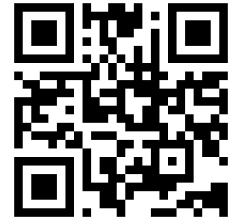
LLMs as a synthesis between symbolic and distributed approaches to language

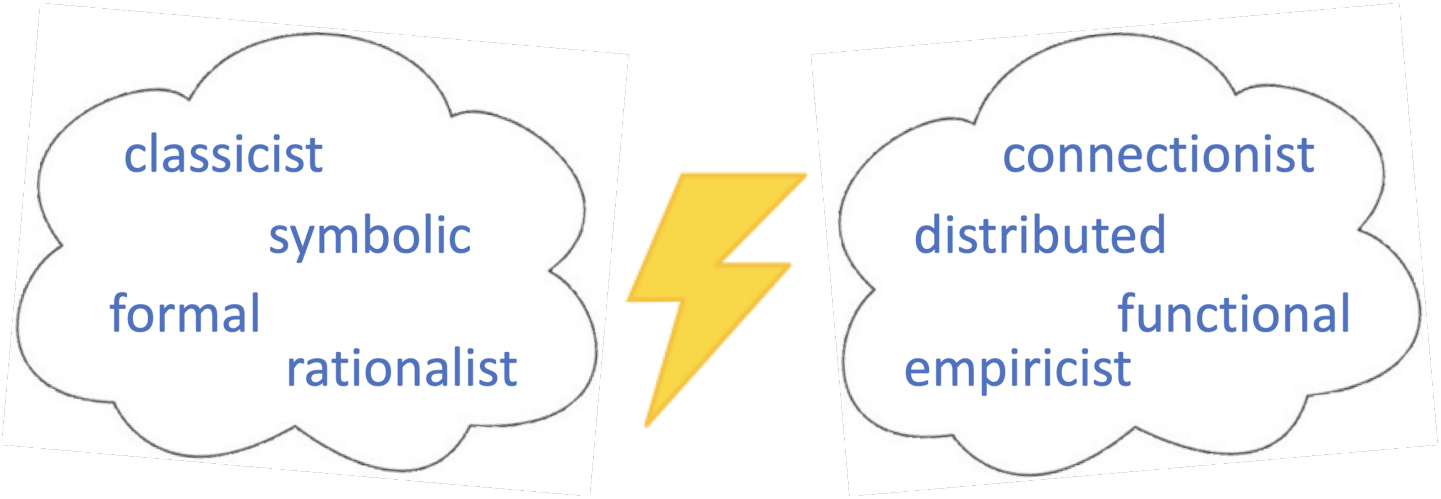
Gemma Boleda
Universitat Pompeu Fabra / ICREA

ILFC Seminar
January 22 2026



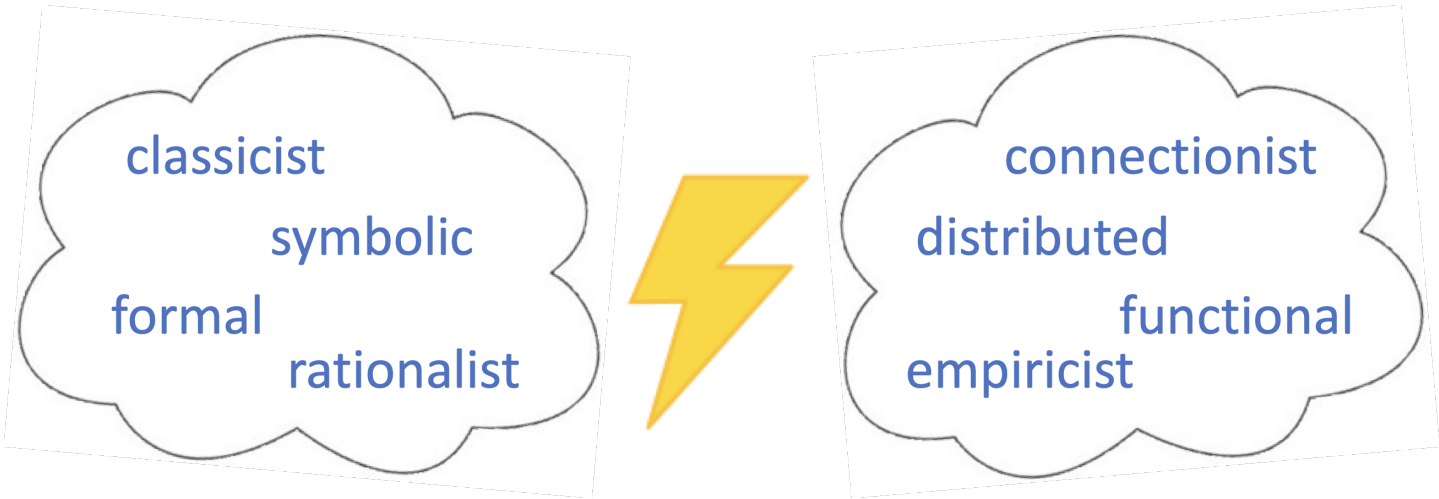
Opposing views on language and cognition

since ∼1950s: conflict



Opposing views on language and cognition

since ∼1950s: conflict

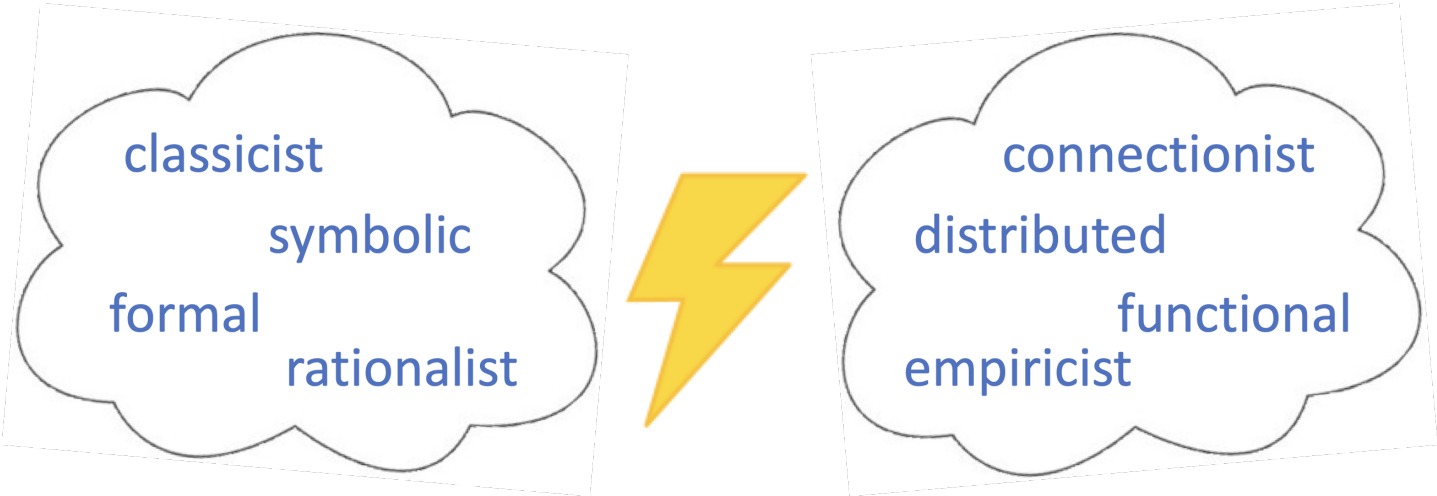


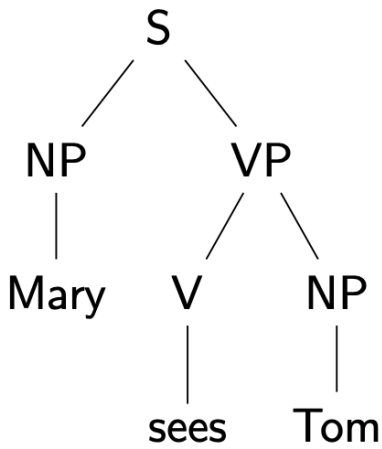
crux:

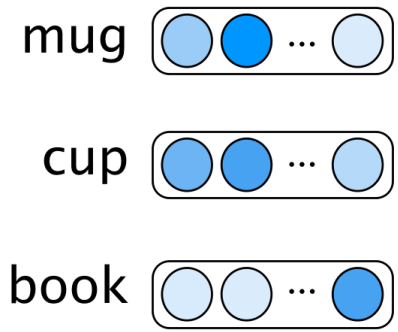
- some researchers focus on regularity
- others on messiness

Opposing views on language and cognition

since ∼1950s: conflict

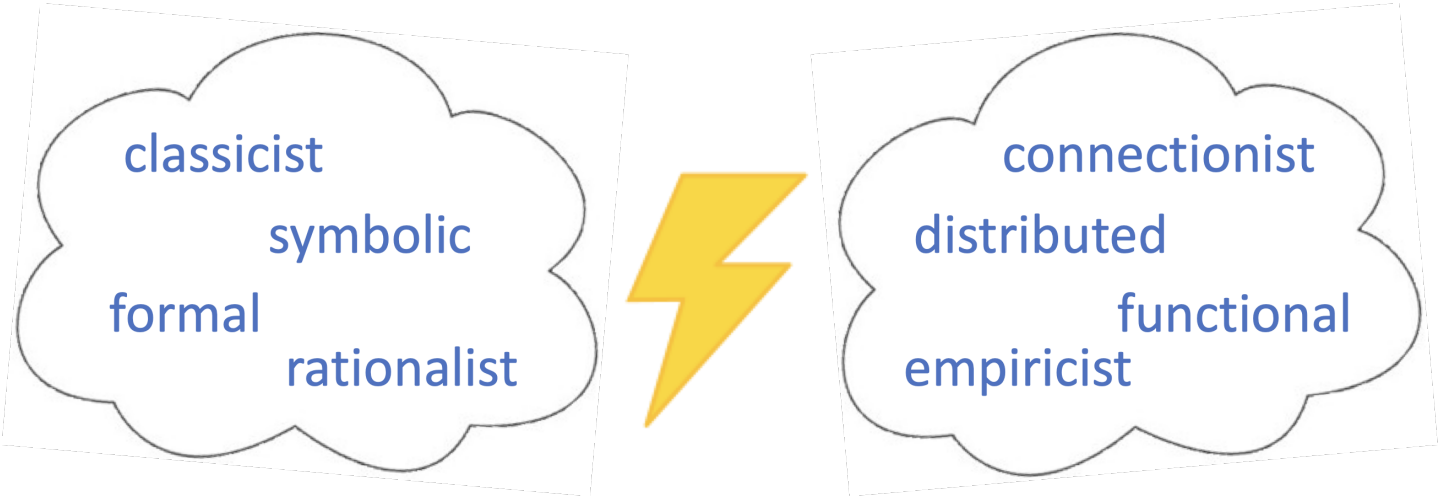






Opposing views on language and cognition

since ∼1950s: conflict



Modern language models refute Chomsky's approach to language

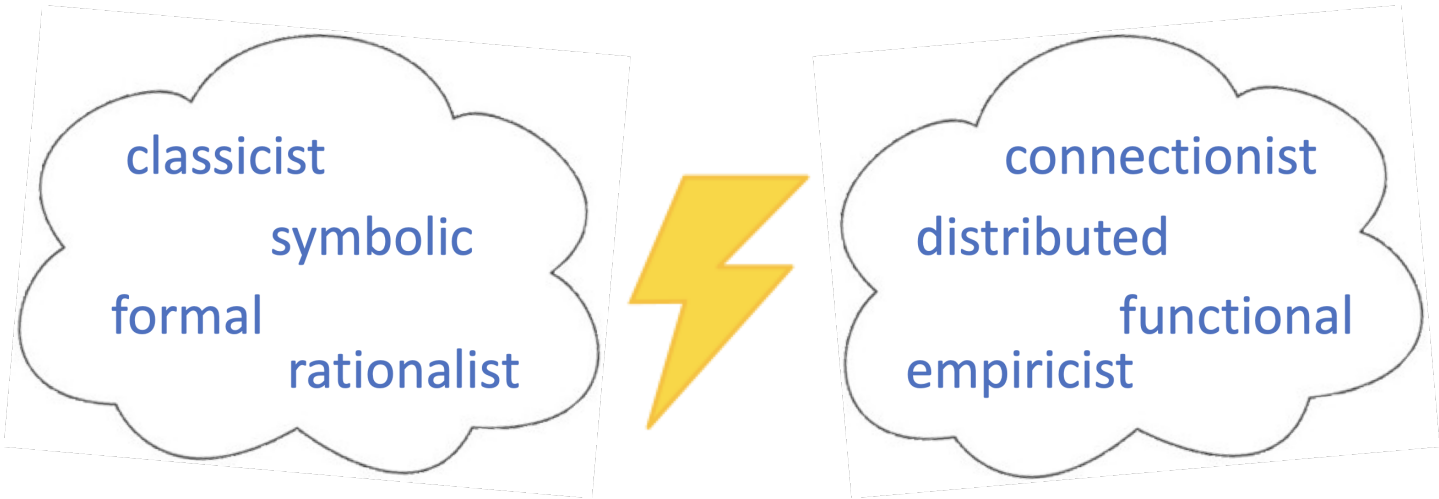
Steven T. Piantadosi

Modern Language Models Refute Nothing

Jon Rawski[1] and Lucie Baumont[2]

Opposing views on language and cognition

since ∼1950s: conflict



Modern language models refute Chomsky's approach to language

Steven T. Piantadosi

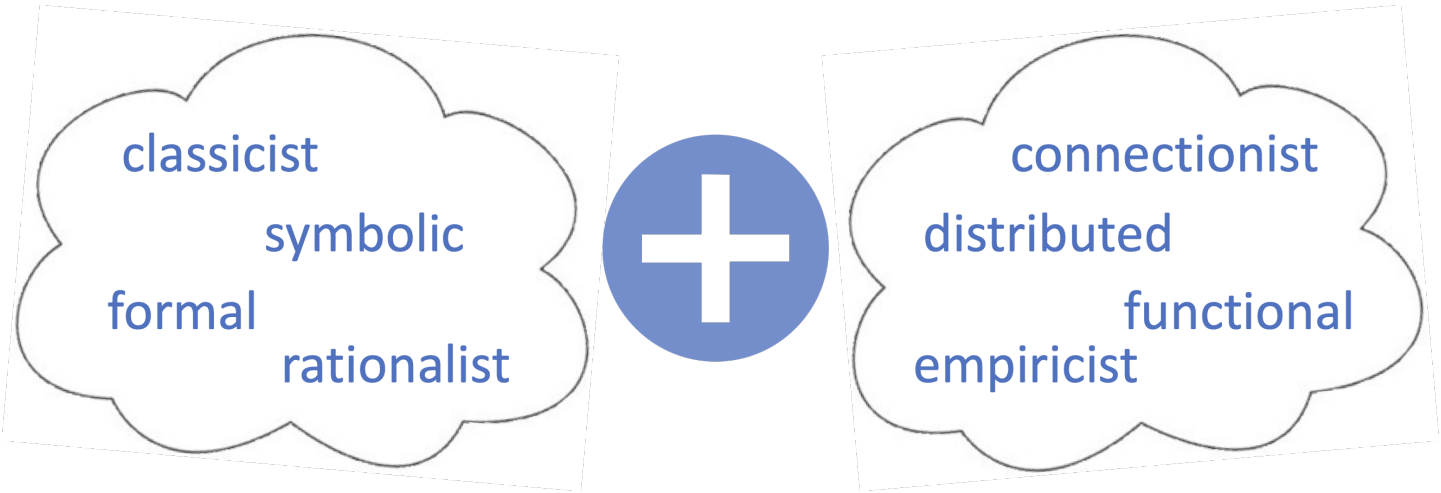
Modern Language Models Refute Nothing

Jon Rawski[1] and Lucie Baumont[2]

the *CL community should participate more in the debate!

Opposing views on language and cognition

since ∼1950s: conflict

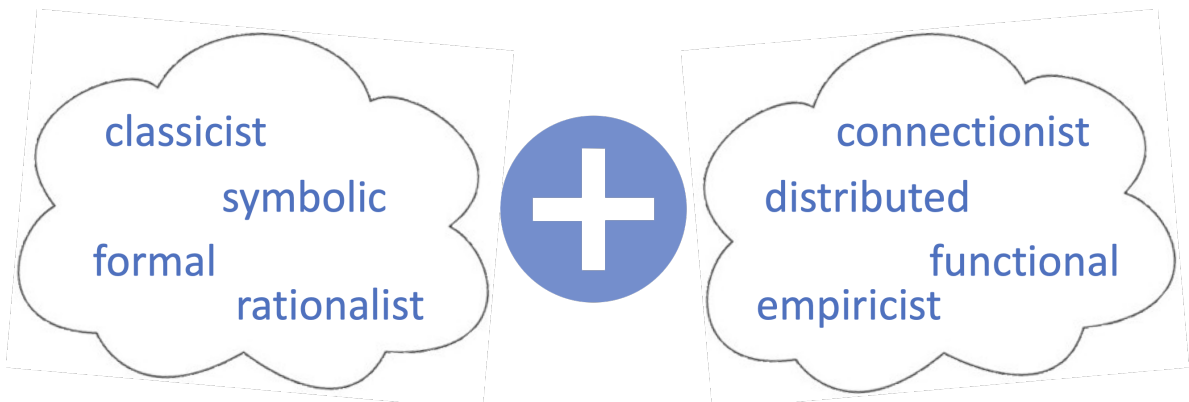


(Manning, 2015; Warstadt and Bowman, 2022; Futrell and Mahowald, 2025, . . .)

Theses

1. language is **both** regular and messy
2. modern LLMs are a **synthesis**

Boleda, G. 2025. LLMs as a synthesis between symbolic and distributed approaches to language. *Findings of the ACL*.



Plan of the talk

1. the "synthesis view"
   - 2 theses
2. challenges to this view
   - 2 studies and 1 review

Plan of the talk

1. the "synthesis view"
   - 2 theses
2. challenges to this view
   - 2 studies and 1 review

Thesis 1: Language is both regular and messy

example: morphosyntax in English

- categorical, rule-like properties
  $\rightarrow$ systematic link between form and function:
  - verbs inflect for tense
    frighten $\rightarrow$ *frightened*, *support* $\rightarrow$ *supported*, ...
  - prepositions don't
    **befored*, **ined*

Thesis 1: Language is both regular and messy

example: morphosyntax in English

- categorical, rule-like properties
  $\rightarrow$ systematic link between form and function:
  - verbs inflect for tense
    frighten $\rightarrow$ *frightened*, *support* $\rightarrow$ *supported*, ...
  - prepositions don't
    **befored*, **ined*
- messiness:
  - irregularities
    *go/*goed/went*, *drink/*drinked/drank*
  - fuzzy borders between categories
    frightened: verb, adjective?

The movie frightened the kid *The kid seemed very frightened*

Thesis 1: Language is both regular and messy

example: morphosyntax in English

- categorical, rule-like properties
  $\rightarrow$ systematic link between form and function:
  - verbs inflect for tense
    frighten $\rightarrow$ *frightened*, *support* $\rightarrow$ *supported*, ...
  - prepositions don't
    **befored*, **ined*
- messiness:
  - irregularities
    go/**goed*/*went*, *drink*/**drinked*/*drank*
  - fuzzy borders between categories
    frightened: verb, adjective?

The movie frightened the kid *The kid seemed very frightened*

She sent the email **The email seemed very sent*

Thesis 1: Language is both regular and messy

example: morphosyntax in English

- categorical, rule-like properties
  → systematic link between form and function:
  - verbs inflect for tense
    frighten → *frightened*, *support* → *supported*, ...
  - prepositions don't
    **befored*, **ined*
- messiness:
  - irregularities
    go/**goed*/*went*, *drink*/**drinked*/*drank*
  - fuzzy borders between categories
    frightened: verb, adjective?

The movie frightened the kid *The kid seemed very frightened*

She sent the email **The email seemed very sent*

- duality generalizes:
  phonology, morphology, syntax, semantics, pragmatics

Language is both regular and messy

- no scholar questions the empirical data
- what changes is the way they are appraised
  - formal linguists: focus on regularities
    - generative linguistics: (often) exclude messiness from linguistics
  - functional linguists: reject abstract rules
    - cognitive linguistics: (often) no formal predictive accounts

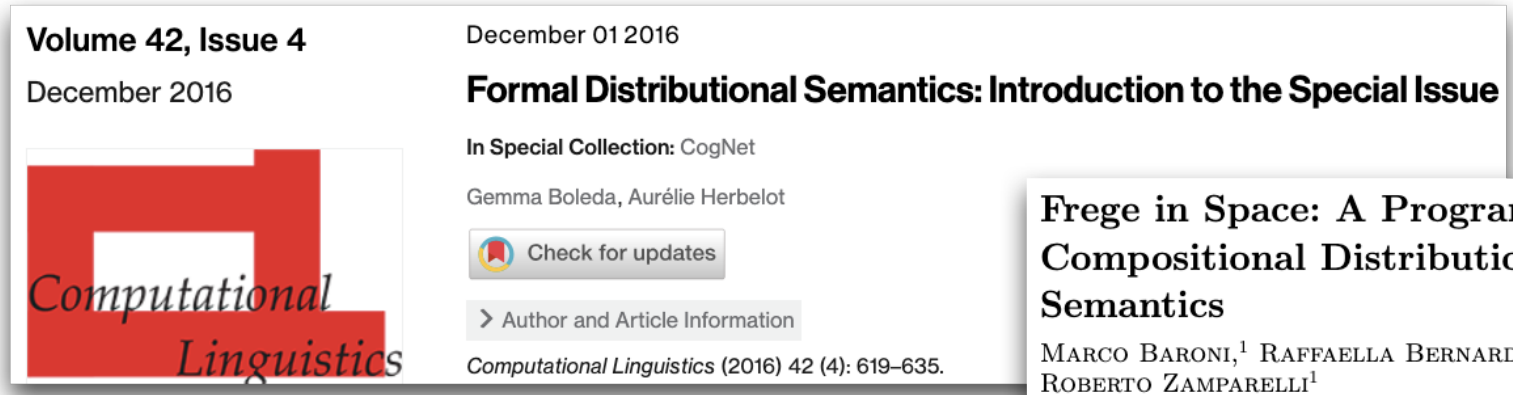
Language is both regular and messy

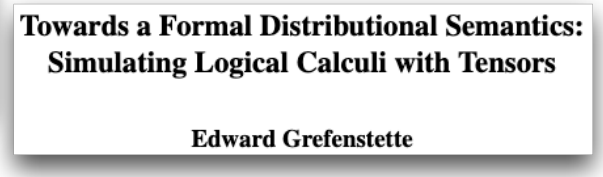
- no scholar questions the empirical data
- what changes is the way they are appraised
  - formal linguists: focus on regularities
    - generative linguistics: (often) exclude messiness from linguistics
  - functional linguists: reject abstract rules
    - cognitive linguistics: (often) no formal predictive accounts

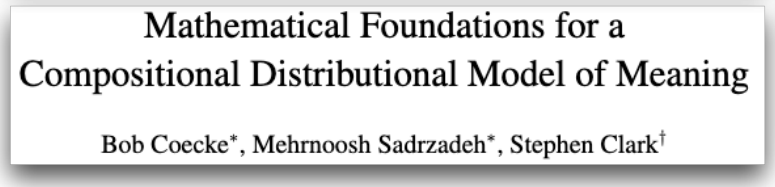
instead, we need

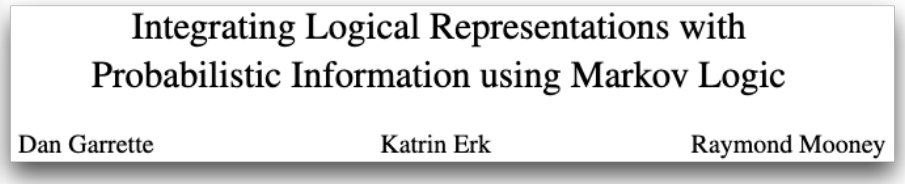
models that natively support both

Formal distributional semantics









Thesis 2: LLMs are a synthesis

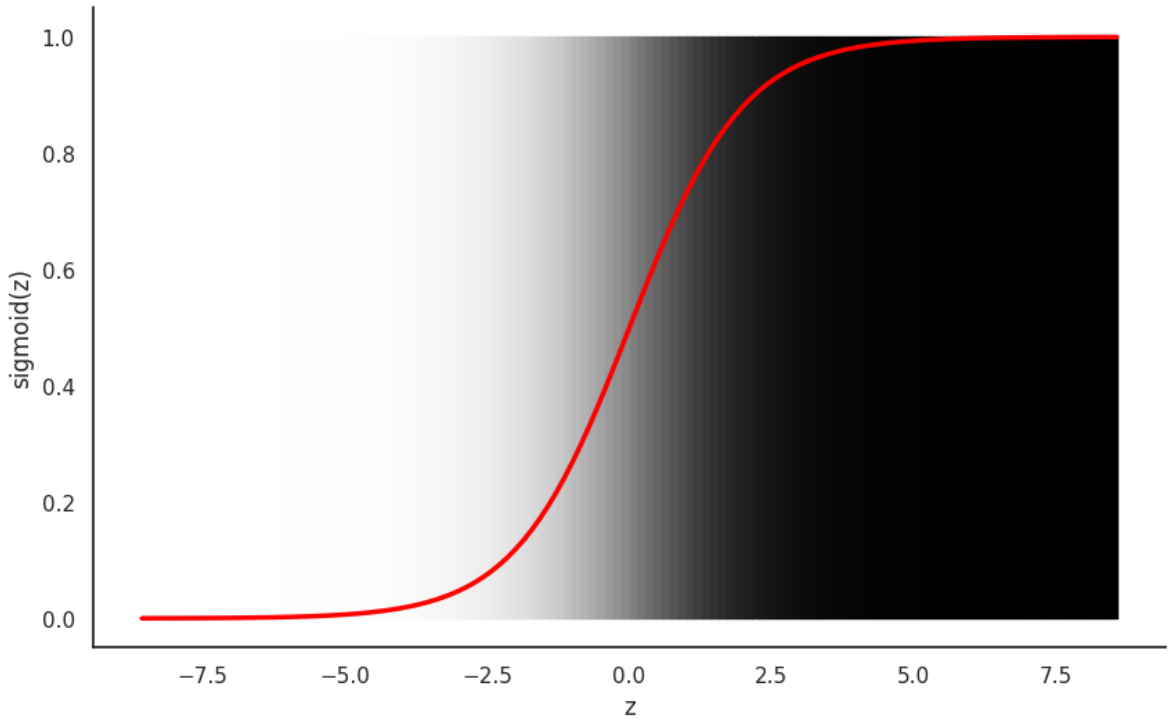
1. language is both regular and messy
2. modern LLMs are a **synthesis**
   between symbolic and distributed approaches

Thesis 2: LLMs are a synthesis

1. language is both regular and messy
2. modern LLMs are a **synthesis**
   between symbolic and distributed approaches

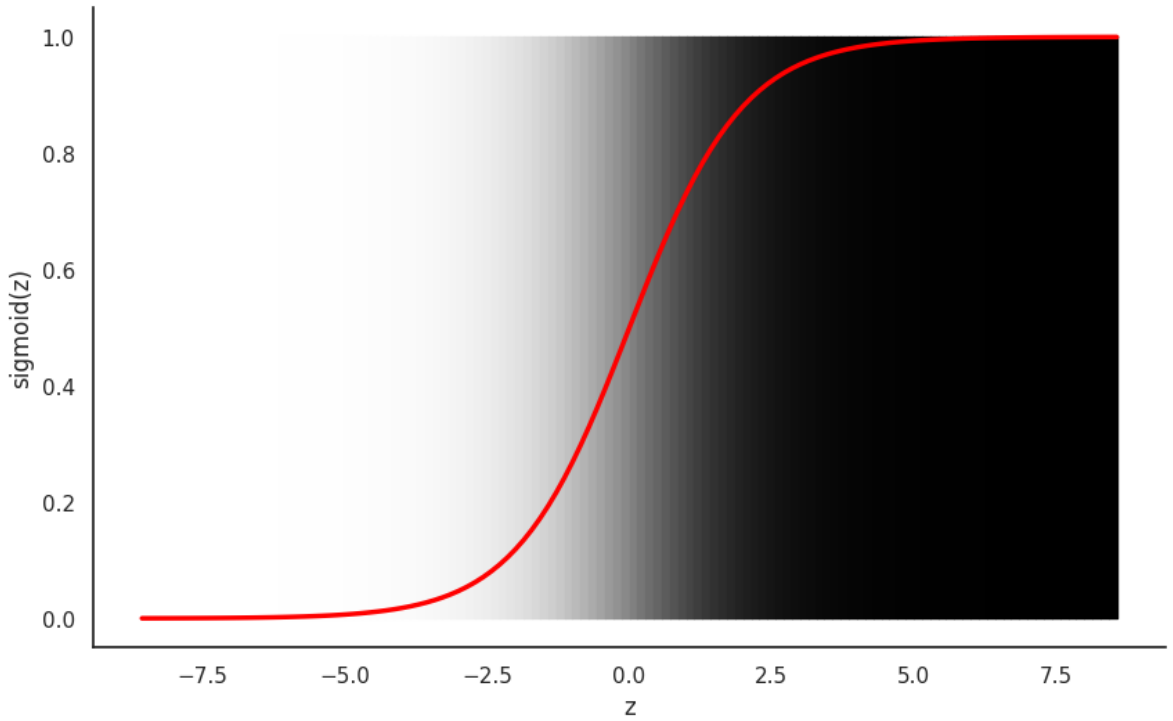
???

The synthesis view: regular AND messy



non-linearities provide the potential for both continuous and near-symbolic behavior

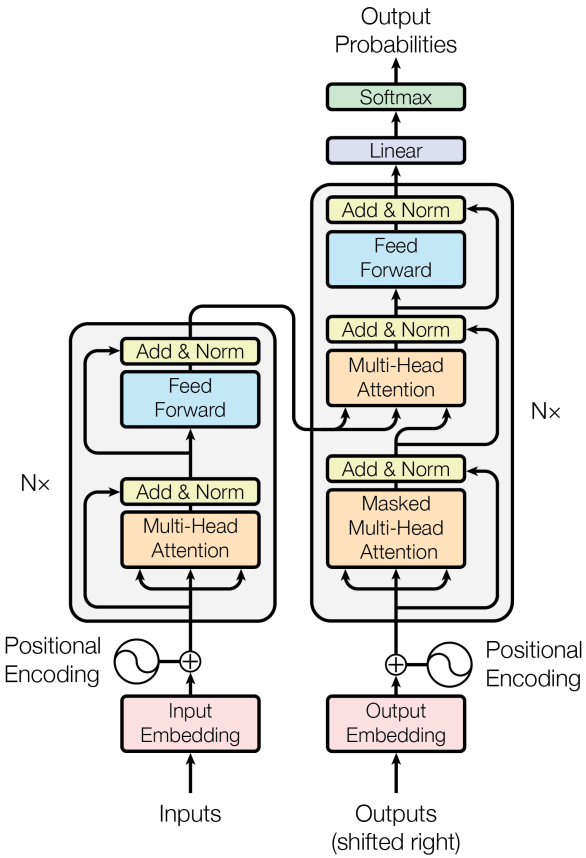
The synthesis view: regular AND messy



non-linearities provide the potential for both continuous and near-symbolic behavior

- known and shown in toy settings in pre-deep learning era
- LLMs?

The synthesis view: regular AND messy

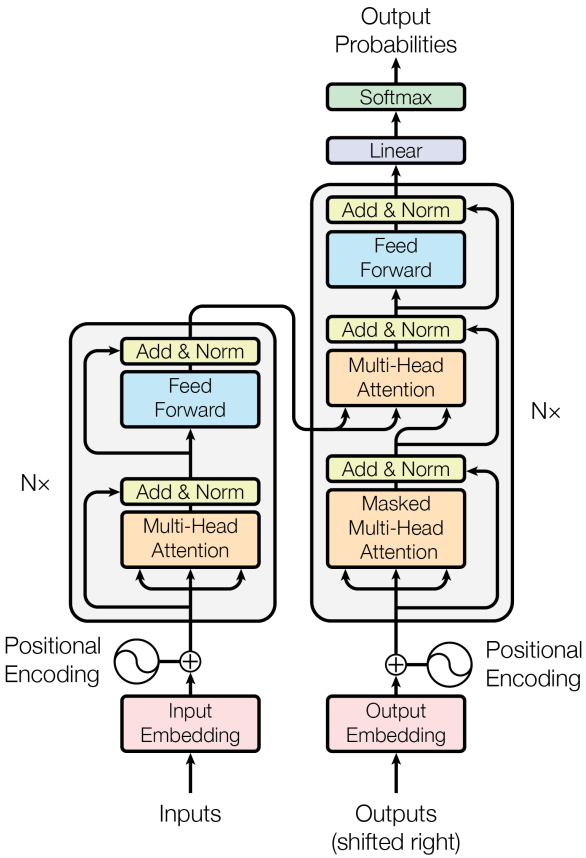


language data provides different kinds of pressures

$\Rightarrow$ hypothesis: differential encoding of phenomena in LLMs

- more symbolic-like encoding for grammar
- more distributed encoding for conceptual aspects of meaning

The synthesis view: regular AND messy



language data provides different kinds of pressures

⇒ hypothesis: differential encoding of phenomena in LLMs

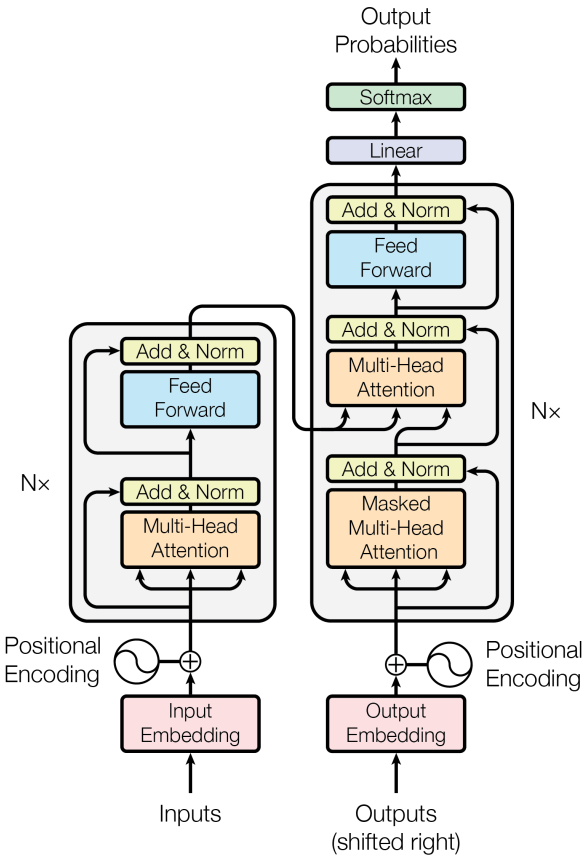
- more symbolic-like encoding for grammar
- more distributed encoding for conceptual aspects of meaning

interpretability studies have identified:

- individual neurons
- specific attention heads
- "circuits"

⇒ causally involved in grammatical processing in a near-symbolic fashion

The synthesis view: regular AND messy



language data provides different kinds of pressures

⇒ hypothesis: differential encoding of phenomena in LLMs

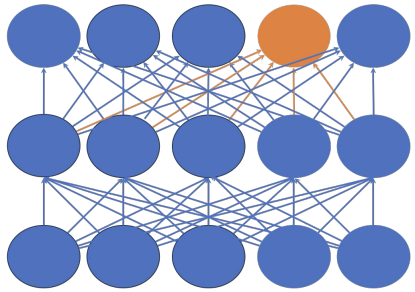
- more symbolic-like encoding for grammar
- more distributed encoding for conceptual aspects of meaning

interpretability studies have identified:

- individual neurons
- specific attention heads
- "circuits"

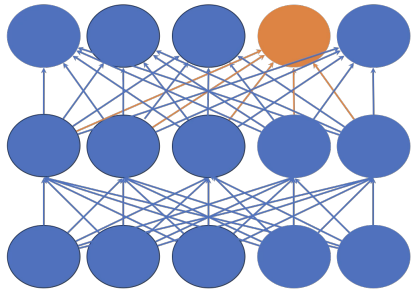
⇒ causally involved in grammatical processing in a near-symbolic fashion

Near-symbolic encoding: neurons



- individual neurons that
  - are selective for morphosyntactic properties
    e.g. number, tense, part of speech
  - or control syntactic relations
    e.g. agreement, syntactic dependencies

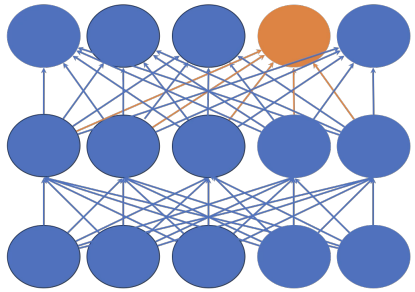
Near-symbolic encoding: neurons



example (Bau et al., 2019): machine translation model: single neuron controls tense of the translation

- *The committee supported the efforts of the authorities*
- original translation:
  Le Comité a appuyé$_{PAST}$ les efforts des autorités

Near-symbolic encoding: neurons



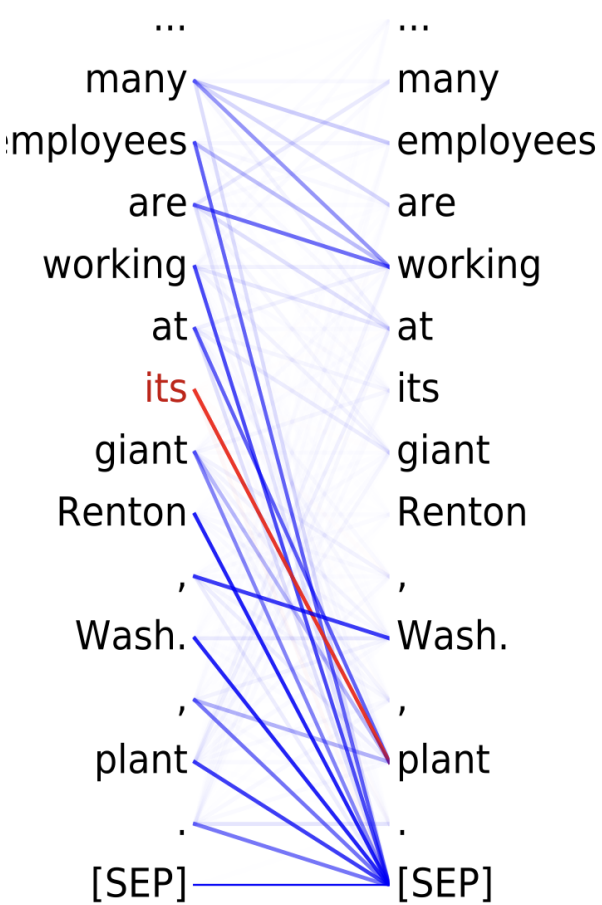
example (Bau et al., 2019): machine translation model: single neuron controls tense of the translation

- *The committee supported the efforts of the authorities*
- original translation:
  Le Comité a appuyé$_{PAST}$ *les efforts des autorités*
- changing neuron value:
  Le Comité appuie$_{PRESENT}$ *les efforts des autorités*

Different modes: attention heads

Clark et al. 2019

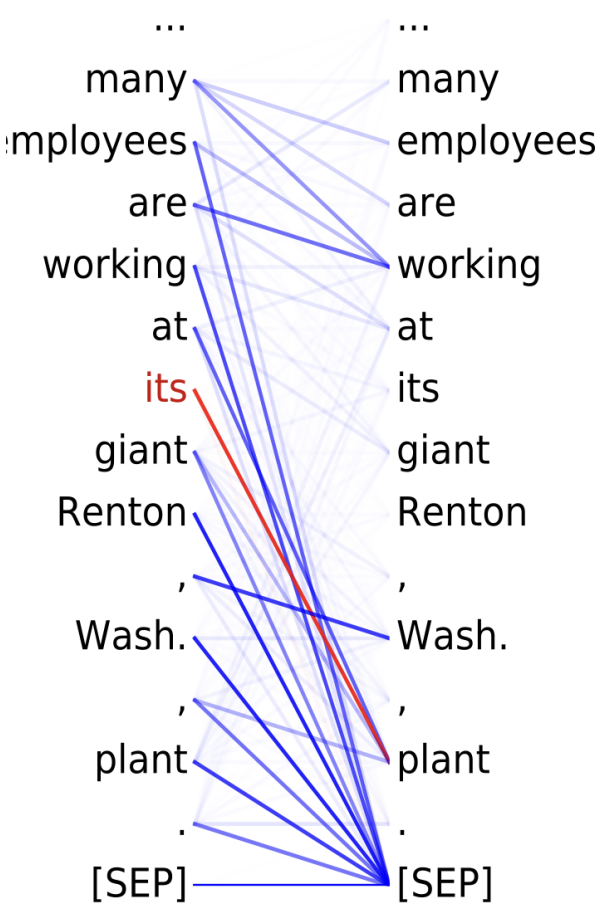
many employees are working at
its *giant Renton, Wash.,* plant



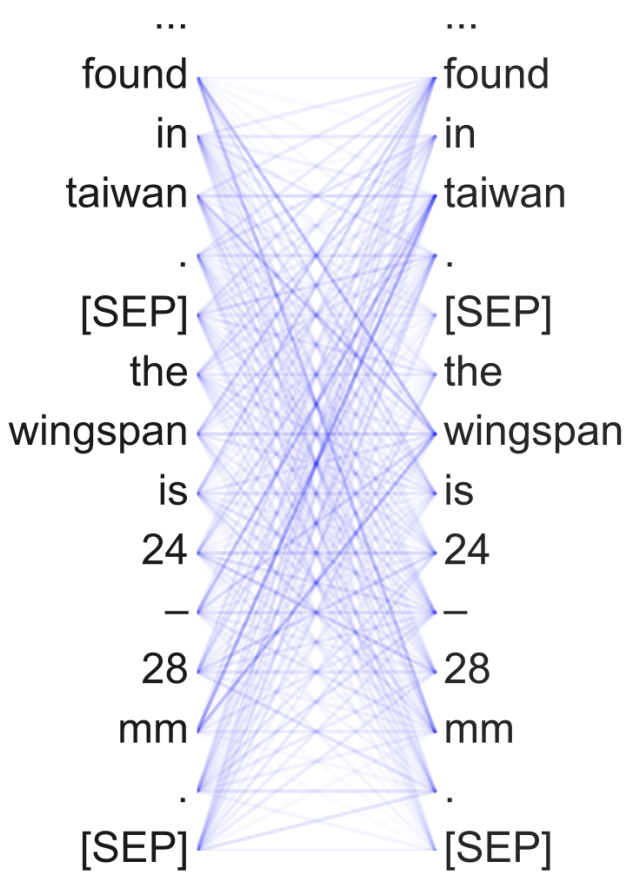
Different modes: attention heads

Clark et al. 2019

many employees are working at its *giant Renton, Wash.,* plant



found in Taiwan. The wingspan is 24-28mm,



Near-symbolic processing: circuits

Wang et al., 2023

When John and Mary went to the store, John gave a drink to $\rightarrow$ *Mary*

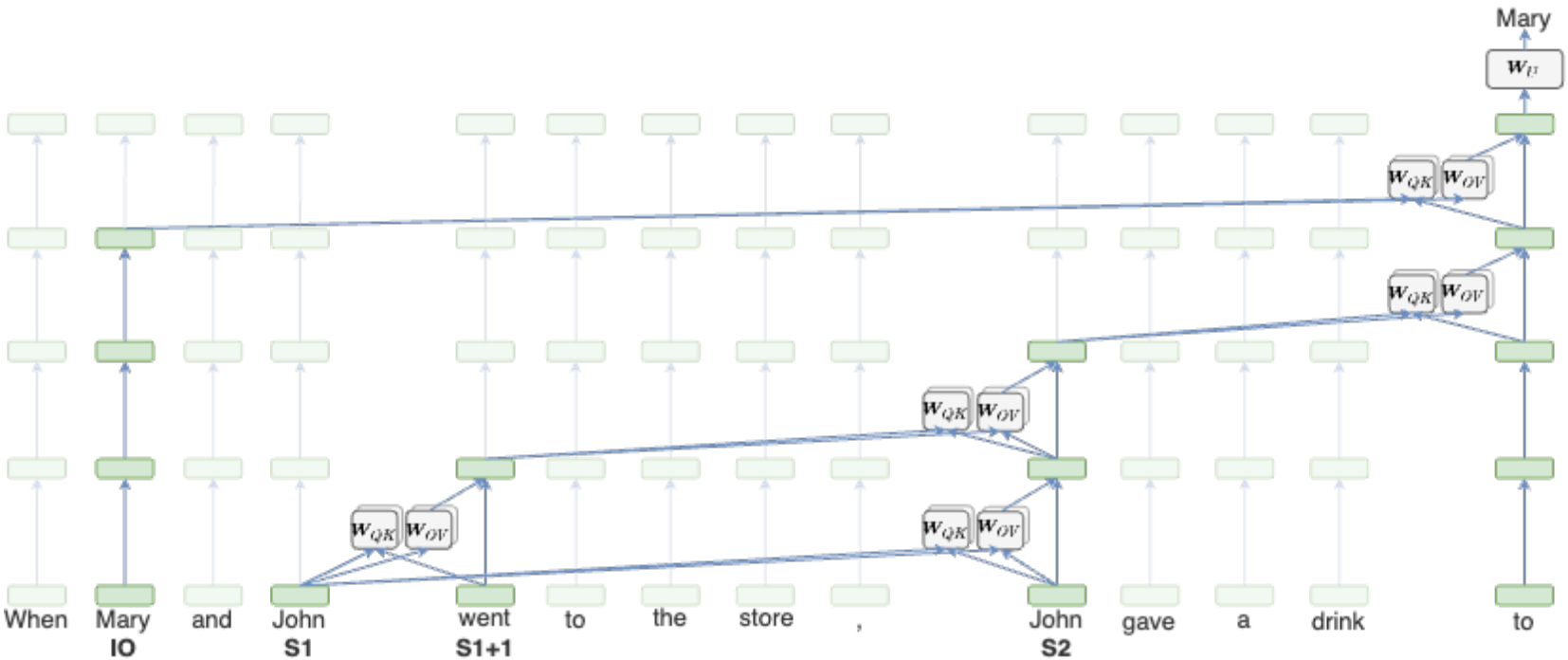


Figure by Javier Ferrando

Indirect-object circuit: how it works

When John and Mary went to the store, John gave a drink to → *Mary*

Algorithm:

1. identify the names in the adjunct clause
2. discard the names that appear in the main clause ("John"),
3. output the remaining name ("Mary").

Indirect-object circuit: how it works

When John and Mary went to the store, John gave a drink to → *Mary*

Algorithm:

1. identify the names in the adjunct clause
2. discard the names that appear in the main clause ("John"),
3. output the remaining name ("Mary").

Mechanism: via different attention heads that have specialized functions

Summing up

- language: both regular and messy
- symbolic representations good for regularities; distributed representations good for messiness
- LLMs exhibit both—plus flexibility

$\Rightarrow$ one of the main reasons for their amazing success at capturing natural language?

Summing up

- language: both regular and messy
- symbolic representations good for regularities; distributed representations good for messiness
- LLMs exhibit both—plus flexibility

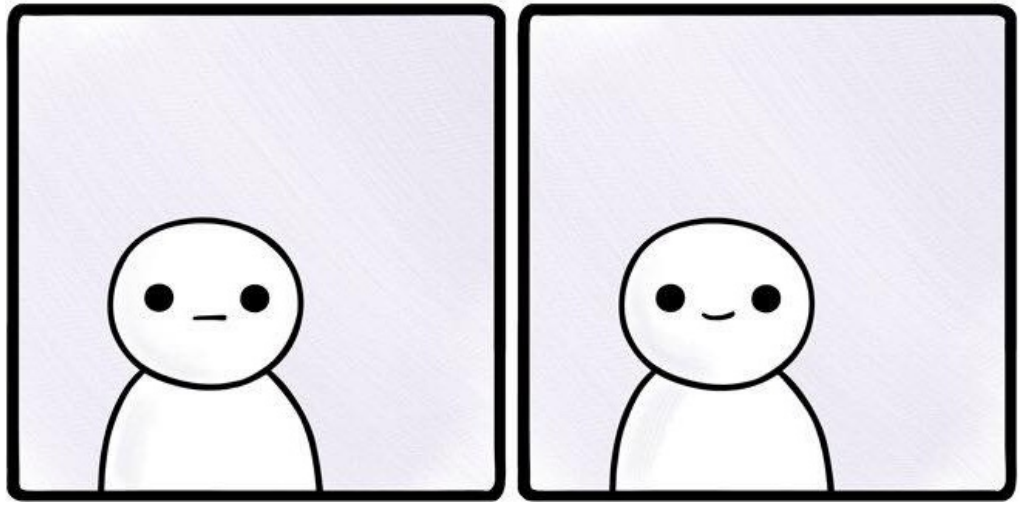
$\Rightarrow$ one of the main reasons for their amazing success at capturing natural language?

- these representations / behavior:

$\rightarrow$ response to pressure from language data

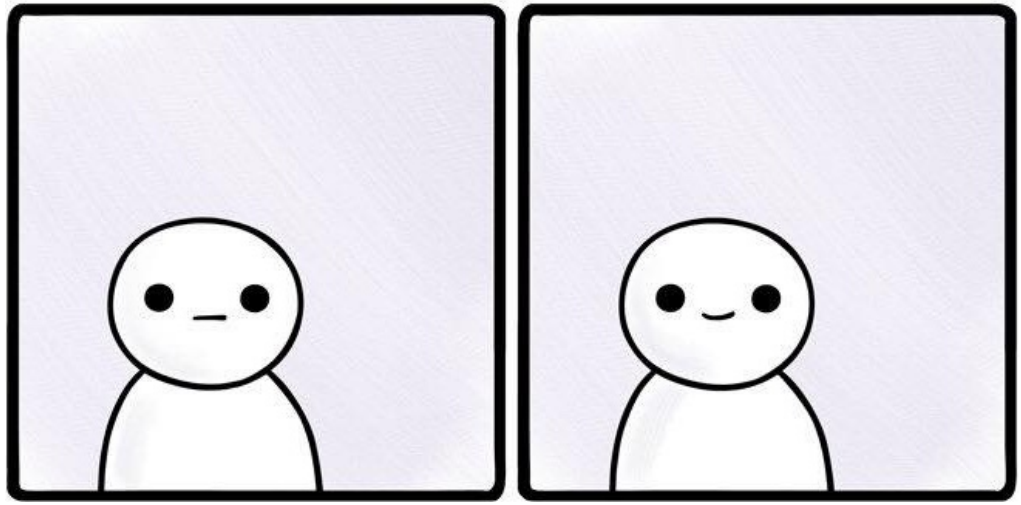
$\Rightarrow$ differential encoding of different phenomena
hypothesis: syntax vs (conceptual) semantics

BUT



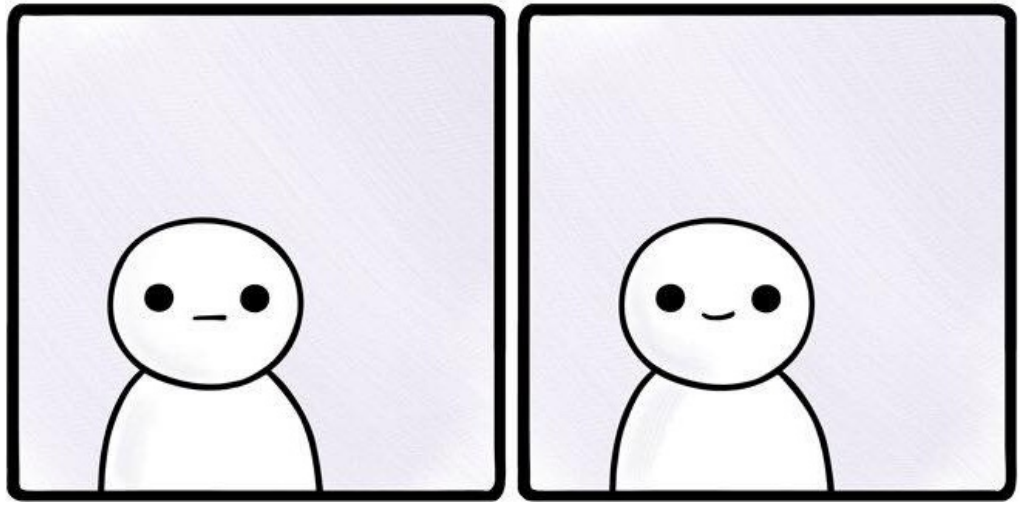
- nice story! does it hold up?
- typical work in linguistic interpretability:
  1. take classical syntactic notion
     (part of speech, agreement, . . .)
  2. check model for whether / how it is encoded

BUT



- nice story! does it hold up?
- typical work in linguistic interpretability:
  1. take classical syntactic notion
     (part of speech, agreement, . . .)
  2. check model for whether / how it is encoded
- cherry-picking? can miss discrete phenomena that are not your standard linguistic / reasoning phenomena

BUT



- nice story! does it hold up?
- typical work in linguistic interpretability:
  1. take classical syntactic notion
     (part of speech, agreement, . . .)
  2. check model for whether / how it is encoded
- cherry-picking? can miss discrete phenomena that are not your standard linguistic / reasoning phenomena
- not clear how prevalent near-symbolic processing is in the first place

Plan of the talk

1. the "synthesis view"
   - 2 theses
2. challenges to this view
   - 2 studies and 1 review

Outlier dimensions

Macocco et al. (2025)

Macocco, Graichen, Boleda, Baroni (2025) Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models. *BlackboxNLP*.

- outlier dimensions: dimensions that display **extreme activations** for the **majority of inputs**

Outlier dimensions

Macocco et al. (2025)

Macocco, Graichen, Boleda, Baroni (2025) Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models. *BlackboxNLP*.

- outlier dimensions: dimensions that display **extreme activations** for the **majority of inputs**
  - given a model and dataset

Outlier dimensions

Macocco et al. (2025)

Macocco, Graichen, Boleda, Baroni (2025) Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models. *BlackboxNLP*.

- outlier dimensions: dimensions that display **extreme activations** for the **majority of inputs**
  - given a model and dataset,
  - extreme activation values: among the 1% most extreme (positively or negatively) across all inputs and dimensions
  - outlier: dimension whose median activation across dataset of sentences is an extreme value

Outlier dimensions

Macocco et al. (2025)

Macocco, Graichen, Boleda, Baroni (2025) Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models. *BlackboxNLP*.

- outlier dimensions: dimensions that display **extreme activations** for the **majority of inputs**
  - given a model and dataset,
  - extreme activation values: among the 1% most extreme (positively or negatively) across all inputs and dimensions
  - outlier: dimension whose median activation across dataset of sentences is an extreme value
- focus: last layer (MLP down-projection matrix)

Outlier dimensions: a near-symbolic/discrete mechanism

model	#ODs
pythia-12b	36
mistral-7b	28
llama-8b	12
olmo2-13b	24
qwen-14b	38
opt-13b	4
gemma-9b	6
stable-12b	23

Outlier dimensions: a near-symbolic/discrete mechanism

model	#ODs	most predicted tokens
pythia-12b	36	_the _a _D
mistral-7b	28	_the _a _ _un _two _large
llama-8b	12	_in , _(_and _
olmo2-13b	24	_the _ , _The _in _A
qwen-14b	38	_ _the ,
opt-13b	4	I
gemma-9b	6	_
stable-12b	23	n.a.

Outlier dimensions: a near-symbolic/discrete mechanism

model	#ODs	most predicted tokens
pythia-12b	36	_the _a _D
mistral-7b	28	_the _a _ _un _two _large
llama-8b	12	_in , _(_and _
olmo2-13b	24	_the _ , _The _in _A
qwen-14b	38	_ _the ,
opt-13b	4	I
gemma-9b	6	_
stable-12b	23	n.a.

outlier dimensions encode a "default to frequent words" heuristic...

Outlier dimensions: a near-symbolic/discrete mechanism

model	#ODs	most predicted tokens
pythia-12b	36	_the _a _D
mistral-7b	28	_the _a _ _un _two _large
llama-8b	12	_in , _(_and _
olmo2-13b	24	_the _ , _The _in _A
qwen-14b	38	_ _the ,
opt-13b	4	I
gemma-9b	6	_
stable-12b	23	n.a.

outlier dimensions encode a "default to frequent words" heuristic. . . in many, but not all models!

How do ODs work?

- outlier dimensions are always activated
- remaining dimensions "conspire" against outliers to produce lower-frequency words

Recap: a "weird" discrete mechanism

- remember: typical work in linguistic interpretability:
  1. take classical syntactic notion
     (part of speech, agreement, . . .)
  2. check model for whether / how it is encoded
- a bit like cherry-picking; can miss discrete phenomena that are not your standard linguistic / reasoning phenomena

$\Rightarrow$ we've uncovered a near-symbolic mechanism in an LLM... which boosts frequent words!

Recap: a "weird" discrete mechanism

- remember: typical work in linguistic interpretability:
  1. take classical syntactic notion
     (part of speech, agreement, . . .)
  2. check model for whether / how it is encoded
- a bit like cherry-picking; can miss discrete phenomena that are not your standard linguistic / reasoning phenomena

$\Rightarrow$ we've uncovered a near-symbolic mechanism in an LLM... which boosts frequent words!

- how prevalent is near-symbolic processing?

Tracing near-discreteness in LLMs

Work in progress with Corentin **Kervadec**, Iuliia Lysova and Marco Baroni

- LLMs: billions of parameters, deep and wide structure
- to what extent do LLMs use all these parameters?

Method

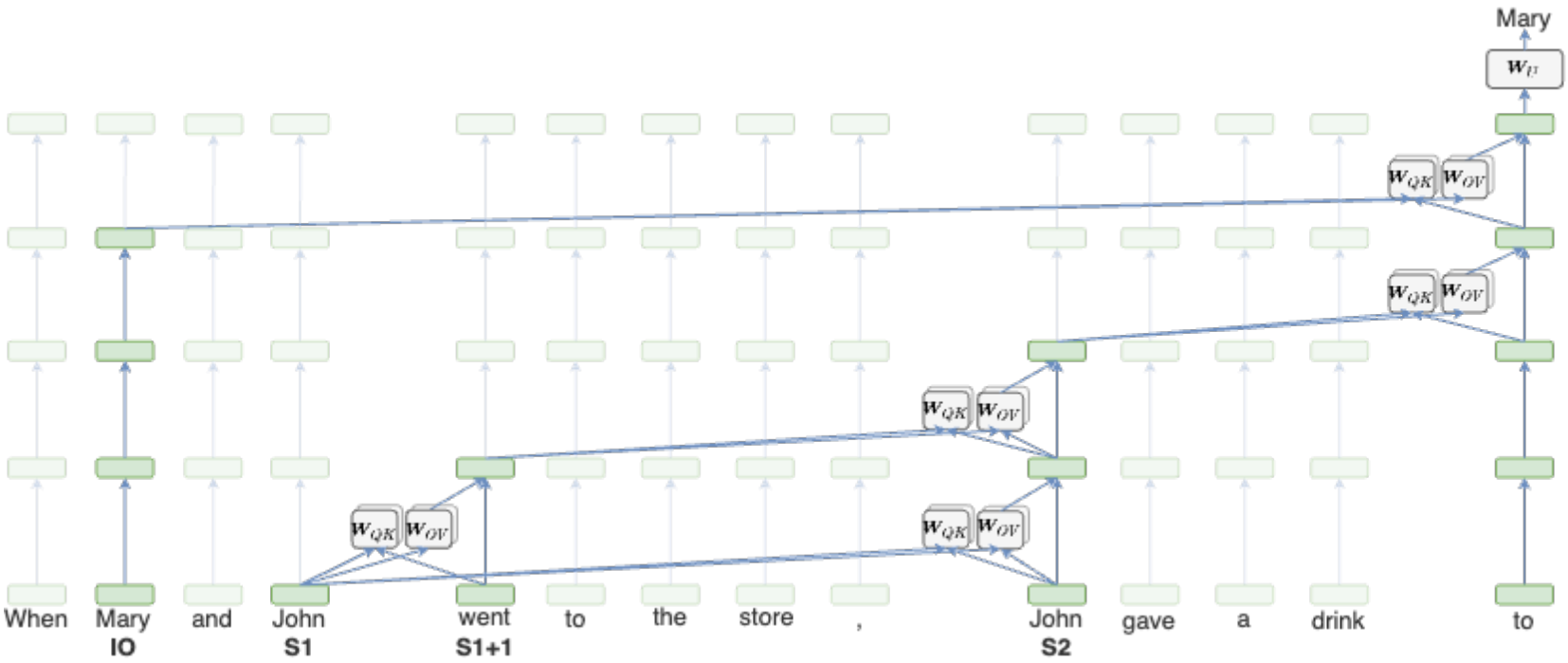
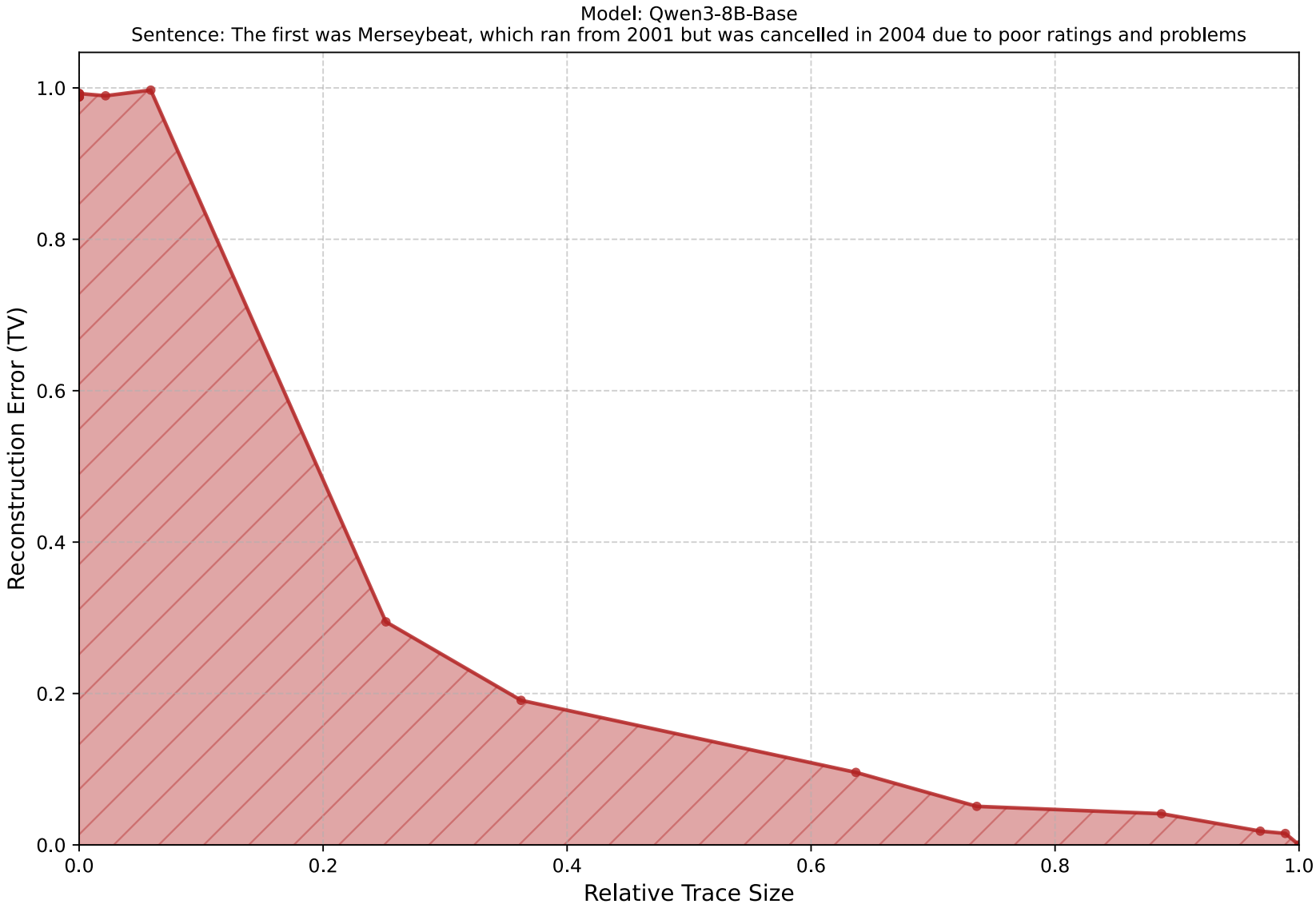


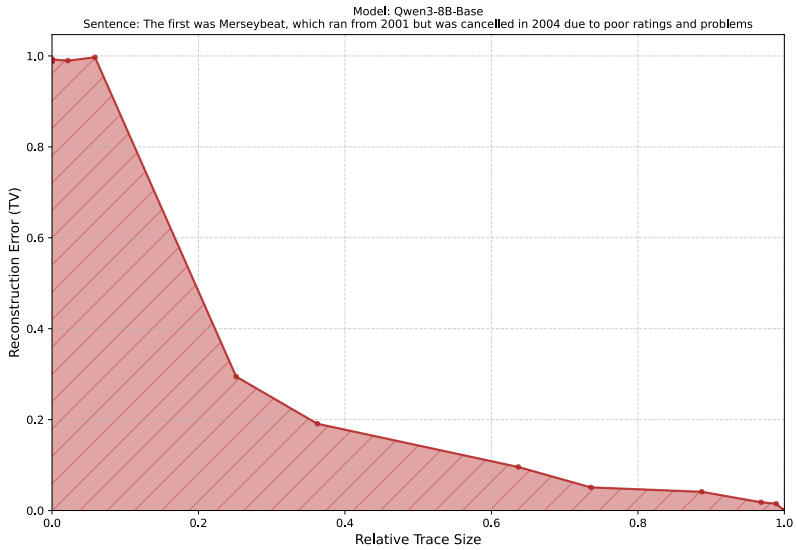
Figure by Javier Ferrando

Method



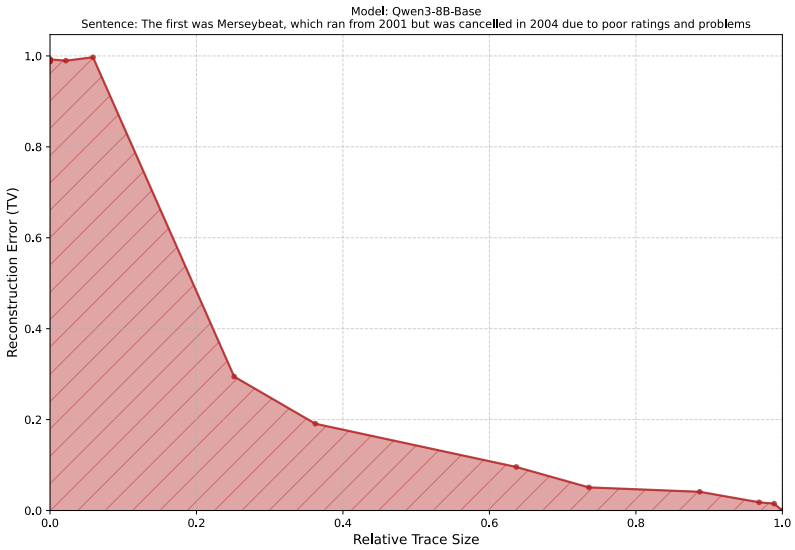
"The first was Merseybeat, which ran from 2001 but was cancelled in 2004 due to poor ratings and problems"

Method

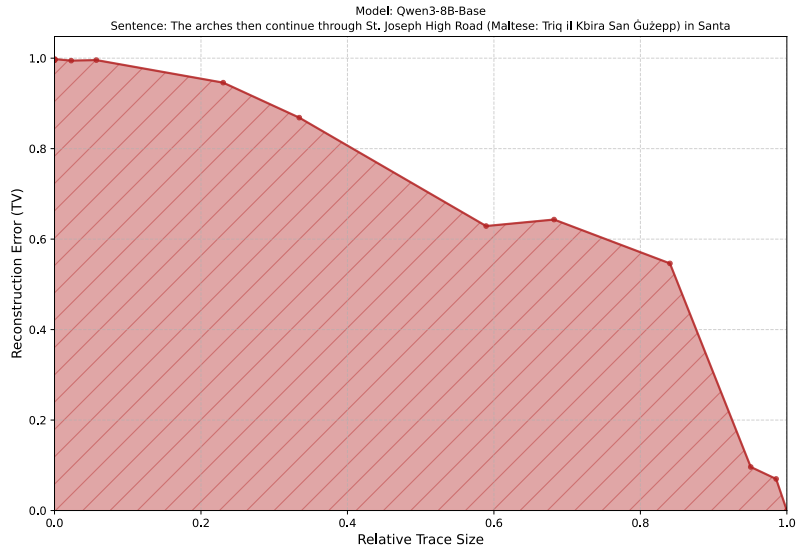


"The first was Merseybeat, which ran from 2001 but was cancelled in 2004 due to poor ratings and problems"

Method

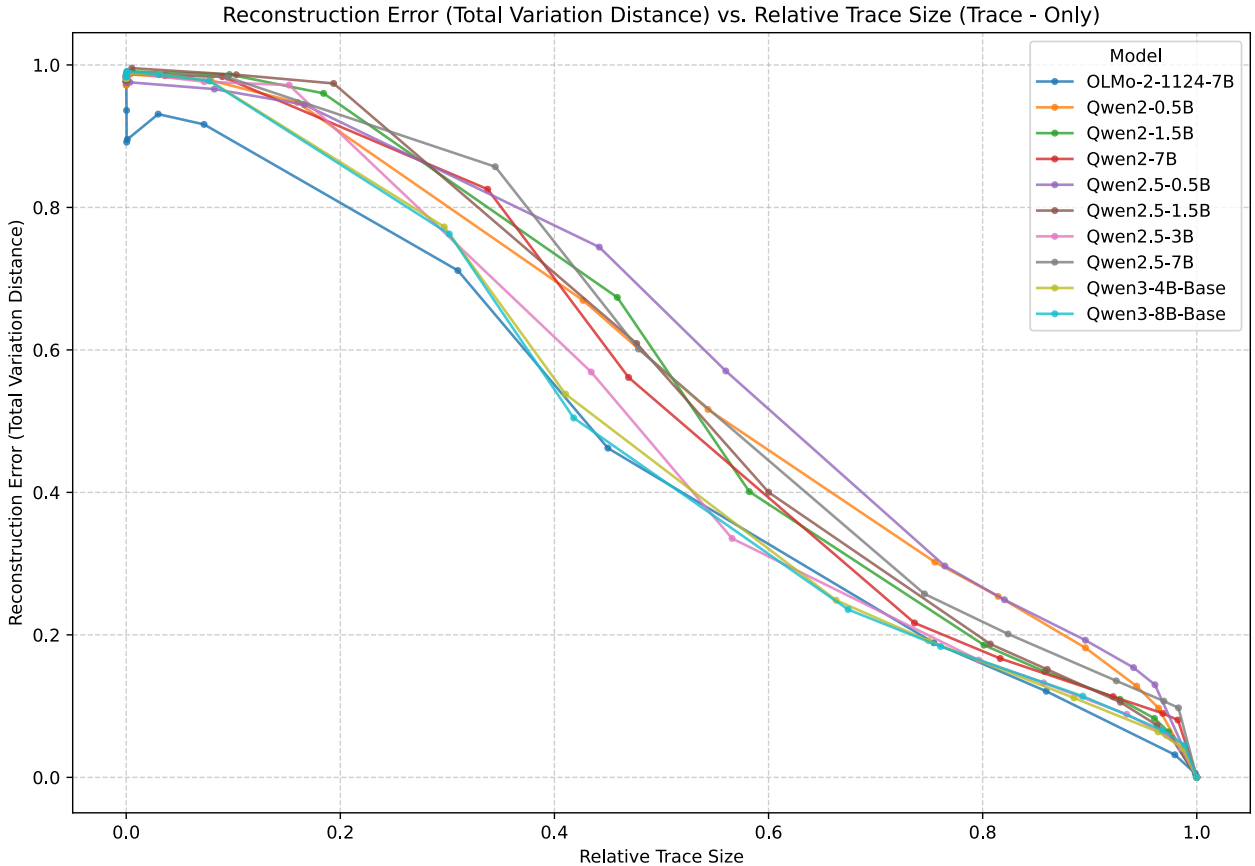


"The first was Merseybeat, which ran from 2001 but was cancelled in 2004 due to poor ratings and problems"

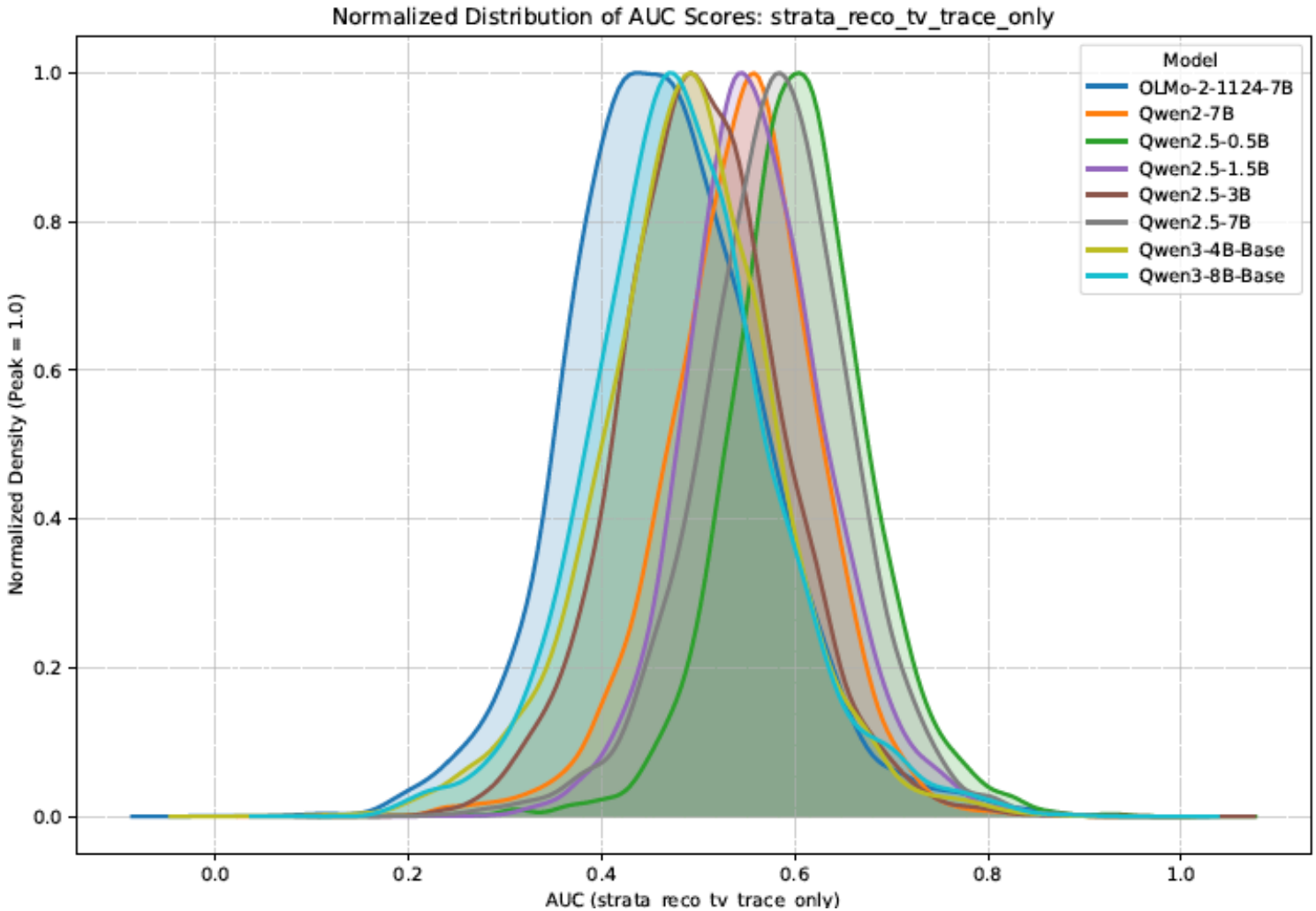


"The arches then continue through St.Joseph High Road (Maltese: Triq il Kbira San Guzepp) in Santa"

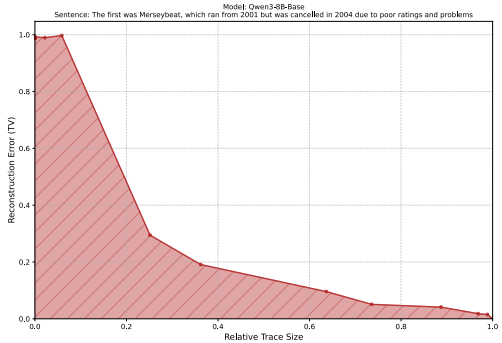
Results: not a lot of evidence for near-symbolic processing

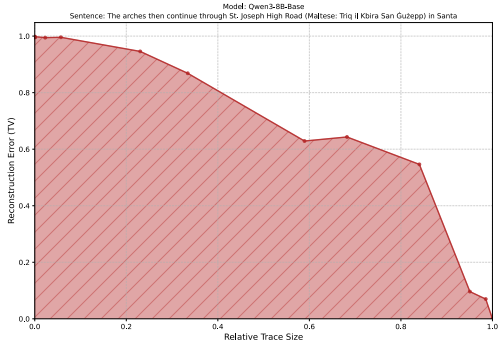


Results: variation



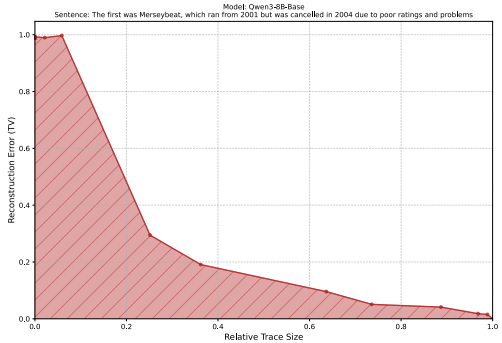
Results: meaningful variation in the trace

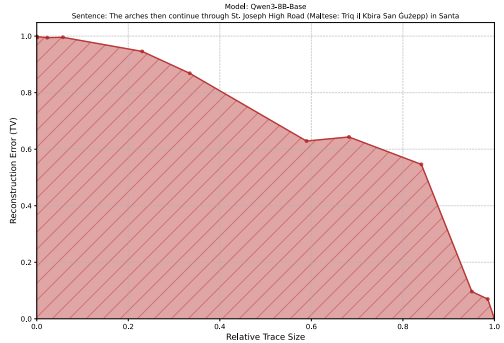




- variation in trace size seems to be meaningful
- substantial agreement among models as to which tokens require more or less computation
  - Pearson correlation in trace size across models: M=0.48, SD: 0.09

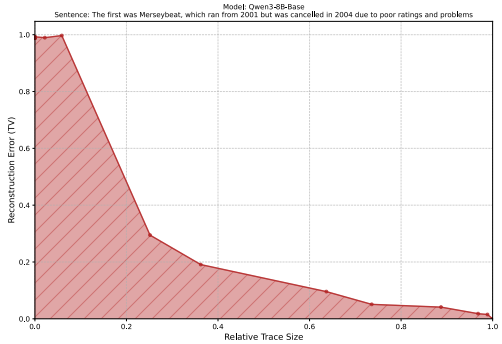
Results: meaningful variation in the trace

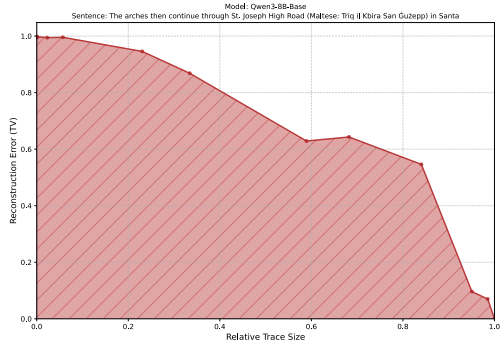




- variation in trace size seems to be meaningful
- substantial agreement among models as to which tokens require more or less computation
  - Pearson correlation in trace size across models: M=0.48, SD: 0.09
- what drives trace size? Could it be that LLMs are indeed sensitive to grammar vs (conceptual) semantics?
  - grammar: more symbolic = smaller trace
  - semantics: more distributed = larger trace

Results: meaningful variation in the trace





- variation in trace size seems to be meaningful
- substantial agreement among models as to which tokens require more or less computation
  - Pearson correlation in trace size across models: M=0.48, SD: 0.09
- what drives trace size? Could it be that LLMs are indeed sensitive to grammar vs (conceptual) semantics?
  - grammar: more symbolic = smaller trace
  - semantics: more distributed = larger trace
- **TBD** (not looking great!)

Plan of the talk

1. the "synthesis view"
   - 2 theses
2. challenges to this view
   - 2 studies and 1 review

Wait a sec. . . how much do we know about syntax in LLMs, really?

Graichen, De-Dios-Flores, Boleda, under review

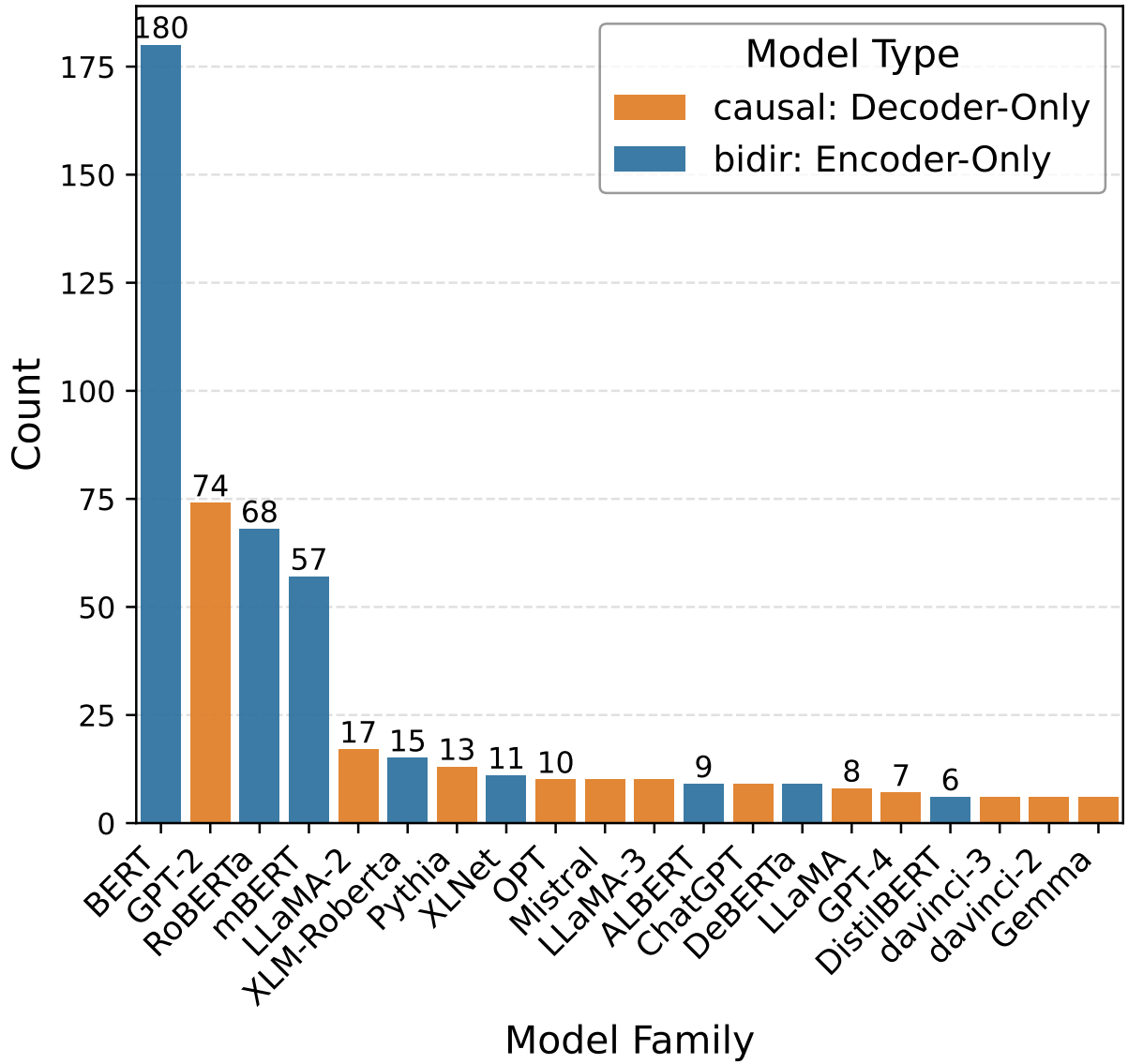
- recall hypothesis:
  - grammar: more symbolic
  - semantics: more distributed
- what do we already know about this?

Wait a sec. . . how much do we know about syntax in LLMs, really?

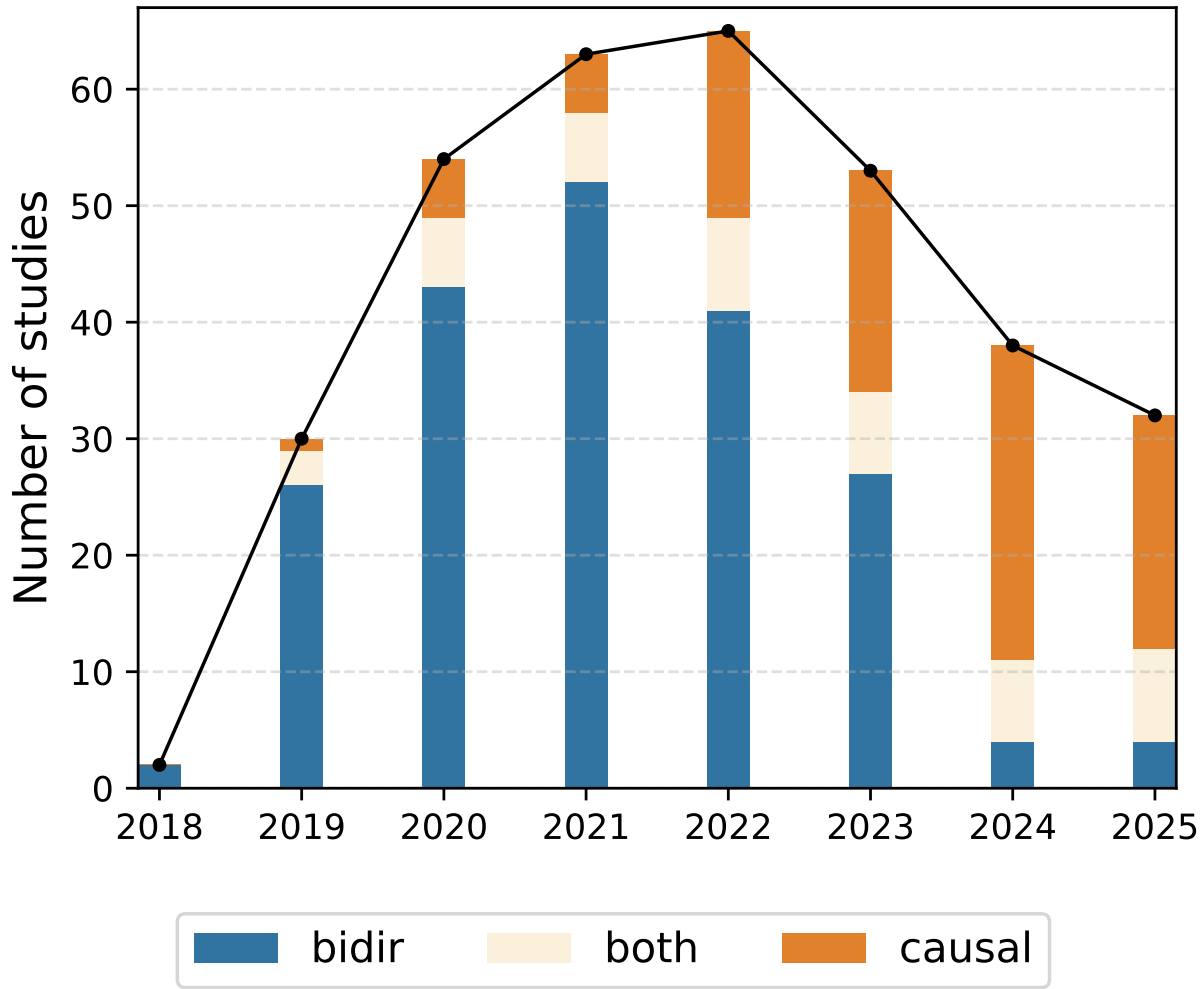
Graichen, De-Dios-Flores, Boleda, under review

- recall hypothesis:
  - grammar: more symbolic
  - semantics: more distributed
- what do we already know about this?
- systematic review
- syntactic abilities of Transformer-based language models
- database: 337 articles, 1015 model results, annotated properties

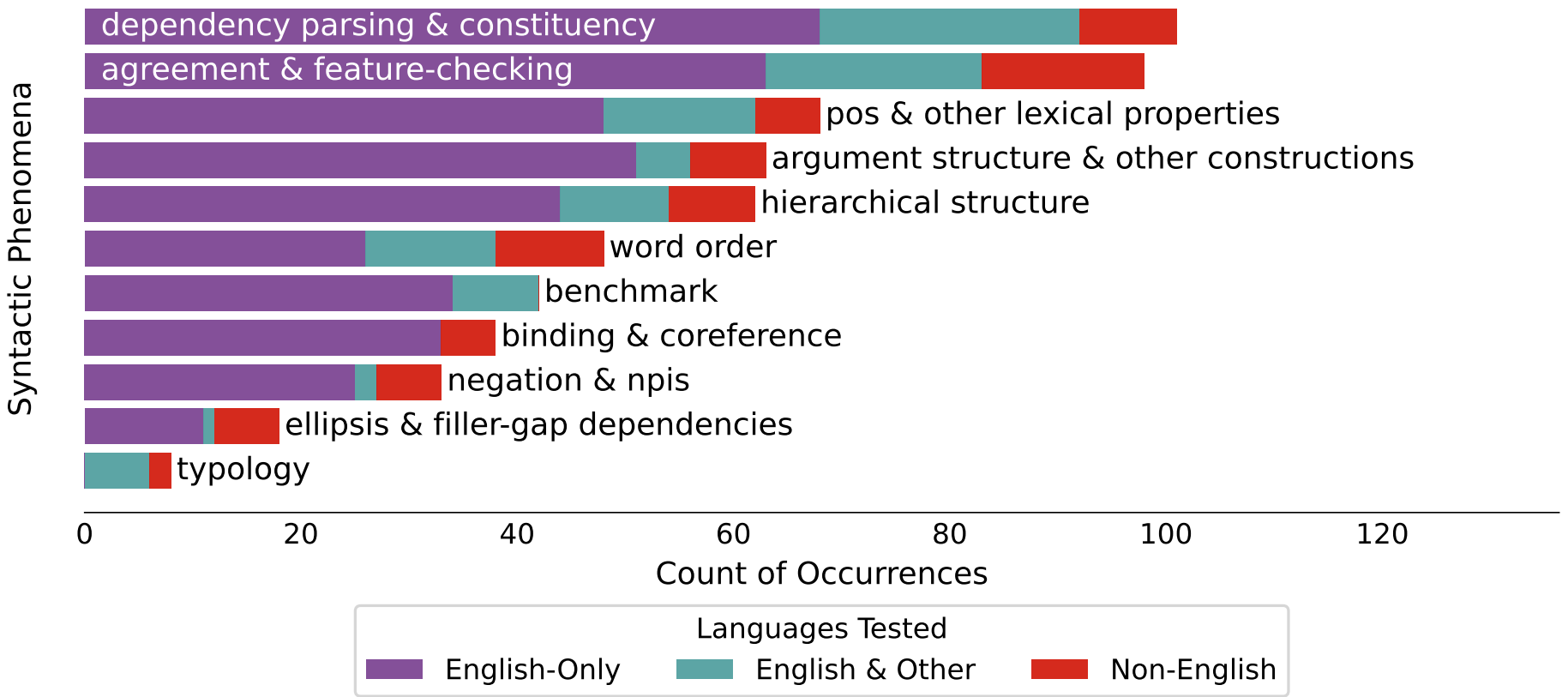
Results: we know quite a bit. . . about BERT :/



Results: evolution by year



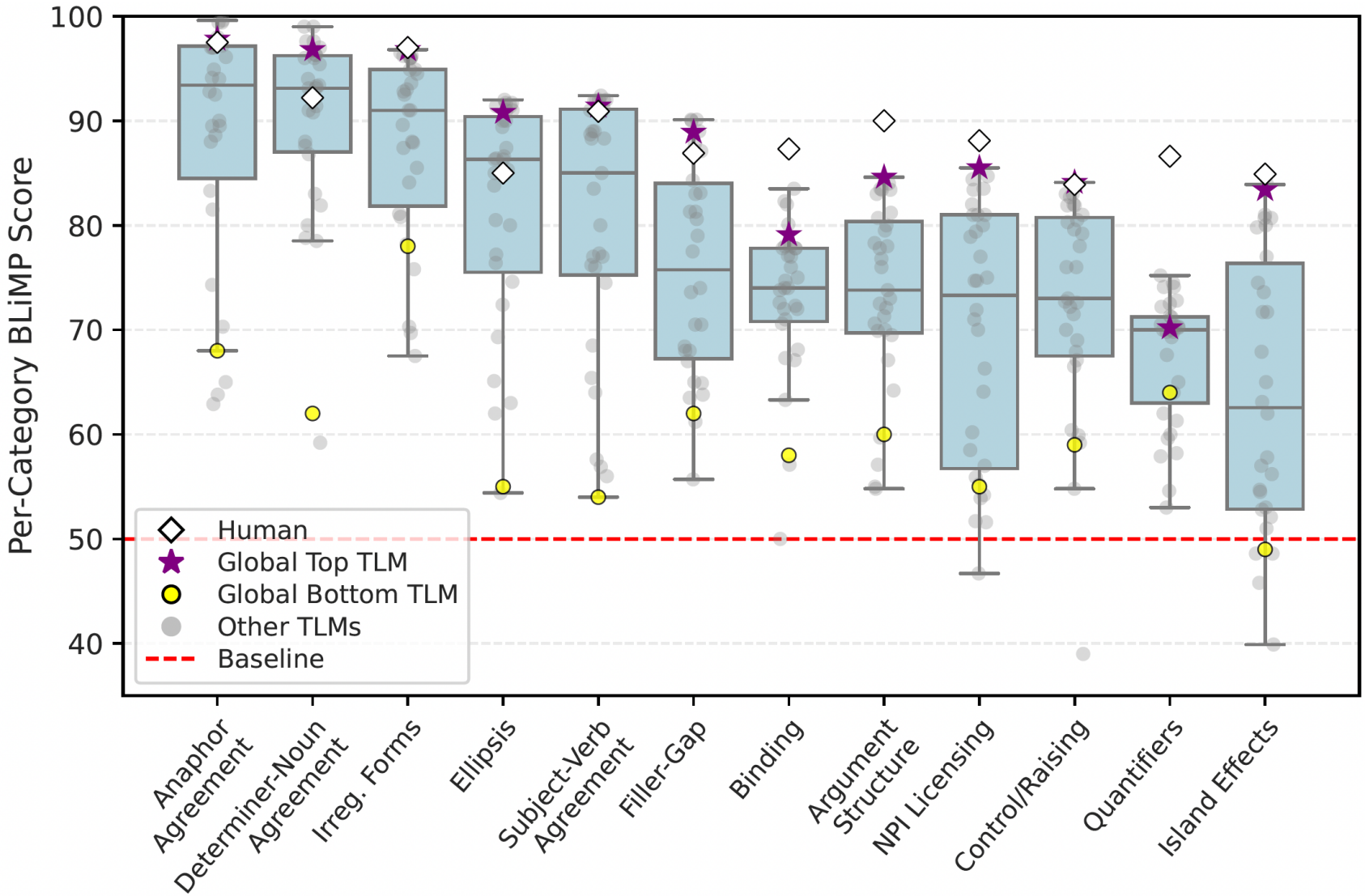
Results: imbalance in languages and phenomena studied



Results: a glimpse at syntactic capabilities

- BLiMP dataset Warstadt et al. (2020) - minimal pairs
- sample stimulus:
  *Brett knew what/*that many waiters find*
- 11 articles, reporting per-category results
- English!

Results: a glimpse at syntactic capabilities



Syntax review: recap

- lots of work on syntax in LLMs
- but narrow in scope!

Syntax review: recap

- languages
  69% of the articles are on English only;
  91% include English

Syntax review: recap

- languages
  69% of the articles are on English only;
  91% include English
- models
  30% of the results concern BERT;
  62%, the four most studied models (BERT, GPT2, RoBERTa, mBERT)

Syntax review: recap

- languages
  69% of the articles are on English only;
  91% include English
- models
  30% of the results concern BERT;
  62%, the four most studied models (BERT, GPT2, RoBERTa, mBERT)
- phenomena
  formal syntax (agreement, parsing, etc.): 50+ articles per category
  syntax-semantics interface: severely understudied

Syntax review: recap

- lots of work on syntax in LLMs
- but narrow in scope!
- so we don't know *that* much about how much syntax LLMs know :/

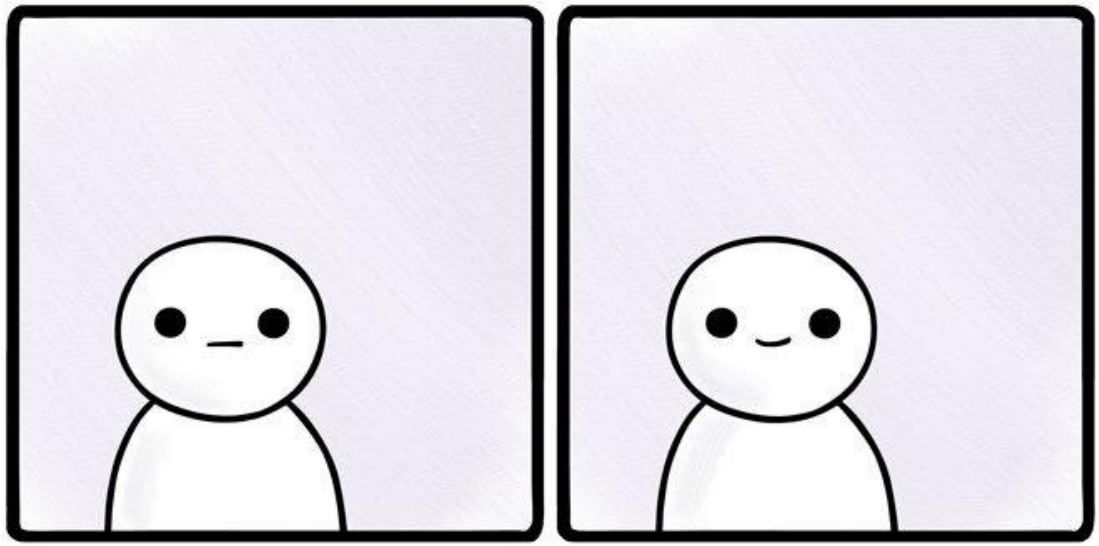
Syntax review: recap

- lots of work on syntax in LLMs
- but narrow in scope!
- so we don't know *that* much about how much syntax LLMs know :/
- let alone how they encode syntactic knowledge!

Plan of the talk

1. the "synthesis view"
   - 2 theses
2. challenges to this view
   - 2 studies and 1 review
3. conclusion

Where I started



Where I started

- language: both regular and messy
- symbolic representations good for regularities;
- distributed representations good for messiness
- LLMs exhibit both—plus flexibility

$\Rightarrow$ they are a synthesis between symbolic and distributed approaches!

Where I started

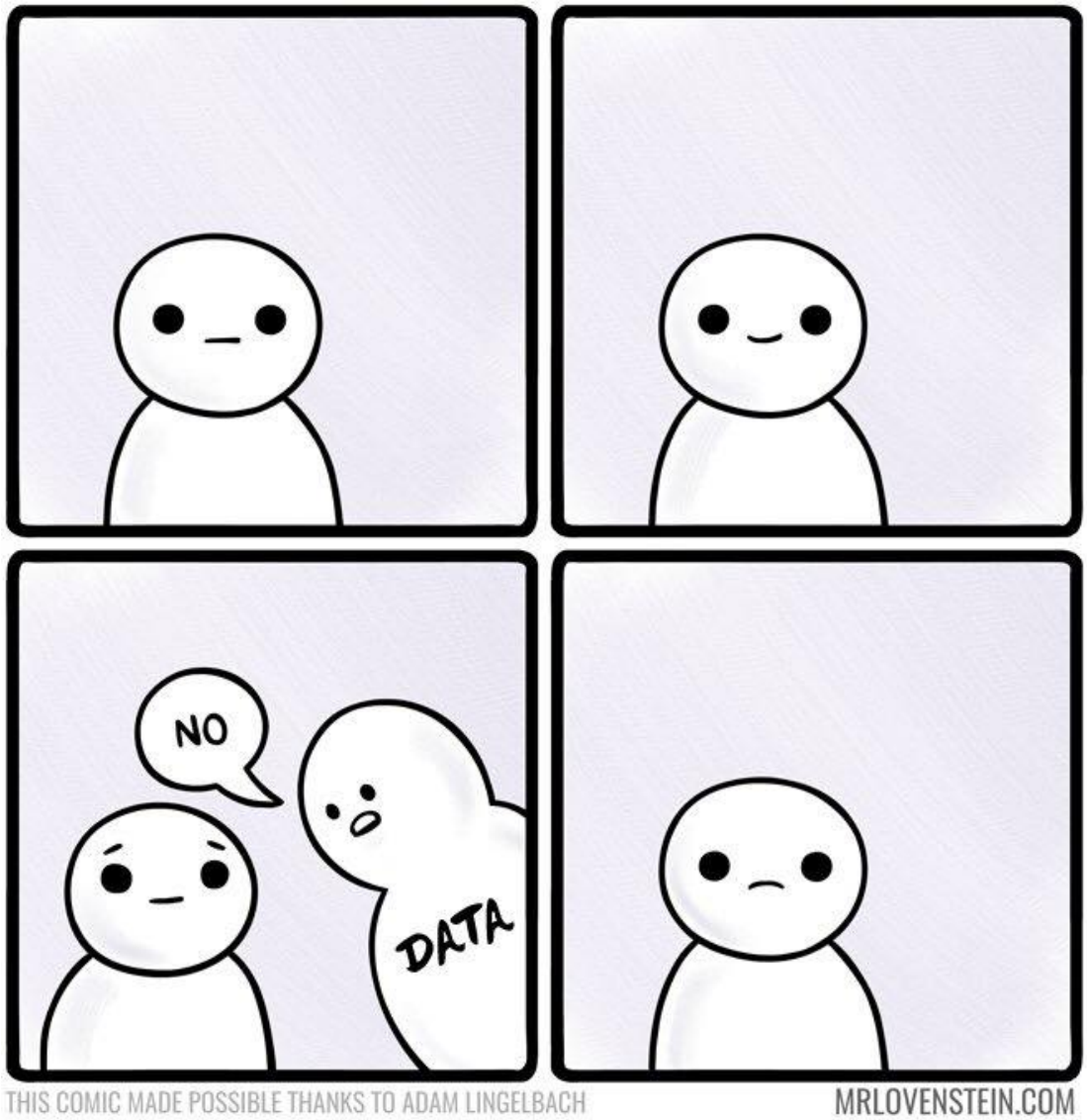
- language: both regular and messy
- symbolic representations good for regularities;
- distributed representations good for messiness
- LLMs exhibit both—plus flexibility

$\Rightarrow$ they are a synthesis between symbolic and distributed approaches!

- these representations / behavior:
- response to pressure from language data
- differential encoding of different phenomena

$\Rightarrow$ clear candidates: syntax vs semantics

Where I am now



Where I am now

- Macocco et al. (2025): LLMs use their near-symbolic power not for grammar, but for frequency

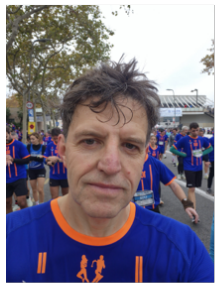
  (sometimes)
- (very tentatively) no evidence for pervasive near-symbolic processing anyway
- and the community has a very partial view of how LLMs process syntax

  —an important linguistic domain where near-symbolic processing is expected

The good news

LLMs as a synthesis between symbolic and distributed approaches to language?

Gemma Boleda & collaborators:



Marco Baroni

Iria De-Dios-Flores

Nora Graichen

Corentin Kervadec

Iuliia Lysova

Iuri Macocco

References

G. Boleda. LLMs as a synthesis between symbolic and distributed approaches to language. In *Findings of the ACL: EMNLP 2025*, pages 9365–9379, 2025. URL https://aclanthology.org/2025.findings-emnlp.498/.

R. Futrell and K. Mahowald. How linguistics learned to stop worrying and love the language models, 2025. URL https://arxiv.org/abs/2501.17047.

I. Macocco, N. Graichen, G. Boleda, and M. Baroni. Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models. In *Proceedings of the 8th BlackboxNLP Workshop*, pages 109–136, 2025. URL https://aclanthology.org/2025.blackboxnlp-1.6/.

C. D. Manning. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707, 2015. URL https://doi.org/10.1162/COLI_a_00239.

A. Warstadt and S. R. Bowman. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press, 2022.

A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *TACL*, 8:377–392, 2020. URL https://aclanthology.org/2020.tacl-1.25/.