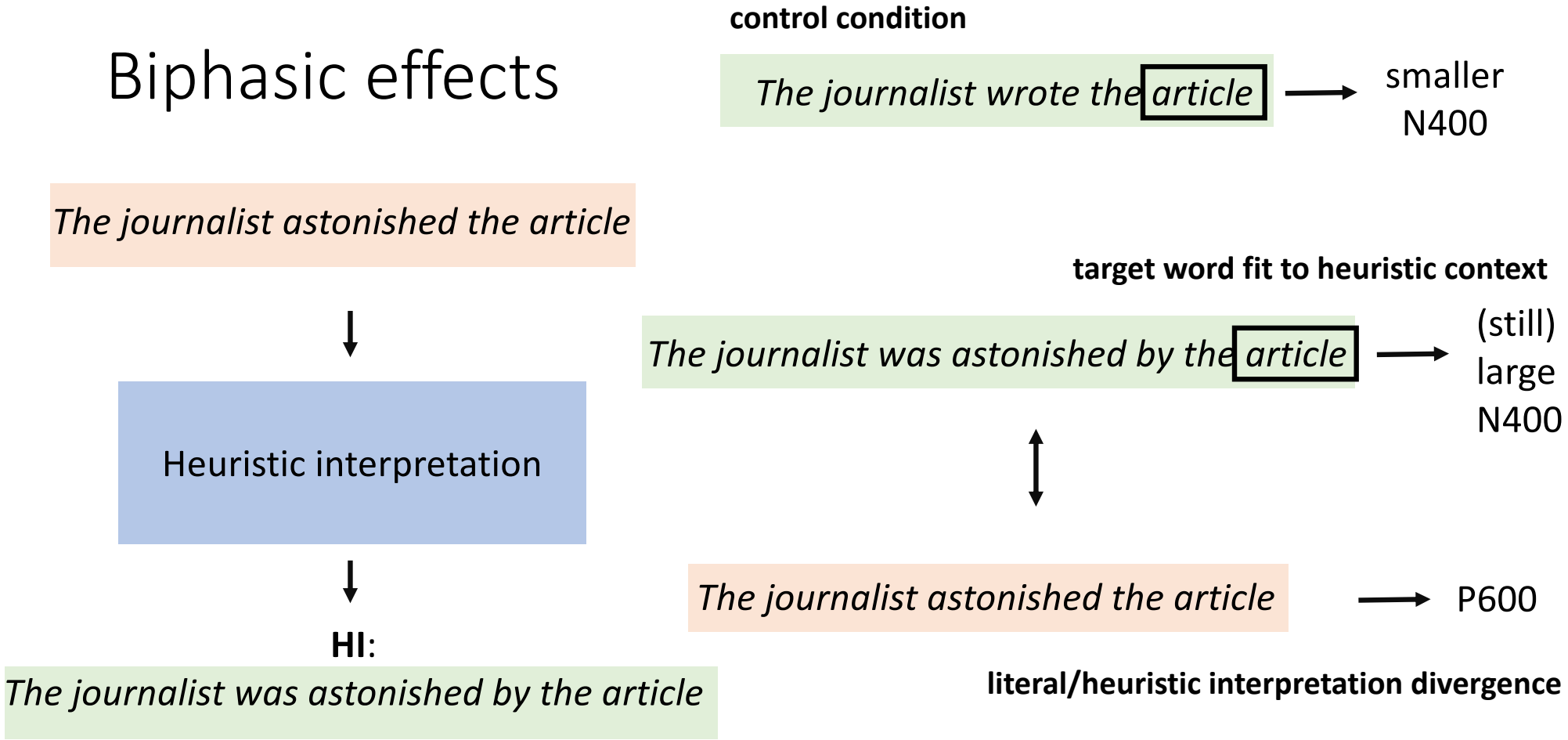
Chow et al. (2015)

Biphasic effects

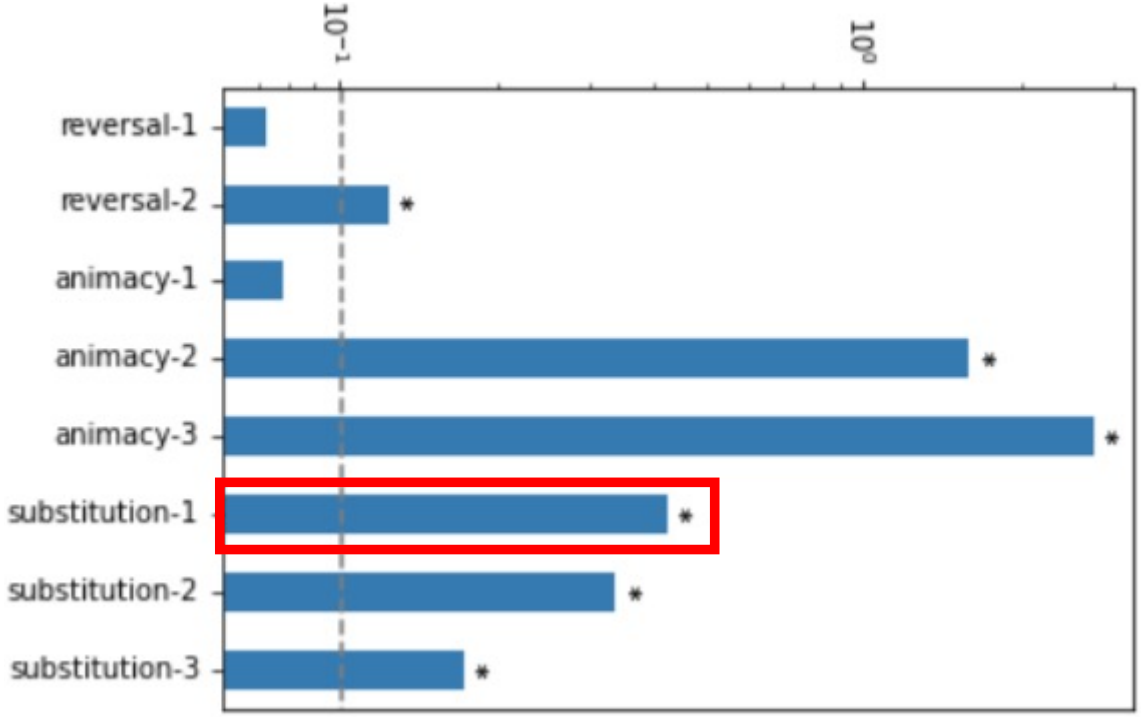
- We see that the dynamics leading to biphasic effects are not uniform across experiments


Simulated N400 effect
10^{0}
10^{-1}
reversal-1
reversal-2
animacy-1
animacy-2
animacy-3
substitution-1
substitution-2
substitution-3
Simulated P600 effect
10^{-1}
10^{-2}
reversal-1
reversal-2
animacy-1
animacy-2
animacy-3
substitution-1
substitution-2
substitution-3

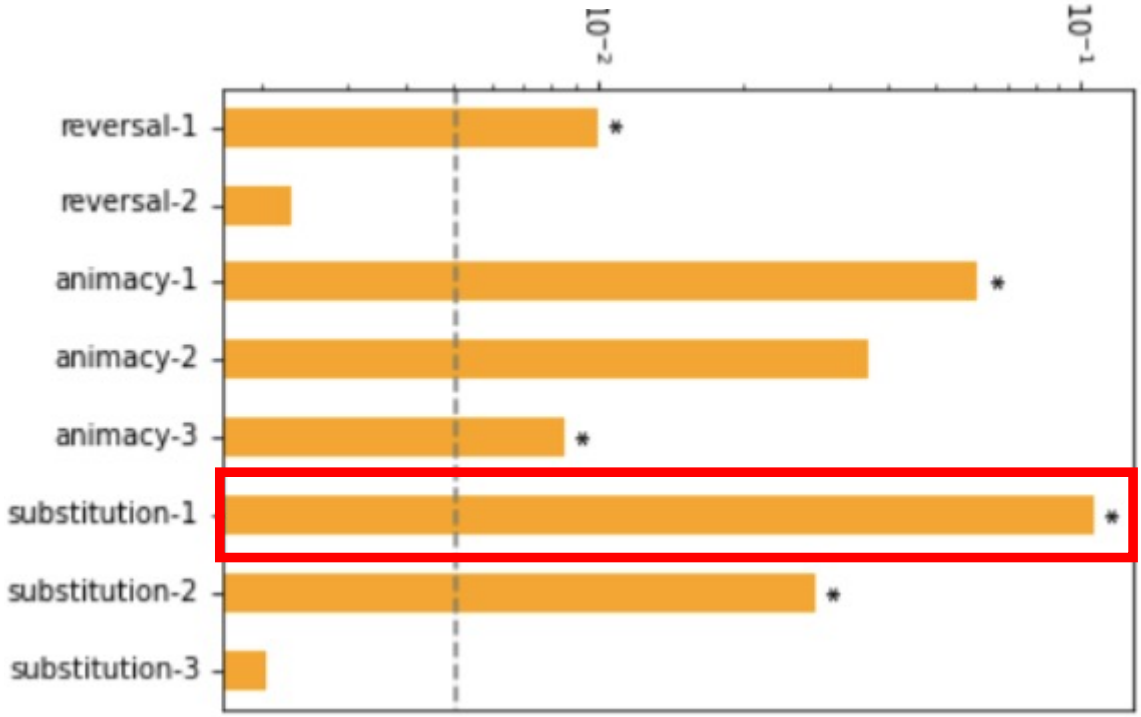
Biphasic effects



Simulated N400 effect



Simulated P600 effect



Biphasic effects

… inquired which exterminator the landlord had evicted

↓

Heuristic interpretation

↓

HI:

… inquired which neighbor the landlord had evicted

control condition

… inquired which neighbor the landlord had evicted → N400

no difference ↕

… inquired which neighbor the landlord had evicted → N400

↕

… inquired which exterminator the landlord had evicted → P600

literal/heuristic interpretation divergence

Biphasic effects

… inquired which exterminator the landlord had evicted

↓

Heuristic interpretation

↓

HI:

… inquired which exterminator the landlord had evicted

control condition

… inquired which neighbor the landlord had evicted → N400

****large difference****

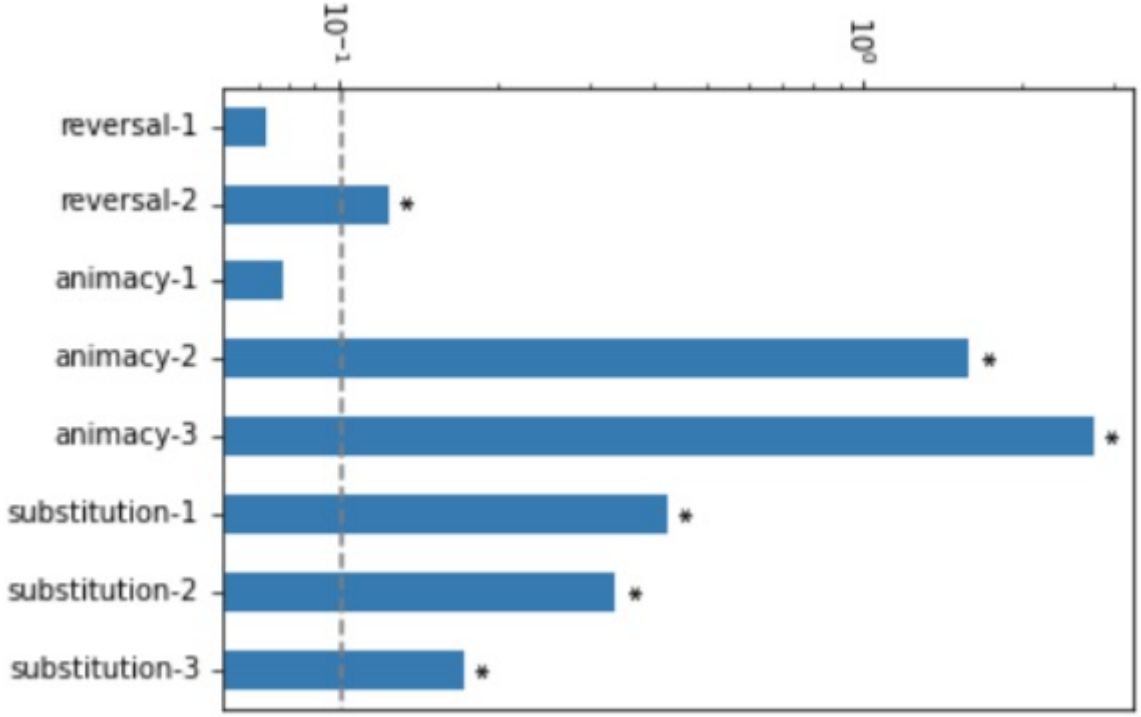
target word fit to heuristic context

… inquired which exterminator the landlord had evicted → N400

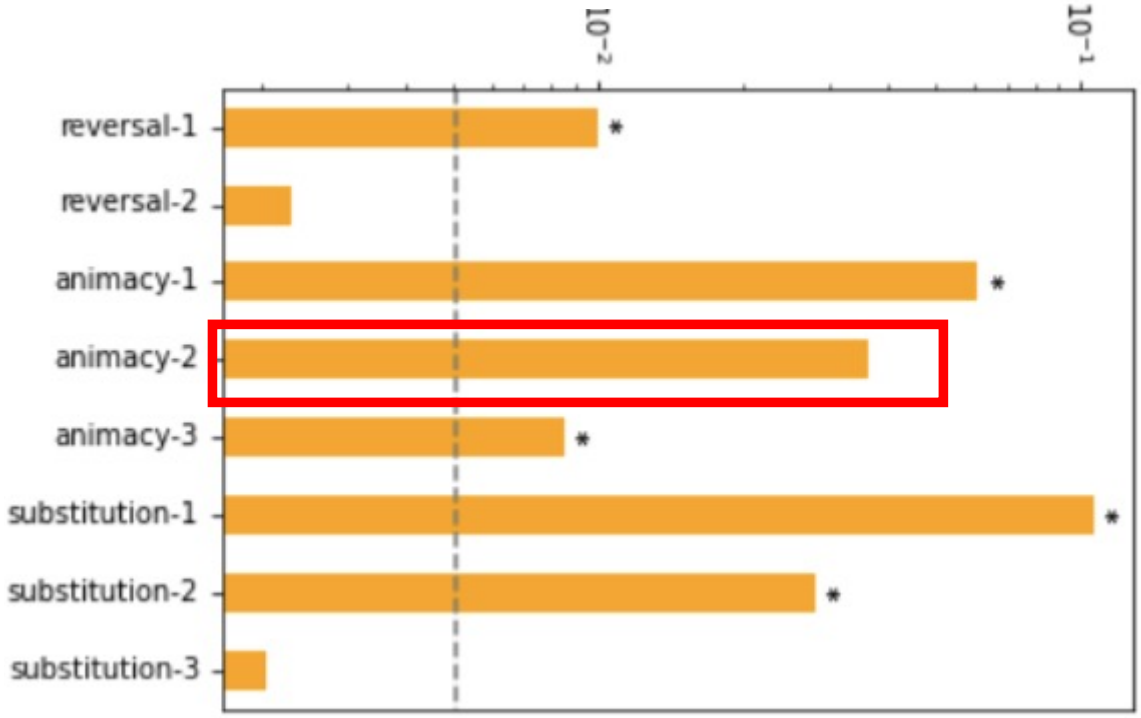
Model-human divergence

- Finally, our one unsuccessful simulation

Simulated N400 effect



Simulated P600 effect



Model-human divergence (human)

The dusty tabletop was devouring …

Heuristic interpretation

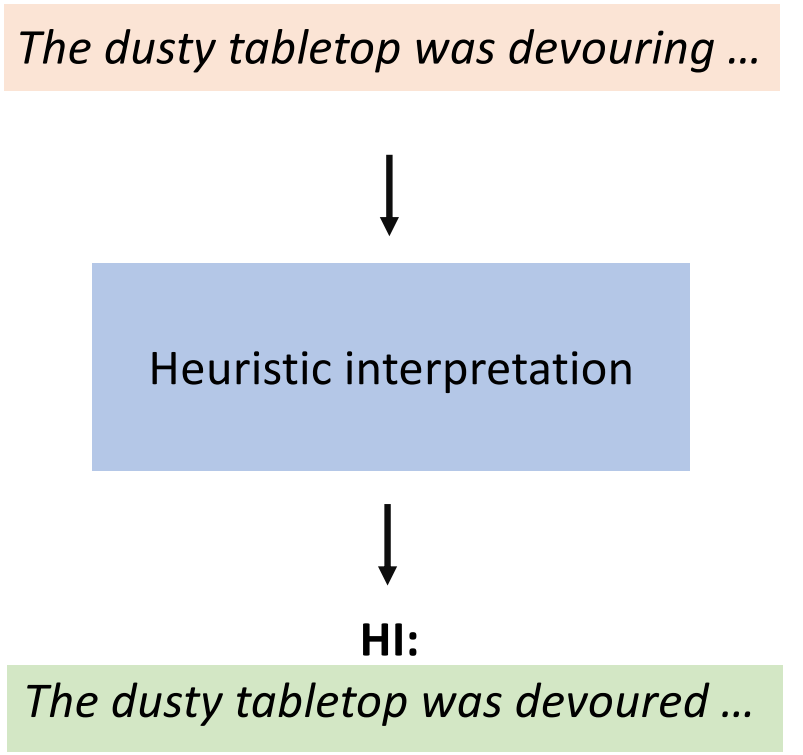
HI:

The dusty tabletop was devouring …

The dusty tabletop was devouring …

The dusty tabletop was devoured …

Model-human divergence (model)



The dusty tabletop was devouring …

The dusty tabletop was devoured …

Model-human divergence

- Suggests divergence between our NN-based plausibility proxy and human plausibility mechanism
- Models use the fact that inanimate subjects are usually associated with passive constructions, and prefer the passive
- Doesn’t appear to trigger reinterpretation in humans
- Area for improvement in aligning plausibility proxy with humans

Discussion

- Able to take existing psycholinguistic theories and strengthen explanatory coverage, shed light on fine-grained interactions between stimulus properties and posited mechanisms

- Successful simulations indicate that model represents strong candidate theory for underlying mechanisms

- Benefits from pre-trained LM measures indicate that processing mechanisms are sensitive to statistical properties of inputs in ways that prior theories did not account for

Cognitive implications

- Processor uses probabilistic inference mechanism to derive heuristic interpretations of input
- These processes are sensitive to fine-grained statistical properties of stimuli, in ways that mirror sensitivities from NLP models
- Heuristic interpretation needs to be reconciled with compositional interpretation, which processor also eventually has access to

NLP implications

- Estimates from NLP models do a decent job of capturing the probabilistic components of human processing
- But if we want to capture meaning extraction, necessary to align with compositional interpretation components of human processing
- Question: optimal to retain heuristic processing for efficiency, but default to compositional processing for robustness?

Key questions

To what extent do NLP models extract meaning, versus relying on shallower predictive heuristics?

How do meaning and prediction processes interact in the brain during real-time language processing?

Summarizing

- Controlled tests suggest that basic problems in meaning understanding remain unsolved in recent NLP models, though heuristics can create illusions of understanding

- Language modeling (word prediction) objective may come with fundamental limitations for forcing models to learn robust meaning

- May instead produce statistical sensitivities bearing resemblance to sensitivities that arise in brain for more efficient predictive processing

Summarizing

- Computational psycholinguistic simulations support probabilistic heuristic interpretation stage sensitive to fine-grained statistical properties of individual stimuli

- These probabilistic sensitivities likely increase efficiency, and can be reasonably approximated using measures from NLP models

- But eventually heuristic interpretations are reconciled with literal, compositional interpretations

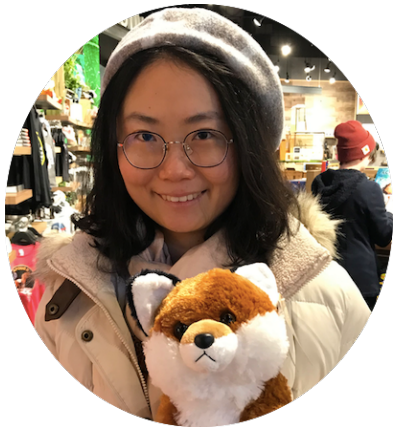
Thank you!

Lang Yu

Lalchand Pandia

Kanishka Misra

NSF Award No. 1941160
Toyota Technological Institute at Chicago



Jiaxuan Li