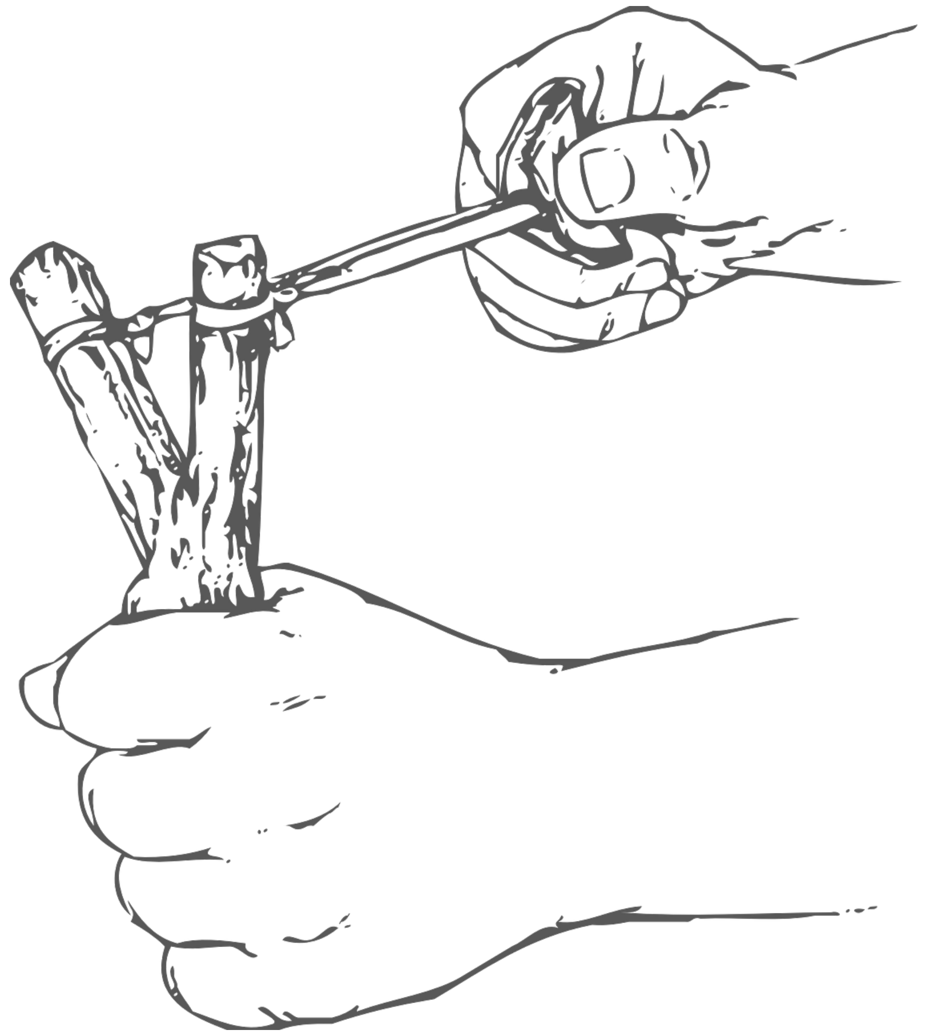
Implications for Linguistics

LLMs: Implications for Linguistics, Cognitive Science & Society

Polina Tsvilodub & Michael Franke, Session 4

Learning goals



1. dive into "**BERTology**"
   - what LLMs "know" about language
   - how LLMs represent "knowledge of language" to do what they do
2. get acquainted with different techniques of "**unblackboxing**"
   a. transfer learning
   b. simple probing (diagnostic classification)
   c. counterfactual probing
   d. targeted behavioral assessment

3. develop opinions about whether LLMs are cognitively plausible or "human-like"

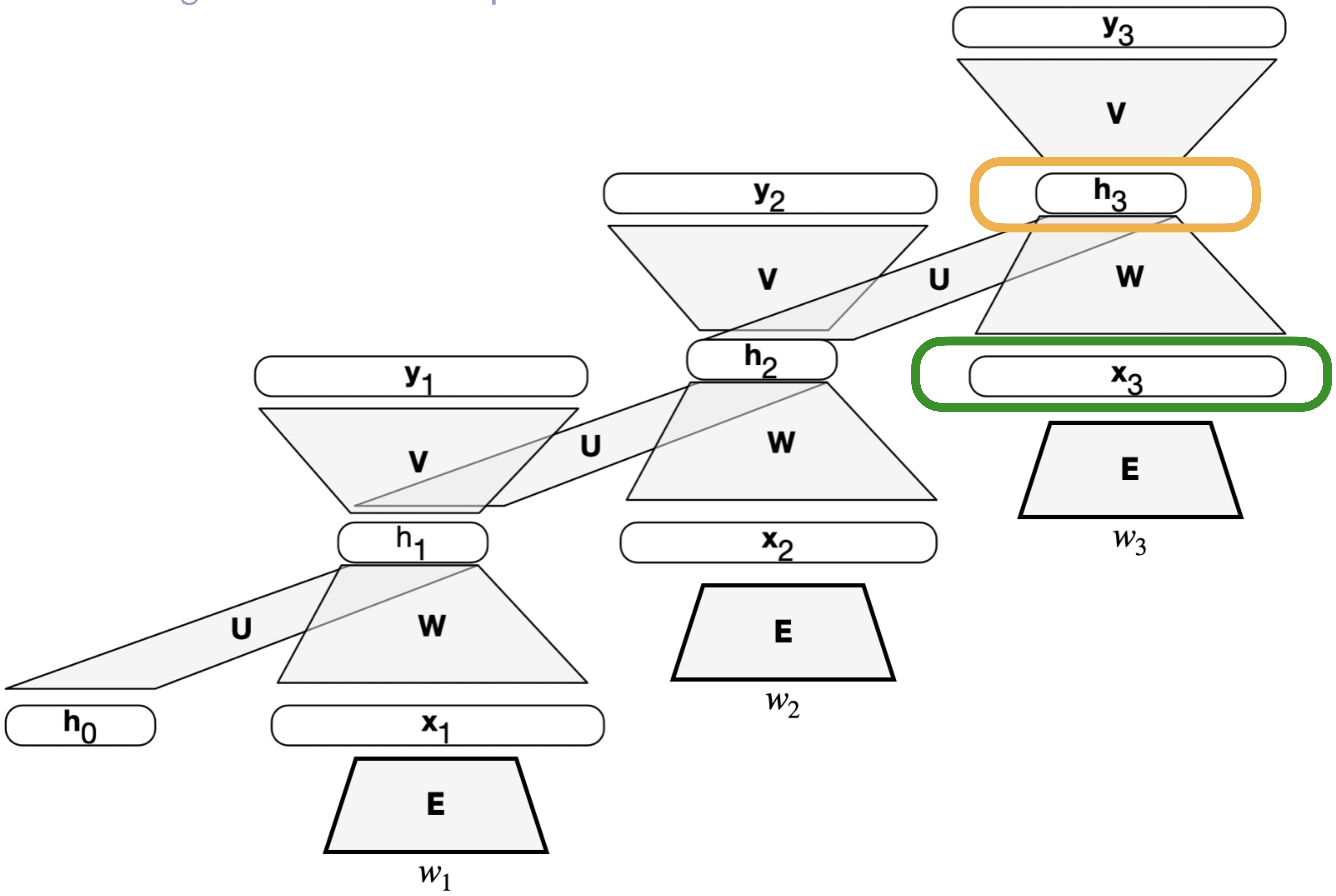
Kicking the elephant out of the room

- human linguistic abilities are much richer and more multi-modal than text input and output
  - intonation, pauses, mimicry, gesture, distance to interlocutor, long-term memory of past interactions, conventional pacts, ...
- nevertheless we want to know what "linguistic abilities" the systems have
  - "LLM-ology" studying machines as a part of (the new) nature with the usual scientific methods

Transfer learning

Recap:

embeddings for words and sequences

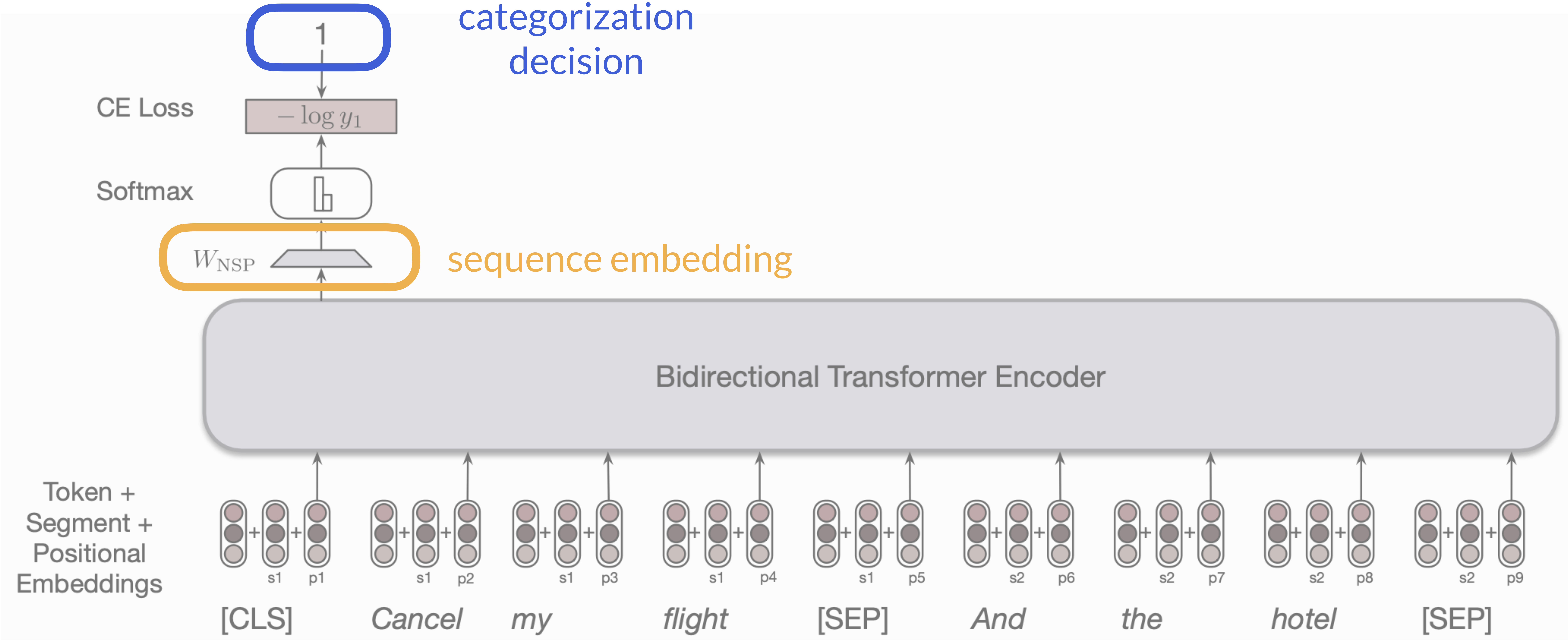


sequence embedding for w_1, w_2, w_3

word embedding for w_3

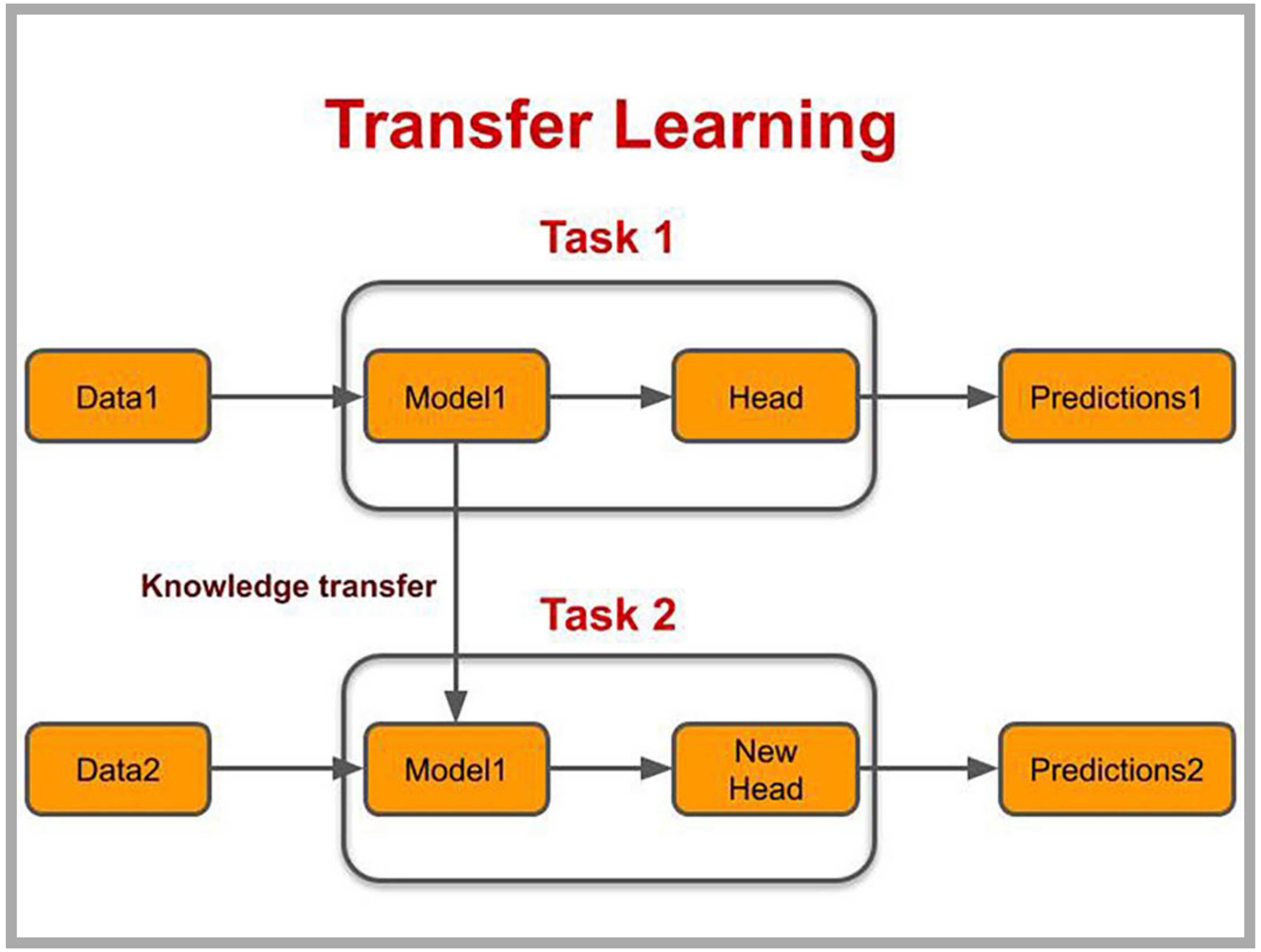
Sequence embeddings

for bi-directional transformers



Pre-training, fine-tuning & transfer learning

- **pretraining:**
  - train model on general large/ huge data set on task T_1
- **fine-tuning:**
  - continue training the model's parameters on a new special case data set on task T_1
- **transfer learning:**
  - apply model pretrained on task T_1 to solve related task T_2
    - option: freeze core model parameters or fine-tune on task T_2

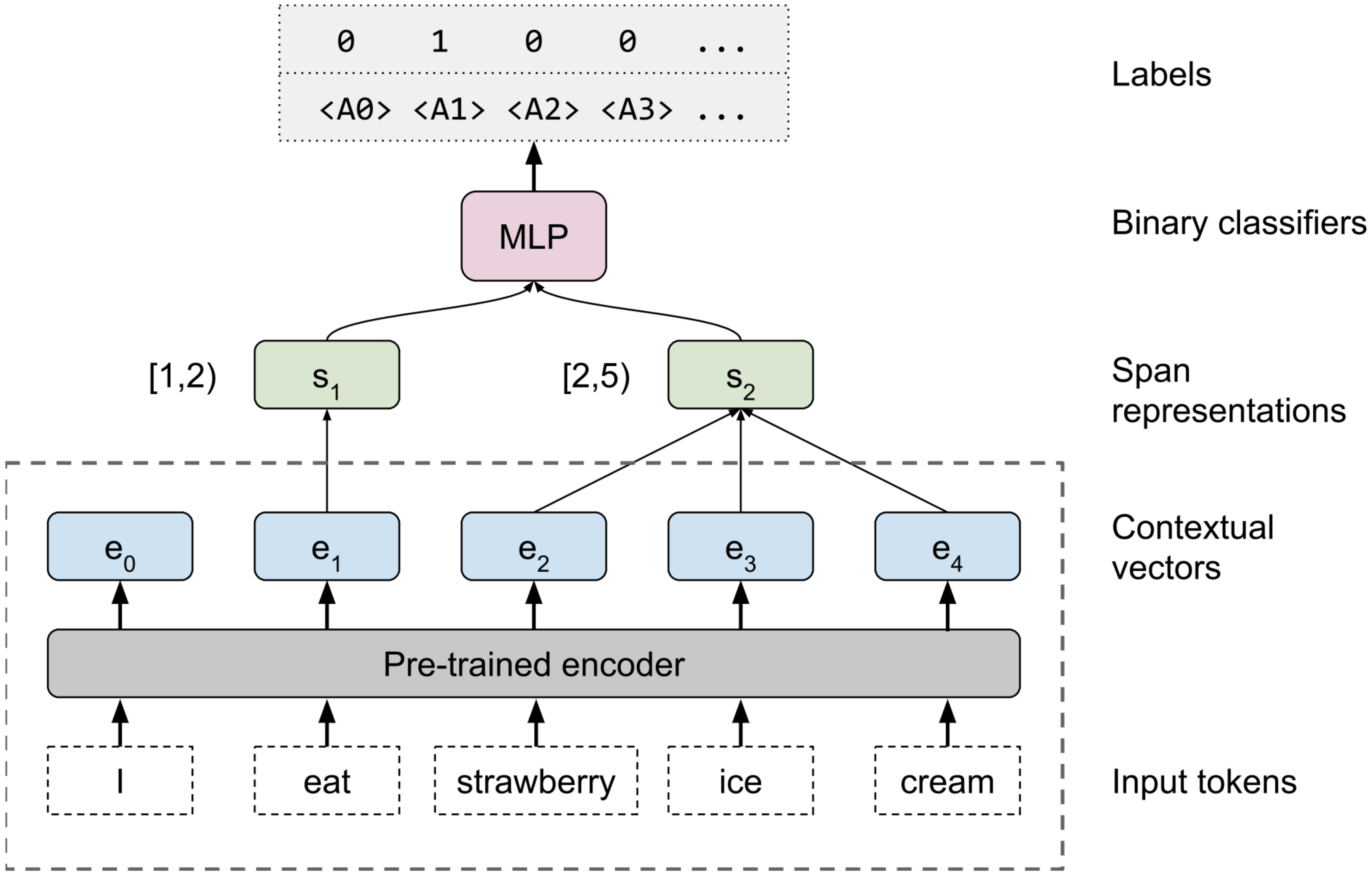


Probing

Probing

aka diagnostic classification

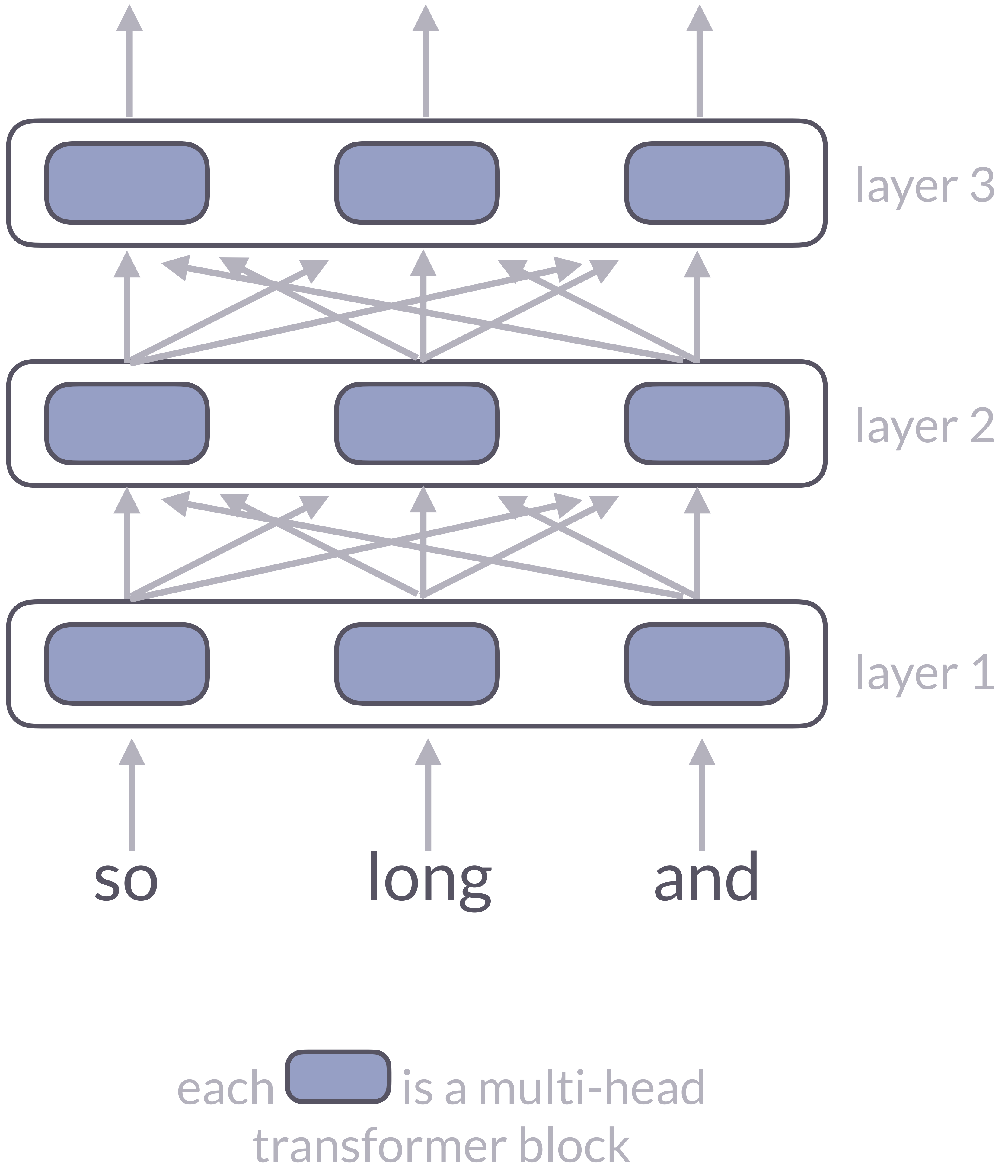
- main idea
  - using transfer-learning w/o fine-tuning to find out which information is contained in different hidden representations
- input:
  - contextual word / span embedding
  - given by LLM
- classifier:
  - linear regression model
  - feedforward neural network (MLP)



Tenney, Xia et al. (2019)

Where is what in BERT?

- **target models:**
  - BERT-base & BERT-large
- **research question:**
  - where (in the hierarchy of transformer layers) is which kind of information processed?
- **method:**
  - edge probing (Tenney, Xia et al 2019)
  - eight tasks:
    - syntax (or low-level semantic):
      part-of-speech, constituents, dependencies, entities
    - (high-level) semantic:
      semantic role labeling, coreference, semantic proto-roles, semantic classification



Tenney, Das & Pavlick (2019) "BERT Rediscovers the Classical NLP Pipeline" *ACL*

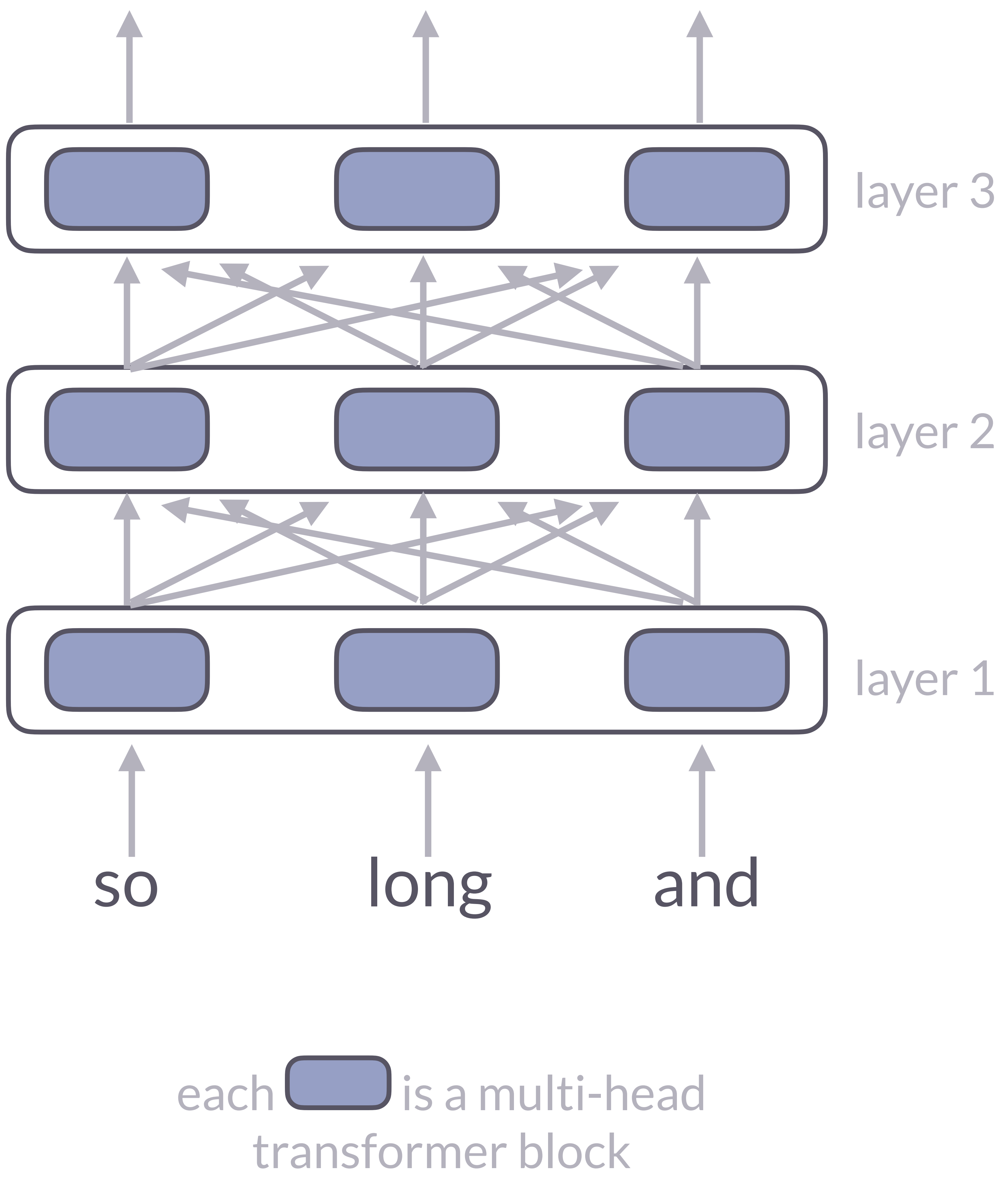
Scalar mixing weights

which layers to combine information from

- consider L layers of stacked embeddings $H^{(0)}, \ldots, H^{(L)}$ (e.g., from BERT)
- given input $w_1, \ldots, w_n$ take vector $\left[\mathbf{h}_0^{(l)}, \ldots, \mathbf{h}_n^{(l)}\right]$ of word embeddings at layer l
- given vector $[s_0, \ldots, s_L]$ of **scalar mixing weights** compute per-token representation vector for w_i as:

$$\mathbf{h}_i = \sum_{l=0}^{L} s_l \mathbf{h}_i^{(l)}$$

- train $[s_0, \ldots, s_L]$ together with MLP classifier

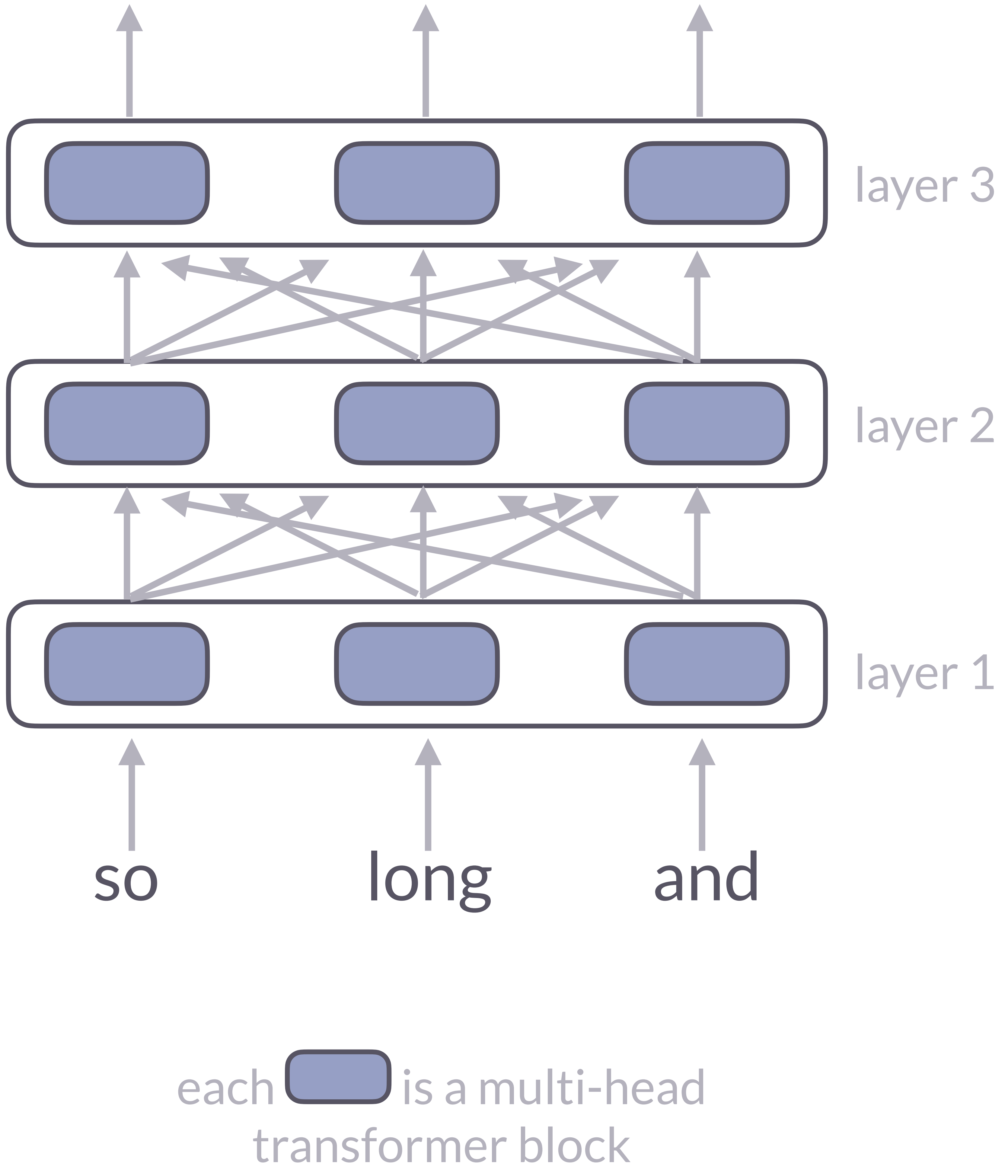


Cumulative scoring

predictive benefit of adding each subsequent layer

- train sequence of classifiers $\{P^{(l)}\}_L$ such that $P^{(l)}$ looks at layer l and below
- cumulative scoring is the difference in F1-score between subsequent classifiers

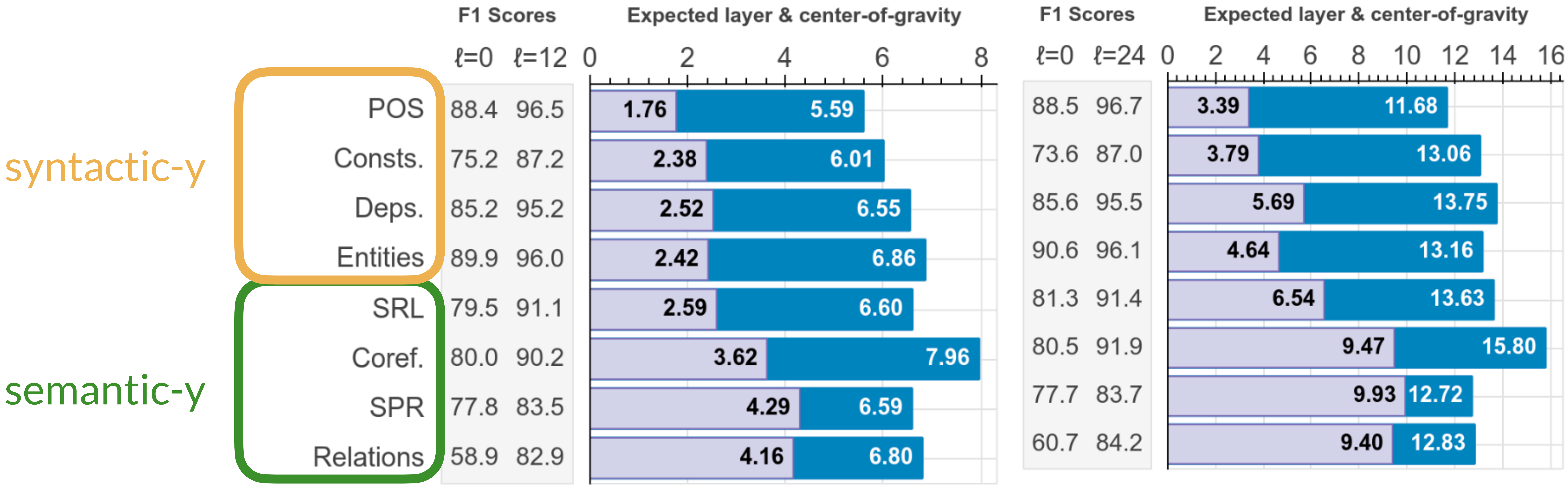
$$\Delta^{(l)} = \text{Score}(P^{(l)}) - \text{Score}(P^{(l-1)})$$



Results

summary statistics

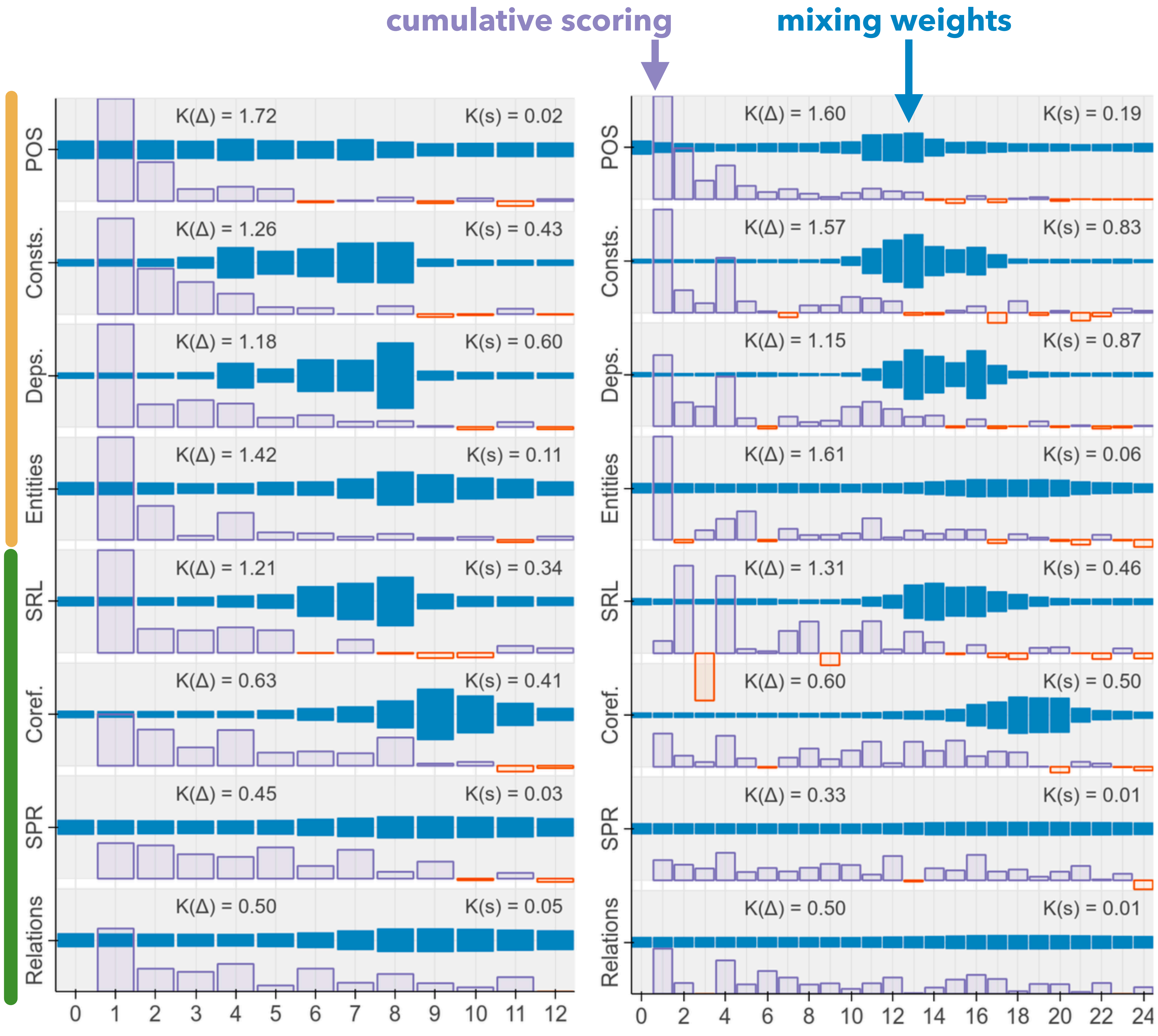
- syntactic information processed earlier in the network than high-level semantic information
  - mixing weights center-of-gravity by layer
  - (pseudo-)expected layer at which model succeeds in classification



(a) BERT-base (b) BERT-large

Results
per layer

- syntactic information processed earlier in the network than high-level semantic information
- syntactic information processing is more localizable / less spread out
- high weights tend to appear with or right after last large delta increase



(a) BERT-base (b) BERT-large

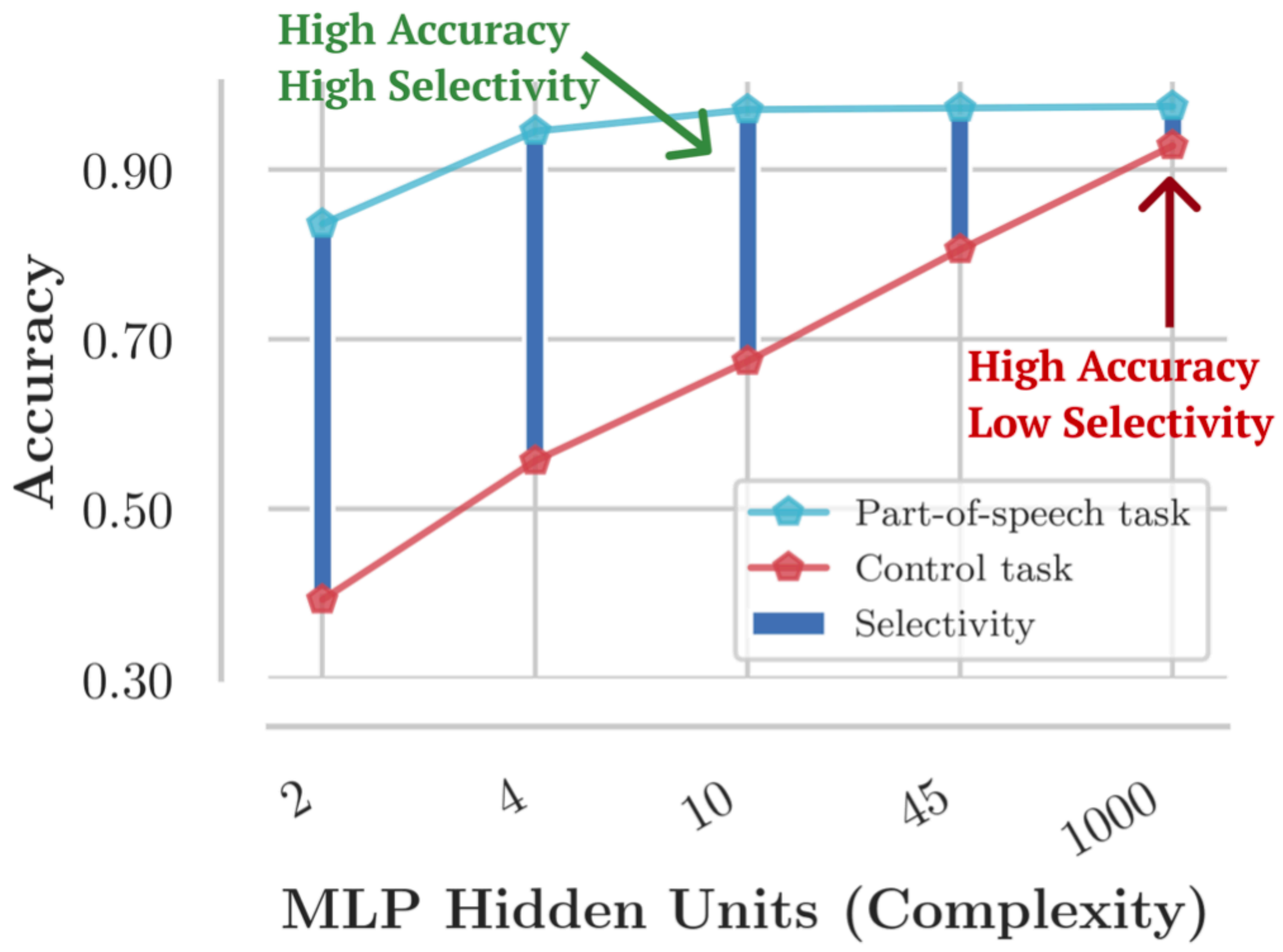
Limitations of probing studies

[O]ur work carries the limitations of all inspection-based probing: the fact that a linguistic pattern is not observed by our probing classifier does not guarantee that it is not there, and the observation of a pattern does not tell us how it is used. For this reason, **we emphasize the importance of combining structural analysis with behavioral studies** (...) to provide a more complete picture of what information these models encode and how that information affects performance on downstream tasks.

Tenney, Das & Pavlick (2019) "BERT Rediscovers the Classical NLP Pipeline" *ACL*

Probe accuracy vs. selectivity

- **probe accuracy:**
  - how well a probe can perform the classification task
- **problem:**
  - models can achieve high accuracy also on entirely random control tasks
- **selectivity:**
  - difference between probe accuracy and accuracy on control task
- **results:**
  - higher selectivity for less powerful classifiers
    - small hidden layer MLPs or linear regression models
  - higher selectivity at deeper layers



Hewitt & Liang (2019) "Designing and Interpreting Probes with Control Tasks" *EMNLP* (video for talk)

Think break

1. How useful a tool is probing to learn about which information is stored where in a neural architecture?
2. What are benefits? What are problems?
3. How could we do better?

Intervention

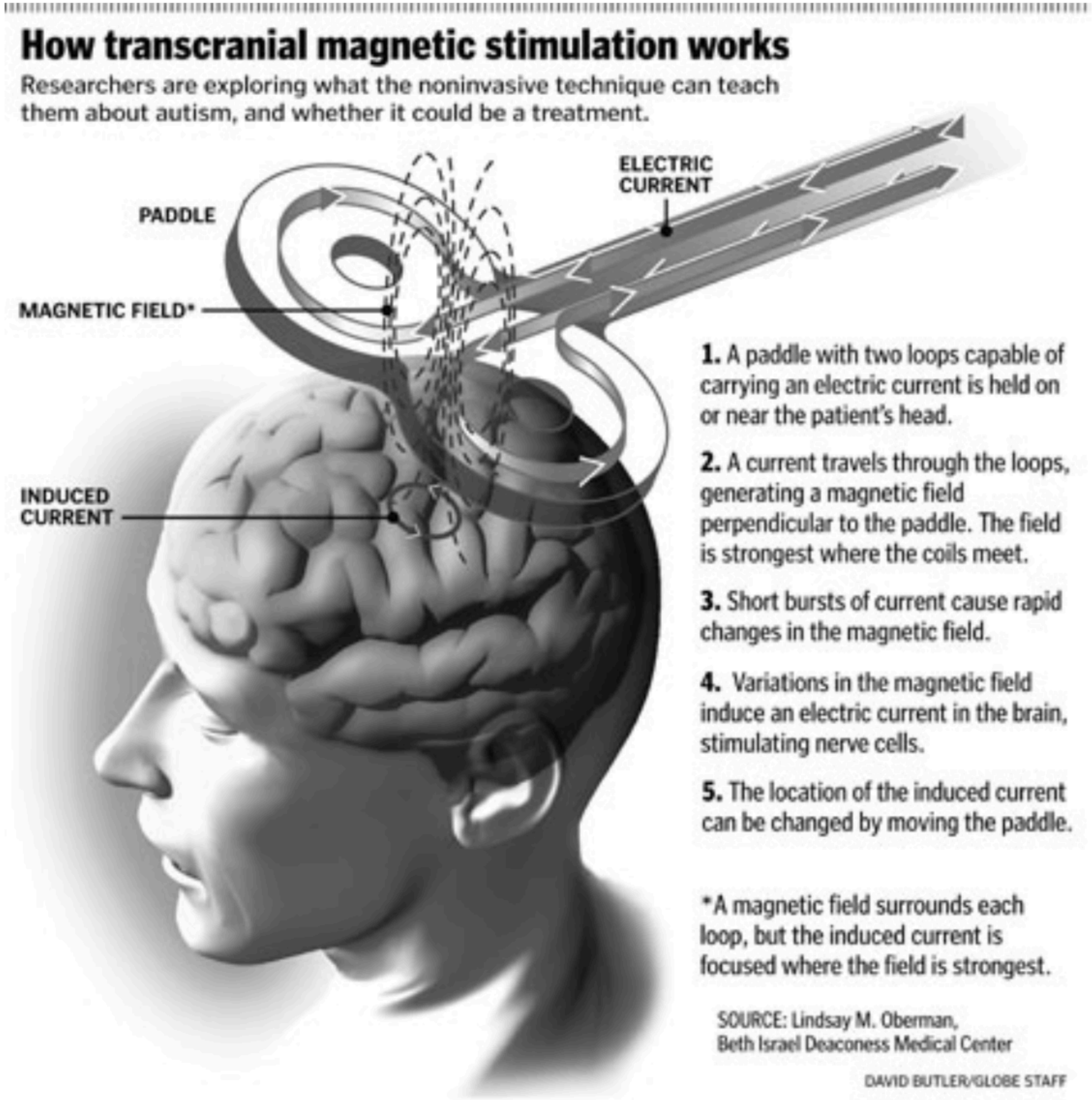
Excursion

cognitive neuroscience

trans cranial magnetic stimulation

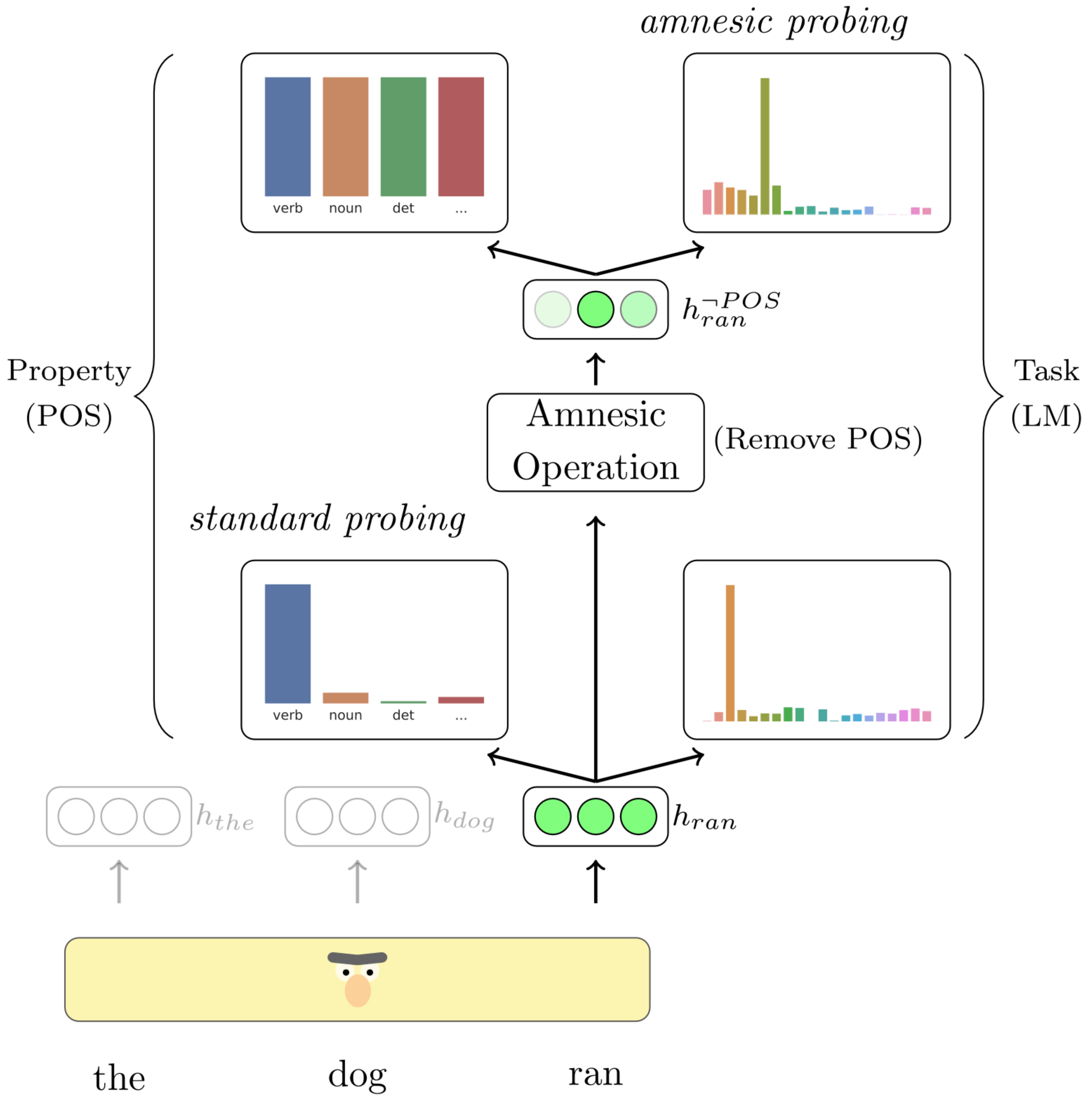
artificial lesioning: short, local disruption of neural activity

allows proper **causal inference** in function ascription



Amnesic probing in neural networks

- systematically intervene with the normal feedforward prediction of a trained model
- check what happens to relevant task performance
- interventions can take place at different locations:
  - input space (Goyal et al. 2019)
  - specific units (Vig. et al 2020)
  - embedding layers (Elazar et al. 2021)

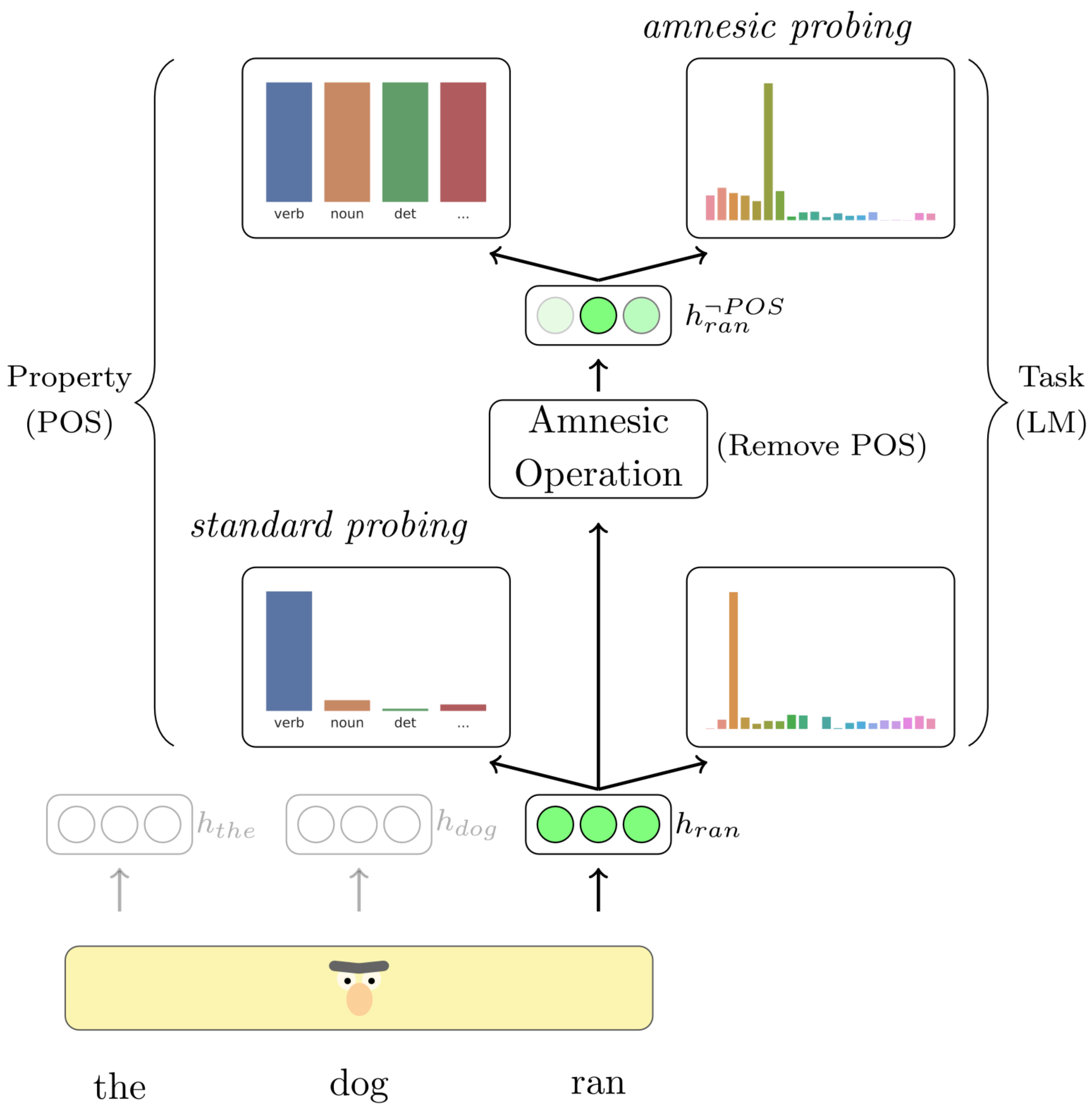


Elazar & al. (2021) "Amnesic probing: Behavioral Explanations with Amnesic Counterfacturals" *TACL*

Iterative null-space projection

Rafvogel et al. (2020)

- sketch of procedure:
  - train a sequence of linear classifiers (SVMs) for task T
  - iteratively remove information useable by classifier for the task
  - terminate when predictive accuracy is at chance level
- include controls (similar amount of deletion but in more arbitrary direction)
  - information
  - selectivity



Elazar & al. (2021) "Amnesic probing: Behavioral Explanations with Amnesic Counterfacturals" *TACL*

Setup & results

- **properties tested (≅ "removed"):**
  - POS (fine and coarse), dependency labels, named-entity labels, constituency boundaries
- **metrics:**
  - (masked) word prediction accuracy
  - KL-divergence between next-word probability before/after lesion

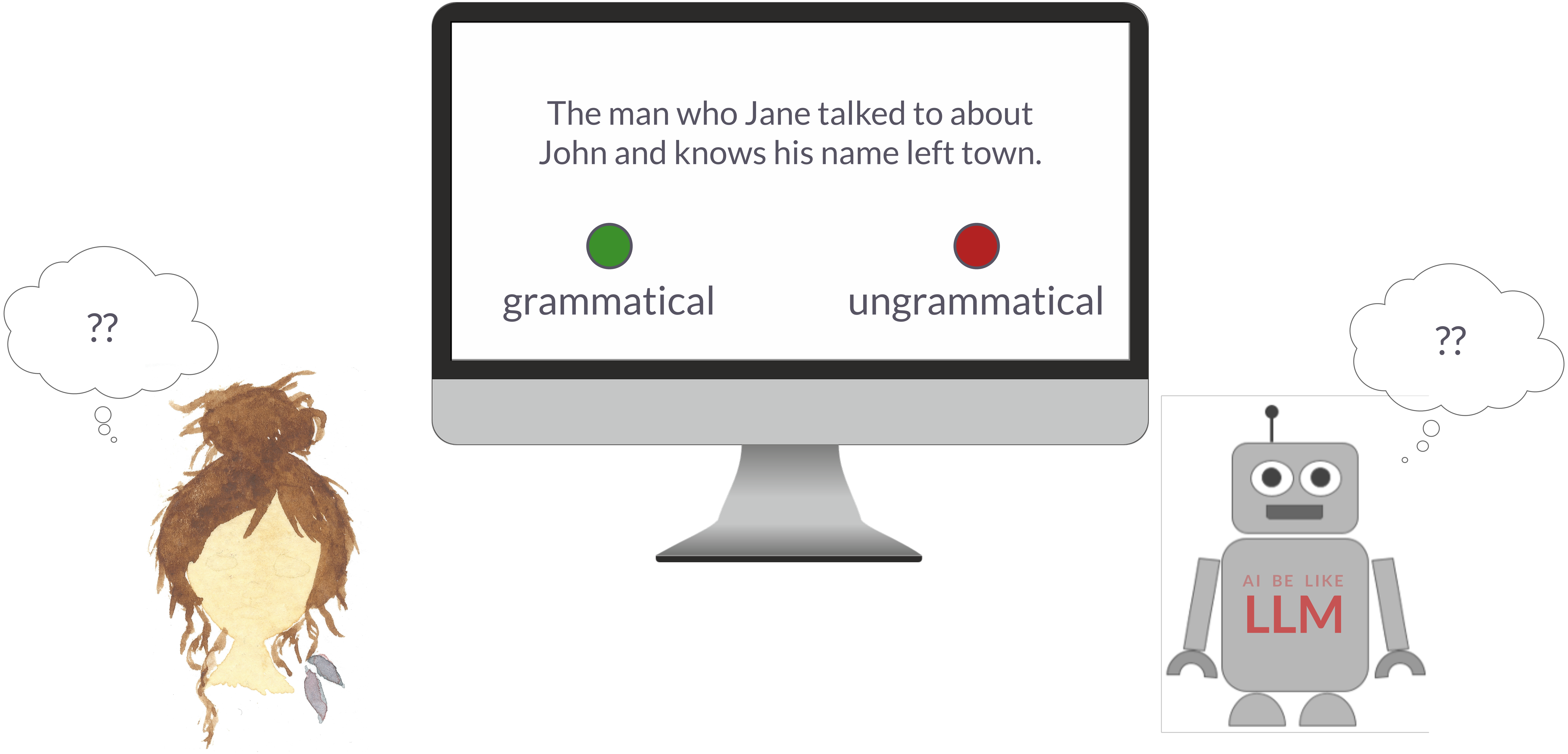
		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Summary
probing

- **probing** resembles transfer-learning, but asks a theoretical question: is there information relevant to task T extractable by a (linear/non-linear) classifier
  - results need to be interpreted with care:
    - better use selectivity than pure accuracy
    - may not be informative about causal role in main task performance
- **intervention / amnesic probing** erases property-specific information (extractable by a linear classifier) and can therefore study how much ("linear") use the model makes of property-specific information
  - does give insights into causal role of property-specific information but …
  - model could still extract information non-linearly

Targeted assessment

Behavioral experiments
w/ minds & machine



Targeted behavioral assessment

- **research question:**
  - does model M accurately predict
    - human (offline) grammaticality judgements and/ or
    - human (online) processing data?
- **method:**
  - curated test suites (informed by theoretical linguistics & psycholinguistics)
    - e.g., benchmark data set BLiMP (Warstadt et al. 2020)
  - derive model predictions from pre-trained models
  - compare against armchair judgements or actual human data

“Targeted Syntactic Evaluation of LMs”

- three LMs are compared against each other and human data
  - n-gram baseline
  - RNN trained on unannotated data
  - same RNN but with additional CCG supertagging
- test set: ~350k automatically generated sentence pairs
  - generated with a non-recursive context-free grammar
- focus on three phenomena:
  - (i) subject-verb agreement, (ii) reflexive anaphora and (iii) negative polarity
- main findings:
  - performance on training data tracks performance of predicting human grammaticality judgements
  - n-gram baseline < simple RNN < multi-trained RNN

resources

- paper
- code
- video

Marvin & Linzen (2018) EMNLP

Test sentence pairs: SV-Agreement 1

- simple agreement
  - The author laughs.
  - * The author laugh.
  - The authors laugh.
  - * The authors laughs.
- agreement in a sentential complement
  - The bankers knew the officer smiles.
  - * The bankers knew the officer smile.
  - ...
- agreement across a prepositional phrase
  - The farmers near the parents smile.
  - * The farmers near the parents smiles.
  - ...

Marvin & Linzen (2018) EMNLP

Test sentence pairs: SV-Agreement 2

- agreement across a subject relative clause
  - * The officers that love the skater smile.
  - * The officers that love the skater smiles.
  - ...
- short VP coordination
  - * The senator smiles and laughs.
  - * The senator smiles and laugh.
  - ...
- long VP coordination
  - * The manager writes a letter every day and likes sweets.
  - * The manager writes a letter every day and like sweets.
  - ...

Marvin & Linzen (2018) EMNLP

Test sentence pairs: Agreement in object relative clauses

more difficult: model would need to tell two subjects apart

- agreement across object relative clauses
  - The farmer that the parents love <u>swims</u>.
  - * The farmer that the parents love <u>swim</u>.
  - The farmers that the parent loves <u>swim</u>.
  - * The farmers that the parent loves <u>swims</u>.
- agreement within object relative clauses
  - The farmer that the parents <u>love</u> swims.
  - * The farmer that the parents <u>loves</u> swims.
  - The farmers that the parent <u>loves</u> swim.
  - * The farmers that the parent <u>love</u> swim.

Marvin & Linzen (2018) EMNLP

Test sentence pairs: Agreement in object relative clauses

more difficult: model would need to tell two subjects apart

- simple reflexive
  - The senators embarrassed <u>themselves</u>.
  - * The senators embarrassed <u>herself</u>.
  - …
- reflexive in a sentential complement
  - The bankers thought the pilot embarrassed <u>herself</u>.
  - * The bankers thought the pilot embarrassed <u>themselves</u>. gender neutral?
  - …
- reflexive across an object relative clause
  - The manager that the architects like doubted <u>herself</u>.
  - * The manager that the architects like doubted <u>themselves</u>. gender neutral?
  - …

Marvin & Linzen (2018) EMNLP

Test sentence pairs: Negative polarity

- simple NPI
  - No students have ever lived here.
  - * Most students have ever lived here.
- NPI across a relative clause
  - No authors the guards like have ever been famous.
  - * The authors no guards like have ever been famous.

Marvin & Linzen (2018) EMNLP

Human data

- 100 participants (MTurk)
- each participant saw 76 pairs of sentences
- on each trial, participants had to choose the grammatical sentence from the pair (forced-choice task)
- 16 participants were excluded due to more than one error on the simple agreement trials

Marvin & Linzen (2018) EMNLP

Think break

1. Given a language model, how would we determine whether the model can or cannot match human grammaticality judgements for any pair of sentences without training the model on the task?
2. If human participants make mistakes, what should we expect an LM to do? Be equally good as humans, or be at ceiling where humans fail to meet the grammatical norm?

Defining grammaticality prediction

- given a contrast pair of sentences like:
  - No students have ever lived here. [$w_{1:n}$]
  - * Most students have ever lived here. [$v_{1:m}$]
- an LM is said to predict the right grammaticality judgement iff:
  $P_M(w_{1:n}) > P_M(v_{1:m})$

Marvin & Linzen (2018) EMNLP

Results

Marvin & Linzen (2018) EMNLP

	RNN	Multitask	*n*-gram	Humans	# sents
SUBJECT-VERB AGREEMENT:					
Simple	0.94	1.00	0.79	0.96	280
In a sentential complement	0.99	0.93	0.79	0.93	3360
Short VP coordination	0.90	0.90	0.51	0.94	1680
Long VP coordination	0.61	0.81	0.50	0.82	800
Across a prepositional phrase	0.57	0.69	0.50	0.85	44800
Across a subject relative clause	0.56	0.74	0.50	0.88	22400
Across an object relative clause	0.50	0.57	0.50	0.85	44800
Across an object relative (no *that*)	0.52	0.52	0.50	0.82	44800
In an object relative clause	0.84	0.89	0.50	0.78	44800
In an object relative (no *that*)	0.71	0.81	0.50	0.79	44800
REFLEXIVE ANAPHORA:					
Simple	0.83	0.86	0.50	0.96	560
In a sentential complement	0.86	0.83	0.50	0.91	6720
Across a relative clause	0.55	0.56	0.50	0.87	44800
NEGATIVE POLARITY ITEMS:					
Simple	0.40	0.48	0.06	0.98	792
Across a relative clause	0.41	0.73	0.60	0.81	31680

demo

how to get surprisal for text passages out of GPT-3

Syntactic generalization scores

Towards systematic assessment of syntactic generalization

- 10 LMs are compared against each other, of which 5 non-pretrained:
  - n-gram baseline, vanilla LSTM, ordered neurons LSTM, RNNG, GTP-2
- 4 different training set sizes (for non-pretrained models)
  - 1, 5, 14 and 42 million tokens
- test set consists of 34 test suits from 6 “syntactic circuits”
  - (i) garden-path effects, (ii) licensing, (iii) agreement, (iv) center embedding
  - (v) long-distance dependencies, (vi) gross syntactic expectation
- introduce **syntactic generalization (SG) score**
- main findings:
  - dissociation between perplexity and SG score
  - model type has more effect on SG than training data size
  - higher SG scores for models with explicit structural training
  - differences in success on different test suits depends on model type

resources

- paper
- code
- video

Hu et al. (2020), ACL

Syntactic generalization (SG) score

- each test suit has a set of predictions
- SG score for test suit X is the proportion of items in X for which the LM matches all predictions associated with X
- example “garden-path sentences”
  - test item example:
    1. The horse raced past the barn <u>fell</u> …
    2. The horse ridden past the barn <u>fell</u> …
    3. The horse which was raced past the barn <u>fell</u> …
    4. The horse which was ridden past the barn <u>fell</u> …
  - associated predictions:

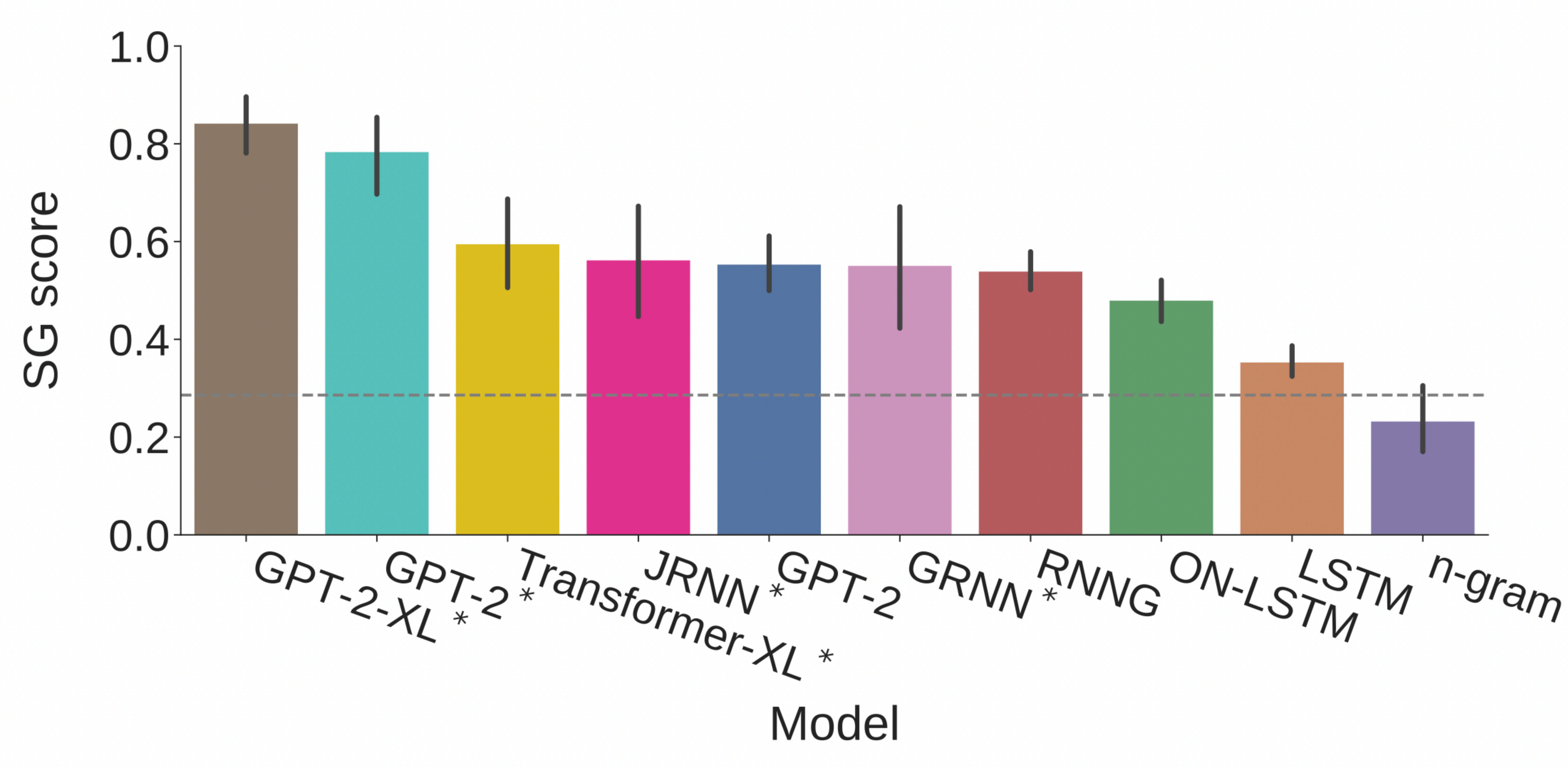
    P(found | (i)) < P(found | (ii))

    P(found | (i)) < P(found | (iii))

    P(found | (i)) - P(found | (ii)) > P(found | (iii)) - P(found | (iv))

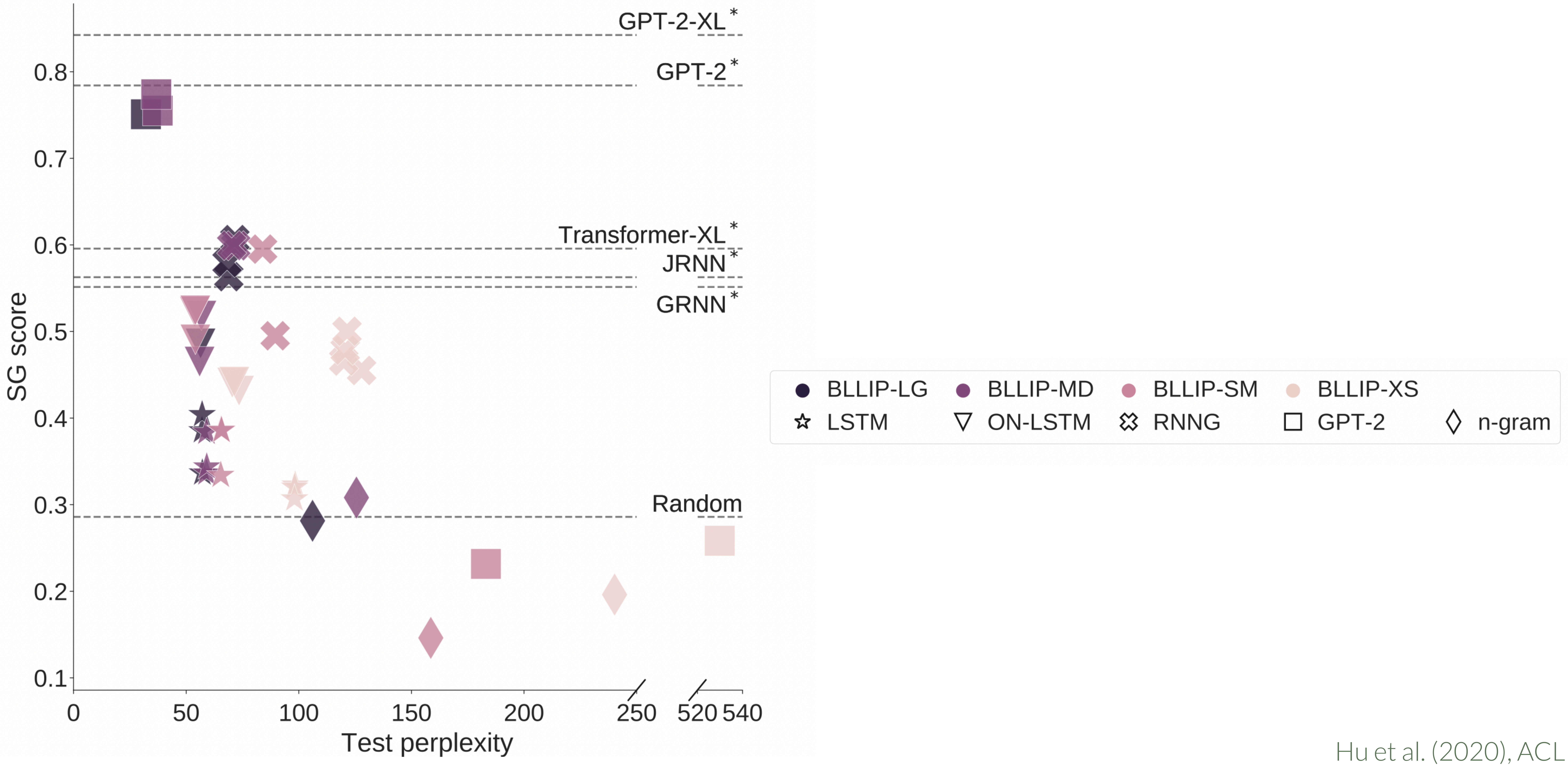
Hu et al. (2020), ACL

Results: Average SG scores by model type



Hu et al. (2020), ACL

Results: Relation SG score vs. perplexity on test set



Hu et al. (2020), ACL

Assessing language processing

Sources of processing difficulty

- limits of working memory

  The dog which the cat which the mouse provoked was chased by barked.

- local ambiguity

  The horse raced past the barn fell.

- interaction w/ semantics & world knowledge

  The cop arrested by the detective was guilty of taking bribes.

Surprisal theory

- **surprisal theory:**
  - $\text{Effort}(w_i, w_{1:i-1}, C) \propto \text{Surprisal}(w_i \mid w_{1:i-1}, C) = -\log P(w_i \mid w_{1:i-1}, C)$
  - compatible with two mechanisms causing processing difficulty:
    - **prediction**: comprehenders actively predict upcoming words; processing difficulty is a form of prediction error
    - **integration**: comprehenders do not actively predict upcoming material, but passive pre-activation leads to easier integration of some material than others
  - empirical evidence for surprisal theory:
    - cloze probability
    - eye-tracked reading
    - self-paced reading
    - EEG during reading
    - maze task

Play break

- go try out the **iMaze task** for yourself:
  - follow this link

Targeted Assessment of Incremental Processing in nLMs & Human

- **language models**:
  - JRNN: large-scale RNN using LSTM units & CNN character embeddings
  - GRNN: from Gulordava et al. (2018)
  - GPT-2: version from `lm-zoo` distribution
  - RNNG: average of three RNNGs from Hu et al. (2020)
- **test set:** 16 test suits adapted from Hu et al. (2020)
- **human data** on sentence processing difficulty: decision times from an **iMaze task**

resources

- paper
- code
- video

Wilcox et al. (2021)

Targeted Assessment of Incremental Processing in nLMs & Human

- measure of interest:
  - *qualitative*: accuracy scores (LM prediction vs armchair grammaticality judgements)
  - *quantitative*: degree of slowdown on critical region (LM prediction vs iMaze data)
  - *generalization*: train linear model to map $P_M(w_i \mid w_{1:i-1}) \mapsto RT_{\text{human}}(w_i \mid w_{1:i-1})$ for each w_i not in a critical region, and use it to explain RTs from words in critical regions
- main findings:
  - *qualitative*: nLMs predict processing difficulty at regions exactly where humans seem to experience it
  - *quantitative*: nLMs are "not surprised enough”
  - *generalization*: nLMs routinely underpredict human RTs / surprisal

resources

- paper
- code
- video

Wilcox et al. (2021)

LLMs and theory of language

Productivity of natural language

"colorless green ideas sleep furiously"

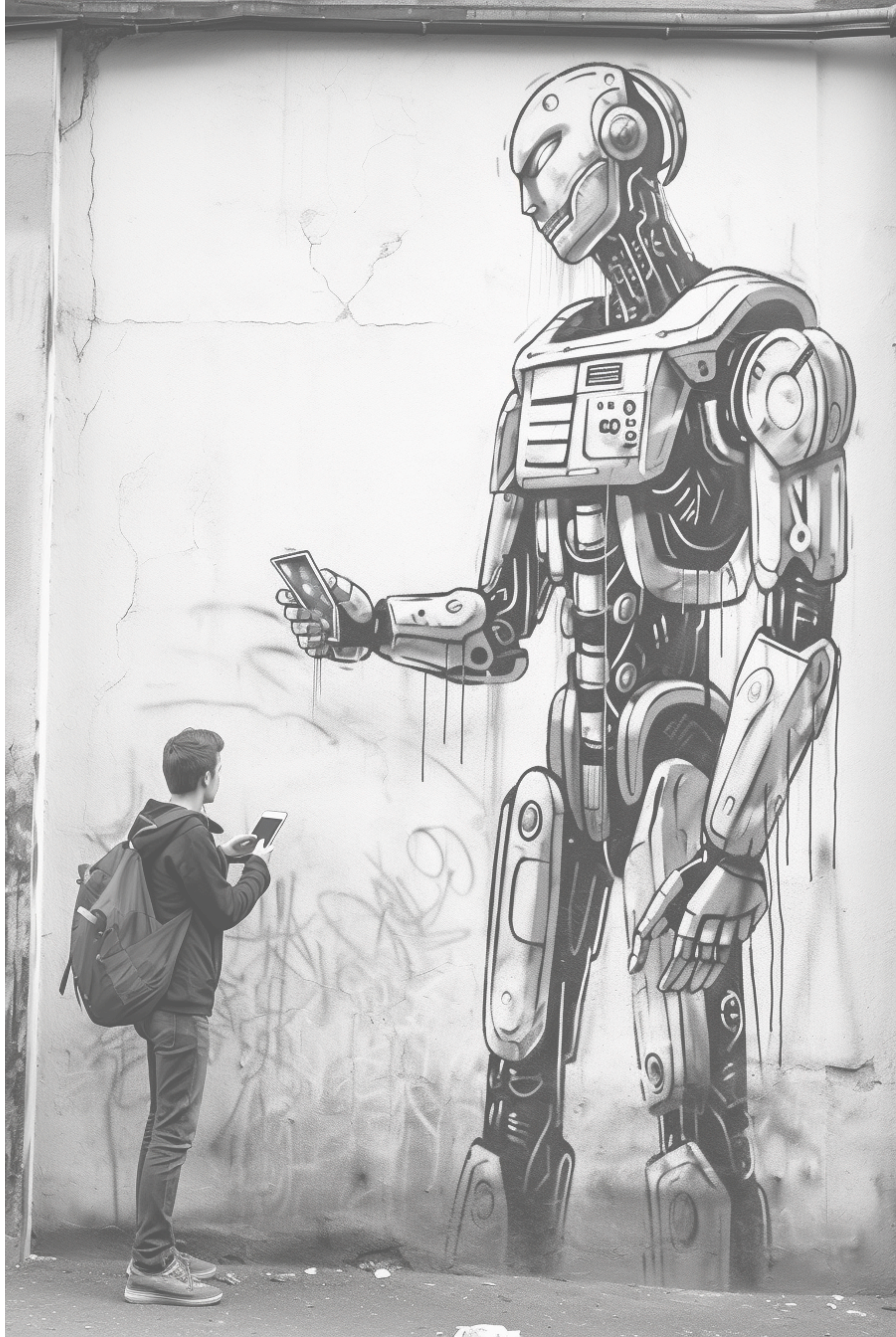
"a knife without a blade whose handle is missing"

Chomsky (1957), Ramesh et al. (2021)

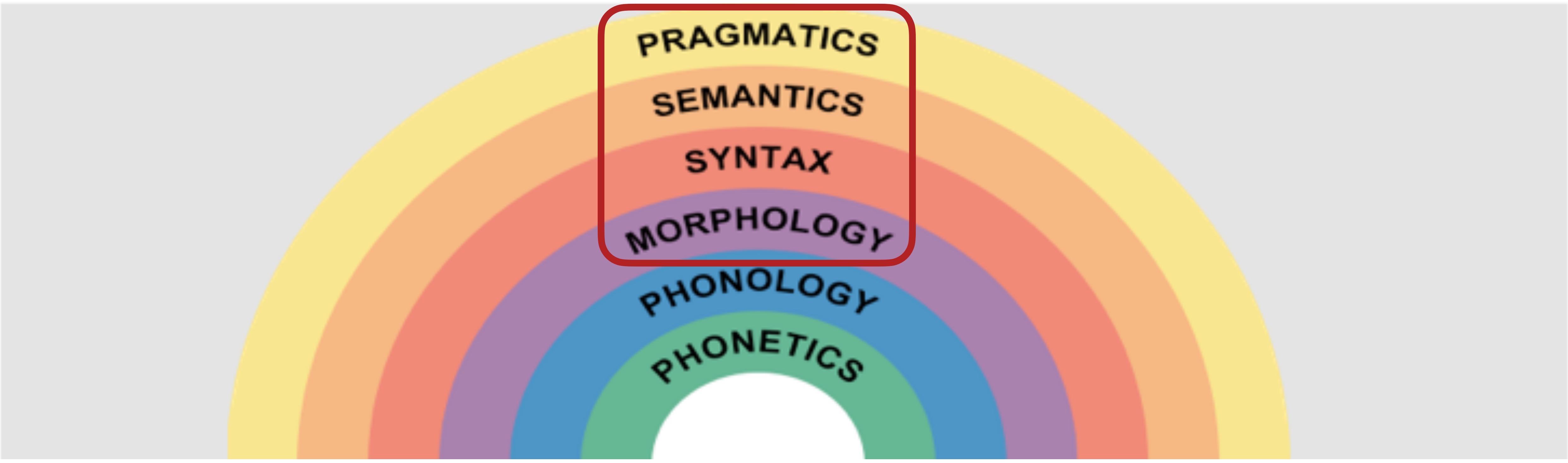
Mastering the impossible
Natural Language

- language use as a hallmark of intelligent behavior
  - Turing test
- uniquely human capability
  - what is the structure of the system that human learn and that makes it so flexible?
  - how are humans able to learn language?
- LLMs are systems that exhibit seemingly human-level language capabilities
  - what does this mean for the study of language and human cognition?

Turing (1950)



Insights from Linguistics
Structure of language (use)

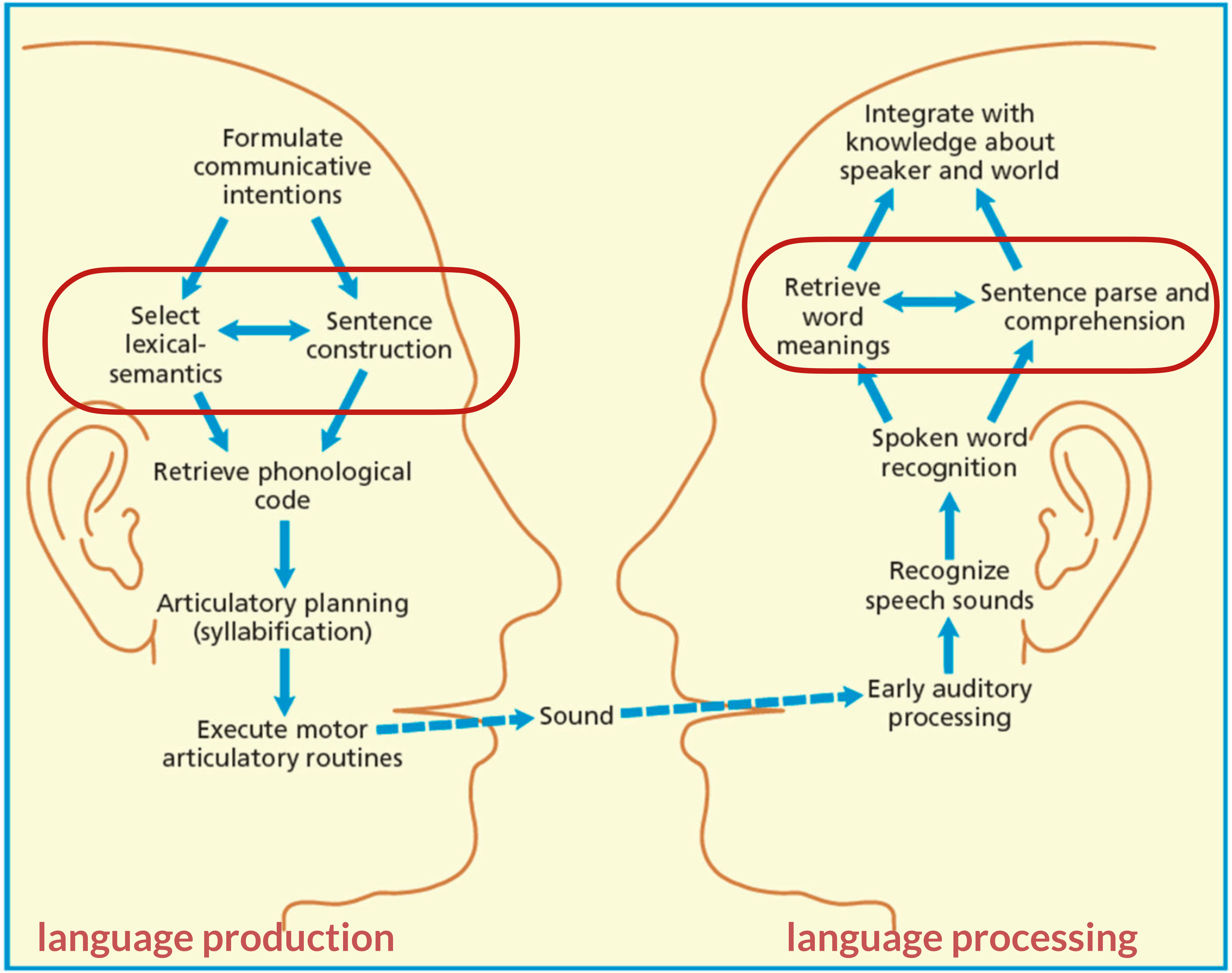


- generativism: language is construed by a grammar from which words and grammatical sentences can be constructed by applying explicit rules and minimal operations
  - humans have **competence** of the grammar, but their **performance** may deviate from the rules
- **compositional** semantics: meaning of larger units = meaning of parts + the syntactic structure
  - explanation of how we can **understand novel sentences**
- Grecian pragmatics: interlocutors infer conversational meaning in context assuming **cooperativity**
  - to **derive non-literal meaning**, interlocutors reason about each other assuming Gricean Maxims

Chomsky (1965), Grice (1975), Zimmerman & Sternefeld (2013)

Natural language in the wild

Psycholinguistics



our focus: can (neural) language models help understand / predict what happens at this stage of language processing?

Natural language in the wild

Psycholinguistics

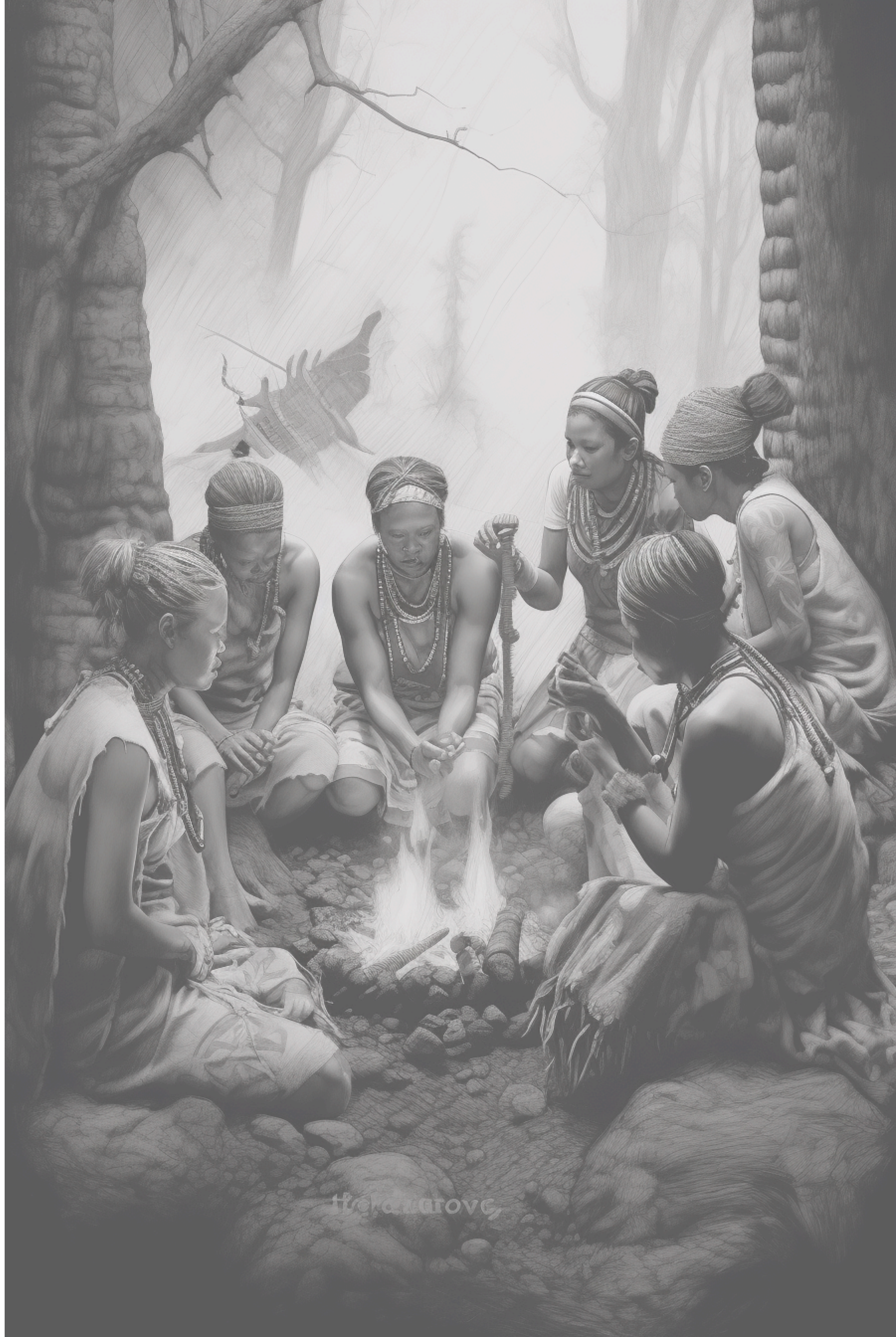
- **language acquisition**: how do humans learn (their native) language?
  - **poverty of the stimulus**: child-directed speed does not provide sufficient evidence for children to learn very feature of their native language; therefore, some structure must be innate (nativism / **UG**)
  - statistical & social learning:
    - infants learn certain properties based on **statistical properties** of child-directed speech
    - **child as a hacker**, language learning via **Bayesian inference**

Chomsky (1965), Perfors et al. (2011), Saffran et al. (1996), Xu & Tenenbaum (2000)

Are LLMs human-like with respect to language production, processing & learning?

In the next 30 minutes, your task is to:

1. think of arguments defending your assigned position (use the text as inspiration)
    a. download reading assigned to your question from Moodle
2. prepare a single slide with your arguments
3. present the slide to the class in max. 2 minutes



Are LLMs human-like with respect to language production? Yes, because..

Arguments collected by Hannes Leier, Lea Krumbach & Yuguang Lin

Evidential observations

- Creativity
- Syntax: use of complex grammatical or syntactic structures
  - (multiple embedded clauses, prepositional phrases, conjunctions etc; It uses pronouns, determiners, quantifiers, adjectives, etc. in realistic ways)
- Semantics: few apparent problems with agreement ir pronoun reference
- Pragmatics: coherent stories (makes some “sense”) -> solving frame problem?
- Long dependencies (across sentences)
- ChatGPT: Empathy with the user?

Technical explanations

- Architecture imitates human neural structures (neurone, attention, but also differences)
- Architecture allows longer and longer dependencies (attention)
- Models improve constantly

Are LLMs human-like with respect to language production? No, because..

Arguments collected by Moritz Goecuekbasi, Tabea Stier & Nion Schürmeyer

- Reasoning is often not human-like (frame problem).
- Lack of real-time information.
- Less likelihood of back references.
- No idea-guided speech production.
- Preference of probability over grammaticality.
- Problems with errors and biases.
- Unable to have personal opinions.

Are LLMs human-like with respect to language processing? No, because..

Arguments collected by students

- Human processing is (also) somewhat unclear, can only observe outcomes / circumstances
- Competence vs. performance
  - models have higher competence/performance ceiling and typically their performance=competence (no fatigue, distraction etc)
- Sequential (incremental) processing (humans always, not all models)
- Humans switch fluently between different input channels (models not yet)
- World knowledge, interaction with non-language knowledge (currently mostly limited to humans)