Signature of (Non-)Human-Like Sentence Processing in Large Language Models

Tatsuki Kuribayashi

(+ Recent linguistic typology alignment works if we have time)

Tatsuki Kuribayashi（栗林樹生）

MBZUAI

- Assistant Prof. at MBZUAI, United Arab Emirates (UAE) (2025/08-)
  - Visiting Researcher at University of Tokyo and Tohoku University, Japan (currently)
  - Postdoc at MBZUAI, UAE (2023-2025; Advisor: Timothy Baldwin)
  - Tohoku University, Japan (-2023; PhD supervisor: Kentaro Inui; Close collaborator: Yohei Oseki)
- Organizer of CMCL workshop (Cognitive Modeling and Computational Linguistics)
  - CMCL 2026 will be co-located with LREC in Palma, Spain (2026/5)!

Tatsuki Kuribayashi（栗林樹生）

- Assistant Prof. at MBZUAI, United Arab Emirates (UAE) (2025/08-)
  - Visiting Researcher at University of Tokyo and Tohoku University, Japan (currently)
  - Postdoc at MBZUAI, UAE (2023-2025; Advisor: Timothy Baldwin)
  - Tohoku University, Japan (-2023; PhD supervisor: Kentaro Inui; Close collaborator: Yohei Oseki)
- Organizer of CMCL workshop (Cognitive Modeling and Computational Linguistics)
  - CMCL 2026 will be co-located with LREC in Palma, Spain (2026/5)!

- The road less traveled:
  - Japan → UAE
    - From one of the most rural areas in Japan
  - Engineering → Language Science

Japan deploys soldiers to contain surge in bear attacks in Akita

Close encounters reported almost daily as bears intrude into residential areas and attack and sometimes kill people

Soldiers will set traps, transport local hunters and help dispose of dead bears. Officials said soldiers would not use firearms to cull the bears. Photograph: AP

MBZUAI…?

https://www.thenationalnews.com/news/uae/2025/09/26/sheikh-khaled-attends-ceremony-giving-honorary-doctorate-to-openais-sam-altman/

- Mohamed bin Zayed University of Artificial Intelligence
  - A newly-built university for AI fields (ML, CV, NLP, Robotics, HCI, CompBio…)
  - Alex visited last month for Embodied AI Symposium
    - I visited ETH last year

#	Institution	Count	Faculty
1	▶ Harbin Institute of Technology	115.6	43
2	▶ Peking University	113.6	51
3	▶ Carnegie Mellon University	111.4	36
4	▶ Tsinghua University	104.0	46
5	▶ University of Edinburgh	102.8	25
6	▶ Chinese Academy of Sciences	93.0	28
7	▶ University of Washington	87.1	19
8	▶ Stanford University	84.0	18
9	▶ Fudan University	77.3	21
10	▶ University of Maryland - College Park	75.8	22
11	▶ MBZUAI	74.7	31
12	▶ Johns Hopkins University	60.8	19
13	▶ Shanghai Jiao Tong University	58.9	42
14	▶ Nanyang Technological University	58.8	24
15	▶ New York University	57.3	13
16	▶ Univ. of Illinois at Urbana-Champaign	55.9	29

(CSRanking in NLP)

MBZUAI…?

- Mohamed bin Zayed University of Artificial Intelligence
  - A newly-built university for AI fields (ML, CV, NLP, Robotics, HCI, CompBio…)
  - Alex visited last month for Embodied AI Symposium
    - I visited ETH last year

- Launched PALM (**P**rocessing and **A**cquisition of **L**anguage in **M**achines and **M**ind) Group 🌴
  - Build a (geographically) new hub for computational linguistics
    - Cognitive modeling and interpretability
      - Eye tracker is going to be installed next year…!
    - Typologically-diverse research
      - Arabic, Japanese, sometimes impossible artificial language… to avoid making comp. psycholing. **WEIRD**
  - Fully-funded MSc/PhD course. Admission deadline: 12/15

My research topics

- Cognitive modeling
  - Are larger language models cognitively plausible?
- Interpretability
  - What information do LMs truly pay attention to?
- Linguistic typology and language acquisition
  - What kind of language design is easy for LMs to learn?
    - Collaborated with Alex as well!
- Past: Automated writing assistance

Lower Perplexity is Not Always Human-Like

Tatsuki Kuribayashi[1,2], Yohei Oseki[3,4], Takumi Ito[1,2],
Ryo Yoshida[3], Masayuki Asahara[5], Kentaro Inui[1,4]
[1]Tohoku University [2]Langsmith Inc. [3]University of Tokyo [4]RIKEN [5]NINJAL
{kuribayashi, takumi.ito.c4, inui}@tohoku.ac.jp ,
{oseki, yoshiryo0617}@g.ecc.u-tokyo.ac.jp ,masayu-a@ninjal.ac.jp

Context Limitations Make Neural Language Models More Human-Like

Tatsuki Kuribayashi[1,2] Yohei Oseki[3,4] Ana Brassard[1,4] Kentaro Inui[1,4]
[1]Tohoku University [2]Langsmith Inc. [3]University of Tokyo [4]RIKEN
{kuribayashi, inui}@tohoku.ac.jp
oseki@g.ecc.u-tokyo.ac.jp ana.brassard@riken.jp

Attention is Not Only a Weight: Analyzing Transformers with Vector Norms

Goro Kobayashi[1] Tatsuki Kuribayashi[1,2] Sho Yokoi[1,3] Kentaro Inui[1,3]
[1] Tohoku University [2] Langsmith Inc. [3] RIKEN
{goro.koba, kuribayashi, yokoi, inui}@ecei.tohoku.ac.jp

Which Word Orders Facilitate Length Generalization in LMs? An Investigation with GCG-Based Artificial Languages

Nadine El-Naggar* Tatsuki Kuribayashi* Ted Briscoe
Mohamed bin Zayed University of Artificial Intelligence
{nadine.naggar, tatsuki.kuribayashi, ted.briscoe}@mbzuai.ac.ae

How to understand human language processing?

- Opening human brains and directly observing their language is not technically or ethically possible
- Reading behavior is an observable interaction between human and text
  - Alternative approach will be analyzing brain signals (although they are sometimes noisy)

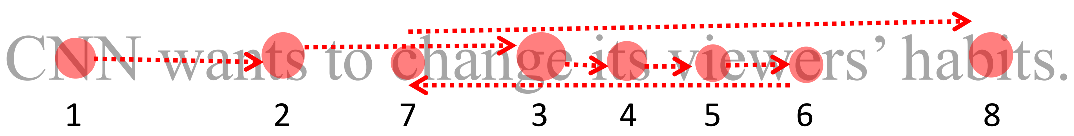
Eye movement

Assumption:
Humans see the parts involved in their processing

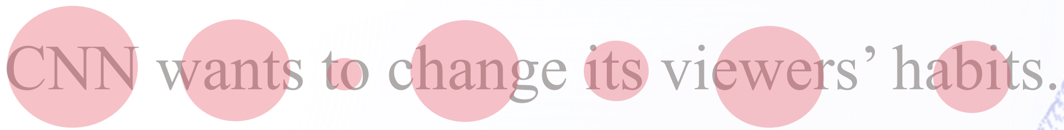
If you were to journey to the North of England, …

LM-based cognitive modeling

- Cognitive modeling×LMs: human reading behavior is analyzed with various information-theoretic measures computed by LMs
  [Hale, 2016][Goodkind&Bicknell,2018][Oh&Shculer, 2022][Kuribayashi+,2024]…



Human reading behavior
(larger circle=took more time)



E.g., predictability of each word in context computed by an LM
(larger circle=highly unpredictable)

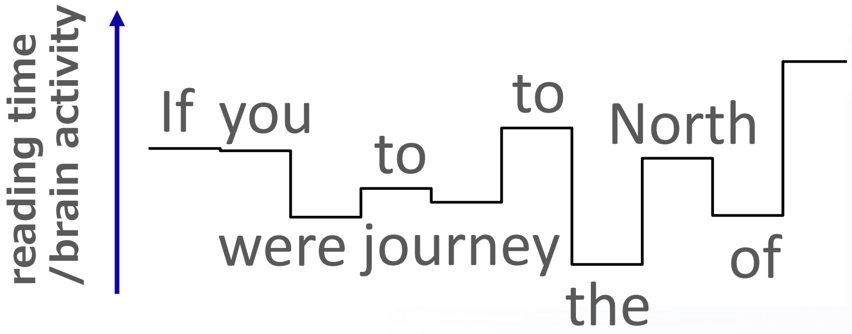
- Proof-of-concept for expectation-based human sentence processing
  [Hale, 2001][Levy, 2008][Smith&Levy, 2013]
  - $\text{ReadingTime}(\text{word}|\text{context}) \sim -\log_2 p(\text{word}|\text{context})$

Figures from https://acl2025-eyetracking-and-nlp.github.io/slides/Eye%20Tracking%20and%20NLP%20ACL%202025%20Tutorial%20no%20builds.pdf

LM-based cognitive modeling

If you were to journey to the North of England, ⋯

humans



Tokens: $\boldsymbol{w} = \{w_1 \ldots w_n\}$

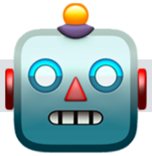
Cognitive load: $\boldsymbol{y} = \{y_1 \ldots y_n\}$

Emergent correlation

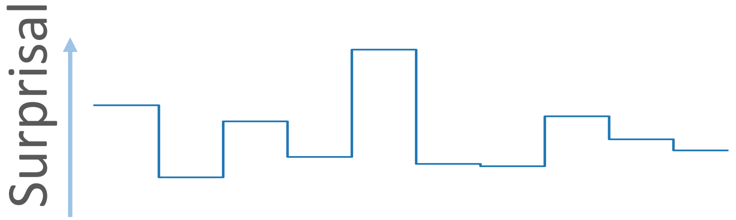
*training a regression model to rule out baseline factors and determine the coefficients, though

Not tuning any part

If you were to journey to the North of England, ⋯



LM

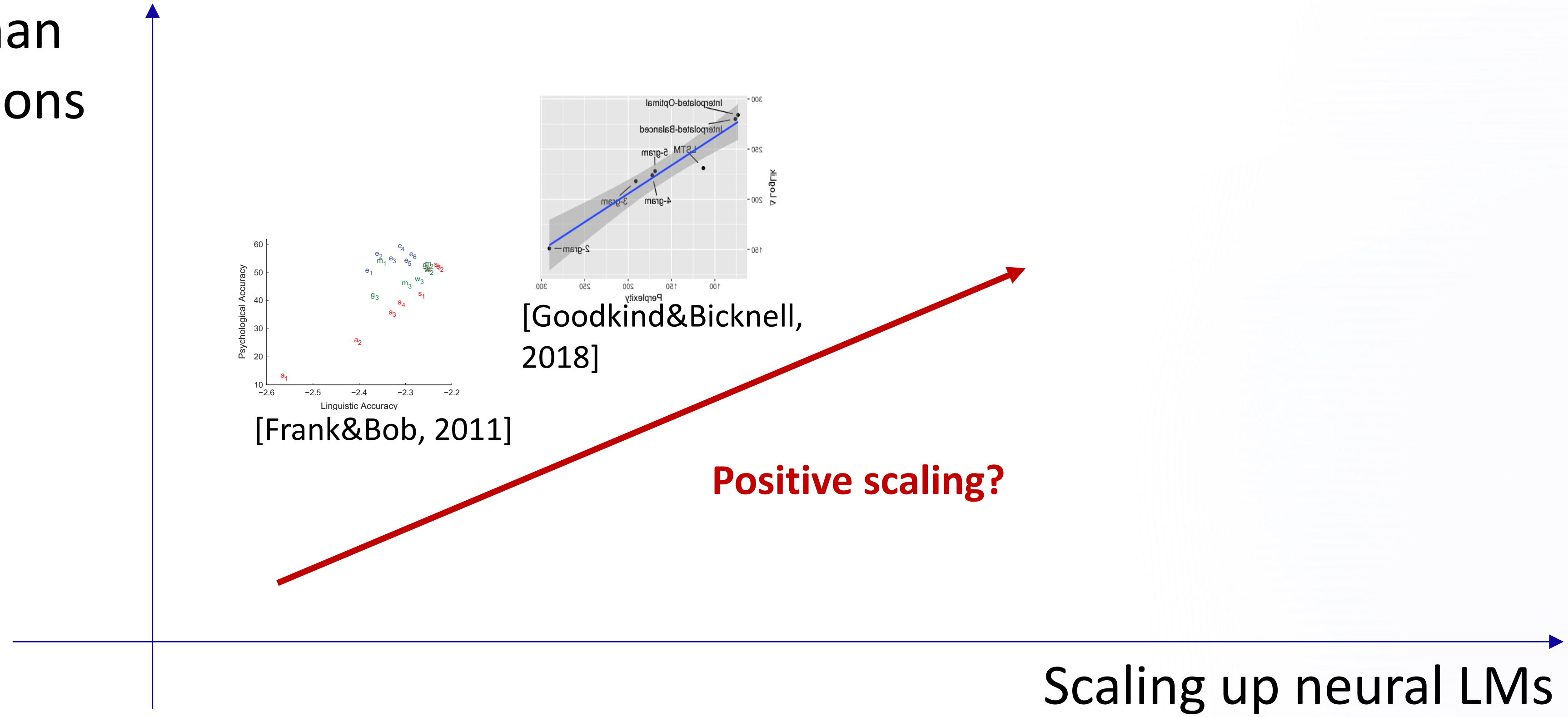


Surprisal: $\hat{\boldsymbol{y}} = \{-\log_2 p(w_1|\boldsymbol{w}_{<1}) \ \ldots - \log_2 p(w_n|\boldsymbol{w}_{<n})\}$

- The more unpredictable a word is, the more cognitive loads humans have
- Next question: what kind of LMs compute more human-like expectation?

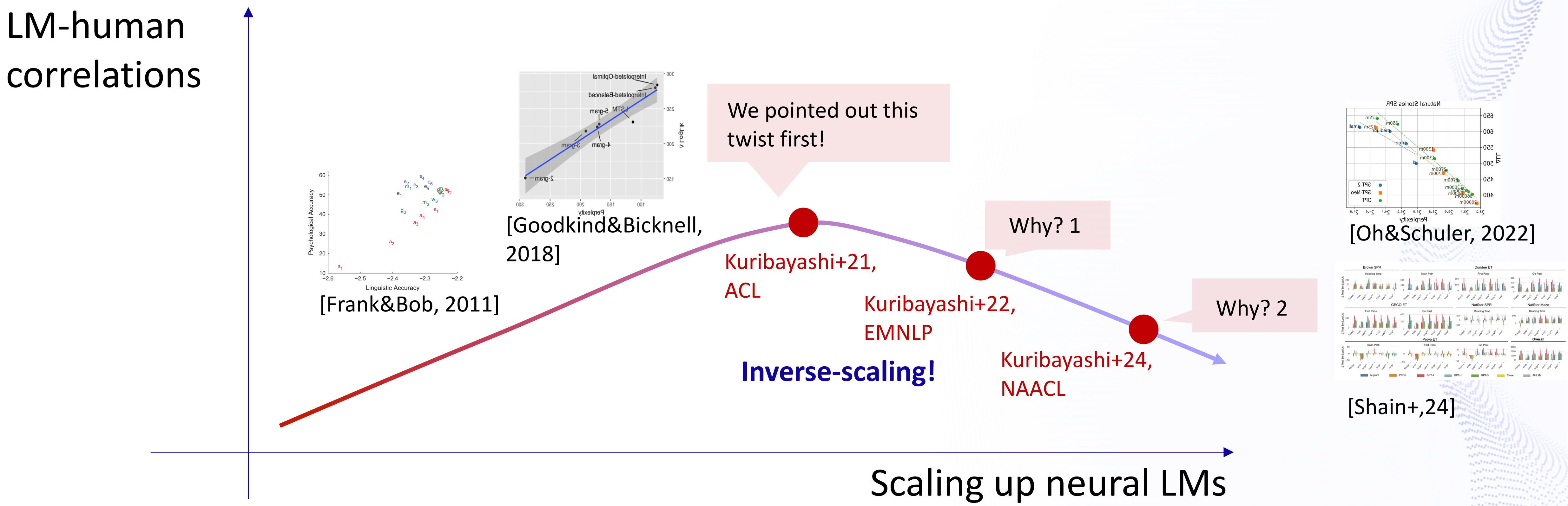
Are LMs approaching to human sentence processing model?
--- scaling law in cognitive modeling

LM-human correlations

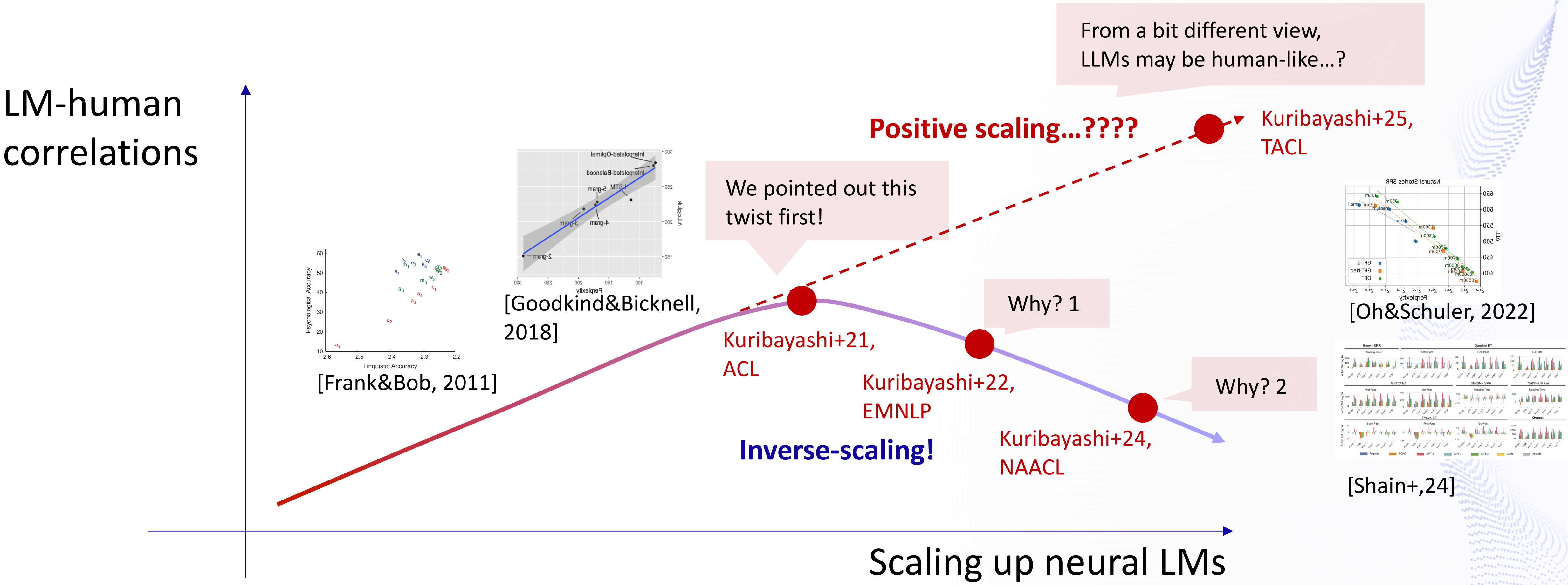


Scaling up neural LMs

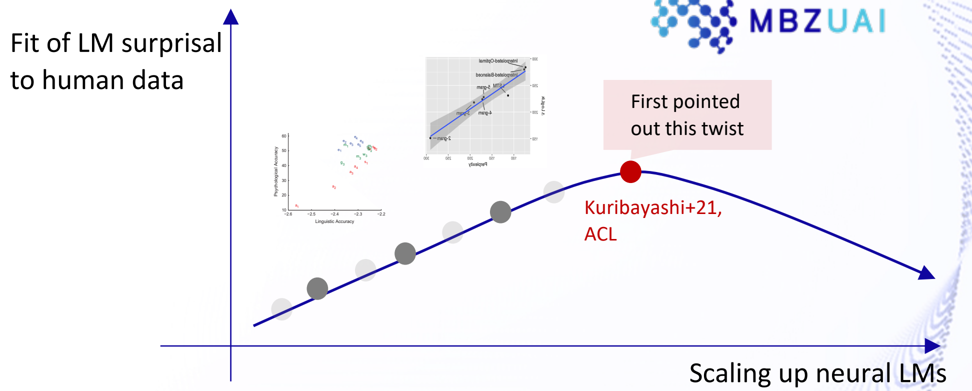
Are LMs approaching to human sentence processing model?
--- scaling law in cognitive modeling



Are LMs approaching to human sentence processing model?
--- scaling law in cognitive modeling



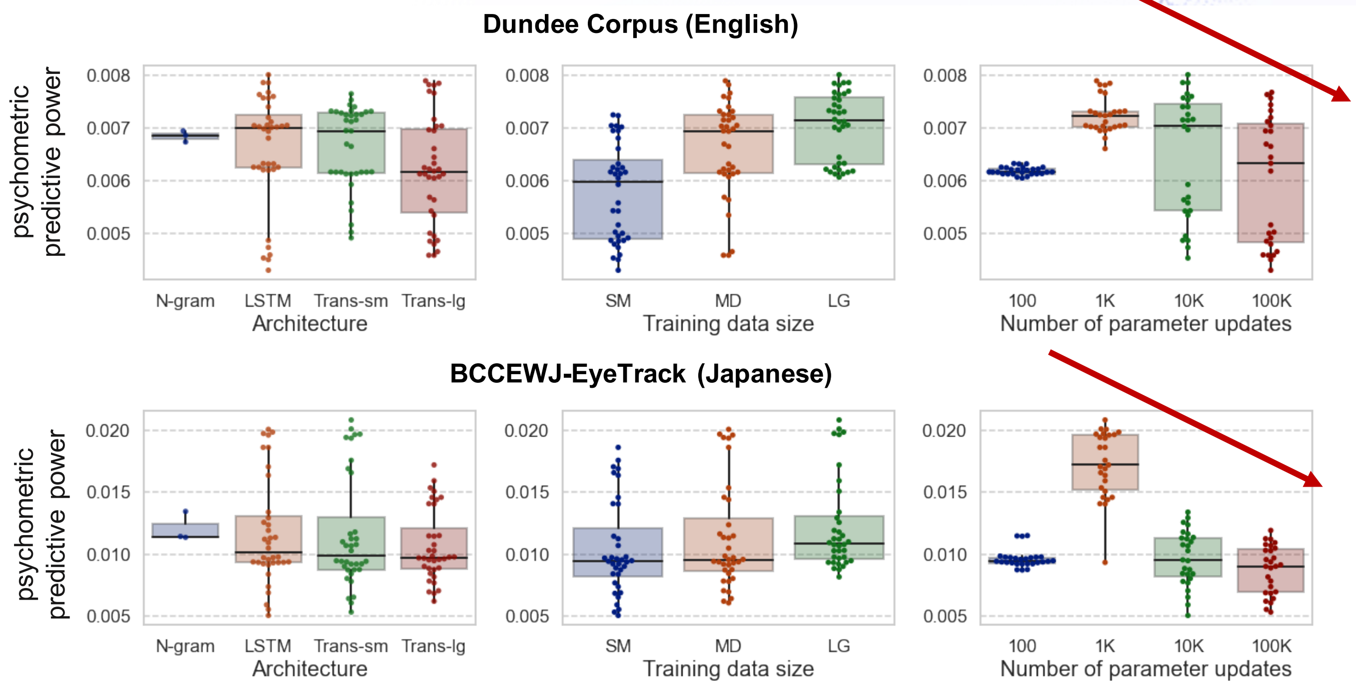
Kuribayashi+21 (ACL)



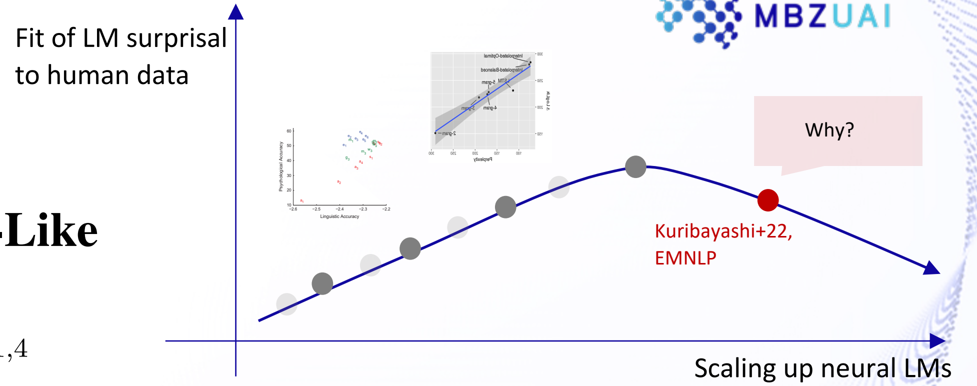
Lower Perplexity is Not Always Human-Like

Tatsuki Kuribayashi[1,2]**, Yohei Oseki**[3,4]**, Takumi Ito**[1,2]**,**
Ryo Yoshida[3]**, Masayuki Asahara**[5]**, Kentaro Inui**[1,4]
[1]Tohoku University [2]Langsmith Inc. [3]University of Tokyo [4]RIKEN [5]NINJAL
{kuribayashi, takumi.ito.c4, inui}@tohoku.ac.jp ,
{oseki, yoshiryo0617}@g.ecc.u-tokyo.ac.jp ,masayu-a@ninjal.ac.jp

- First systematic, cross-linguistic evaluation of psychometric predictive power (PPP) of surprisal from neural LMs



Kuribayashi+22 (EMNLP)



Context Limitations Make Neural Language Models More Human-Like

Tatsuki Kuribayashi[1,2] **Yohei Oseki**[3,4] **Ana Brassard**[1,4] **Kentaro Inui**[1,4]
[1]Tohoku University [2]Langsmith Inc. [3]University of Tokyo [4]RIKEN
{kuribayashi, inui}@tohoku.ac.jp
oseki@g.ecc.u-tokyo.ac.jp ana.brassard@riken.jp

- Why did LMs' surprisal deviate from human reading?
- LMs (Transformers w/ self-attention) may be too good to consider wide contexts, compared to human working memory

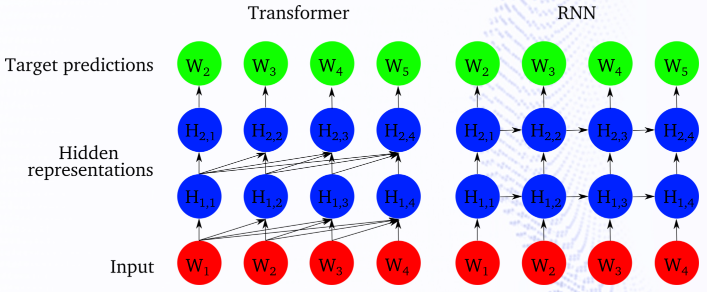


Figure 1: Comparison of sequential information flow through the Transformer and RNN, trained on next-word prediction.

[Merkx&Frank, 21]

SOV creates a long dependency
DLT [Gibson, 2000]

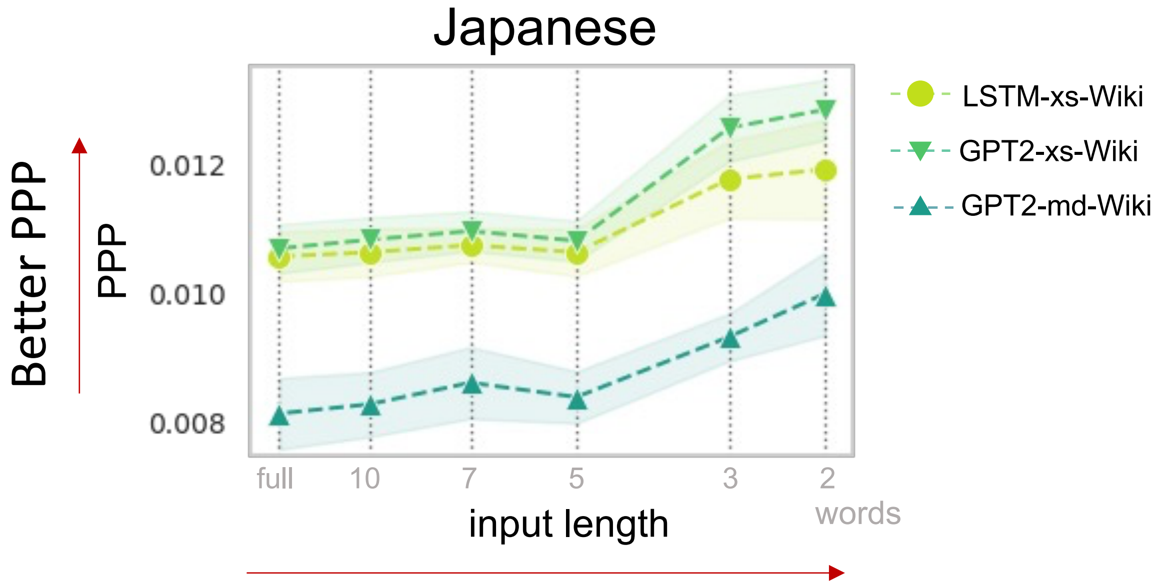
Ja: That man blue hat with clerk to very loud voice with spoke

En: That man spoke to a clerk with a blue hat with very loud voice

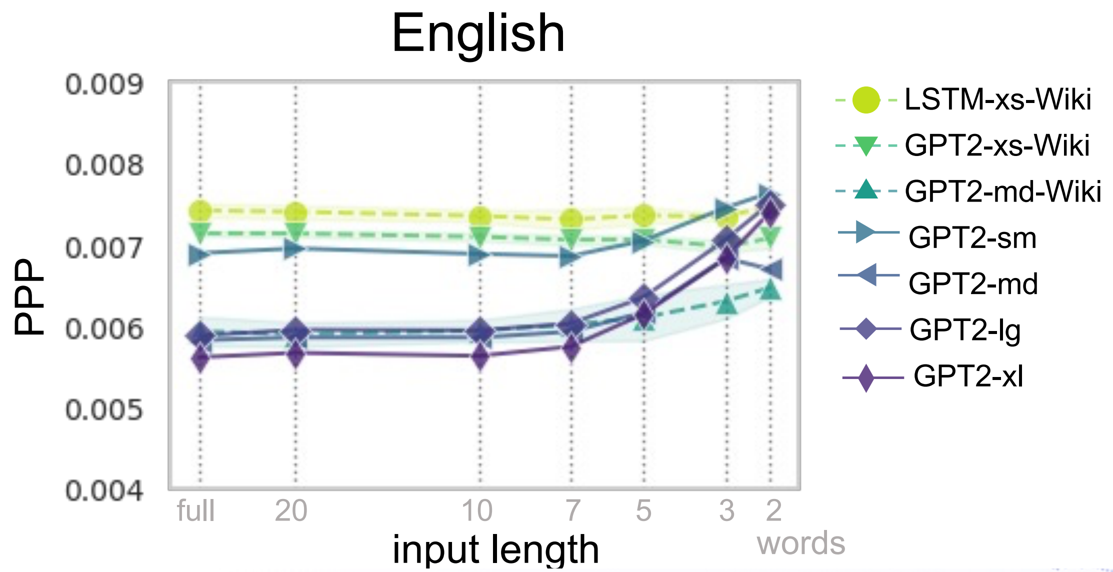
Kuribayashi+22 (EMNLP)

- Limiting LMs memory capacity aligns with human reading time
  - simple erasure of distant contexts works well surprisingly

$$\mathrm{ReadingTime}(w_t) \propto - -\log_2 p(w_t | \cancel{w_0}, \cancel{w_1}, \ldots w_{t-2}, w_{t-1})$$



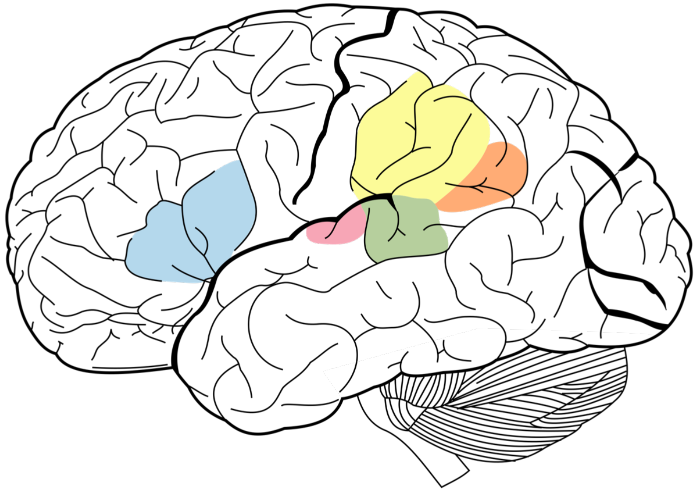
Sever memory limit.



Finding human-like sentence processing module in LLMs

- An LLM, as a whole model, is not like a model of human sentence processing

- But, is there any **part/circuit** that simulates human-like sentence processing?
    - We humans also may not use the entire brain for sentence processing (i.e., modularity of the brain)



Let’s go to Kuribayashi+, 25 (TACL)

Large Language Models Are Human-Like Internally

Tatsuki Kuribayashi[1,4] **Yohei Oseki**[2] **Souhaib Ben Taieb**[1,3]
Kentaro Inui[1,4,5] **Timothy Baldwin**[1,6]
[1]MBZUAI [2]The University of Tokyo [3]University of Mons
[4]Tohoku University [5]RIKEN [6]The University of Melbourne
{tatsuki.kuribayashi,souhaib.bentaieb,
kentaro.inui,timothy.baldwin}@mbzuai.ac.ae
oseki@g.ecc.u-tokyo.ac.jp

Background: Transformer is a stack of layers

- Temporal dynamics
  - Second layer must be computed after first layer, third layer must be computed after second layer…

Background: how LMs process input layer-by-layer

MBZUAI

BERT Rediscovers the Classical NLP Pipeline

Ian Tenney[1] Dipanjan Das[1] Ellie Pavlick[1,2]
[1]Google Research [2]Brown University
{iftenney,dipanjand,epavlick}@google.com

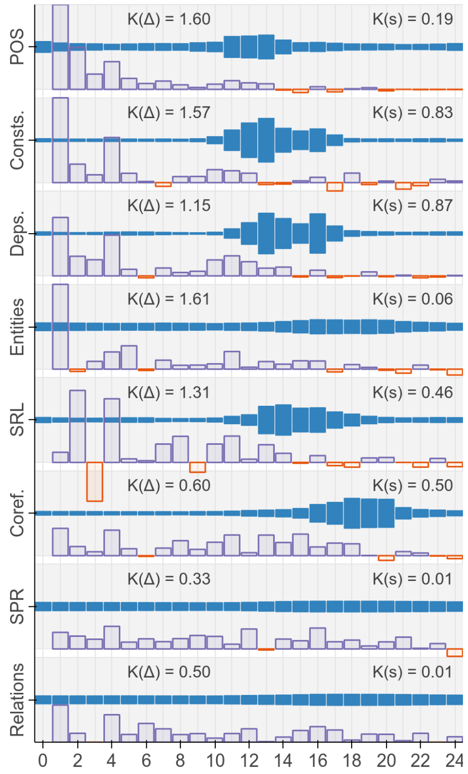
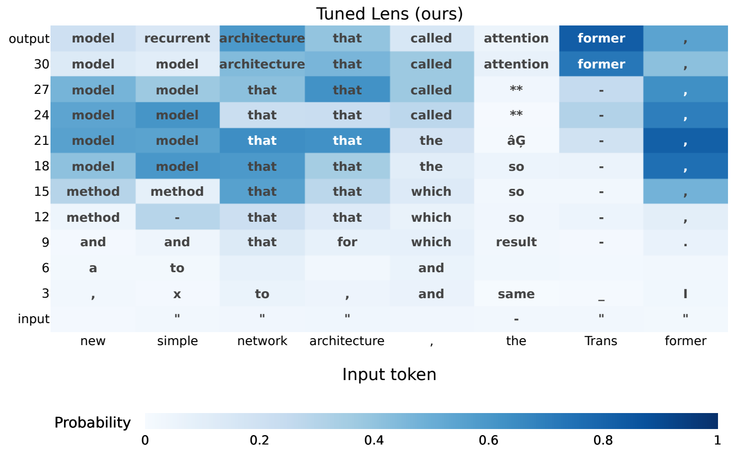


Figure 2: Layer-wise metrics on BERT-large. Solid (blue) are mixing weights $s_\tau^{(\ell)}$ (§3.1); outlined (purple) are differential scores $\Delta_\tau^{(\ell)}$ (§3.2), normalized for each task. Horizontal axis is encoder layer.

Eliciting Latent Predictions from Transformers with the Tuned Lens

Nora Belrose[1,2] Igor Ostrovsky[1] Lev McKinney[3,2] Zach Furman[1,4] Logan Smith[1] Danny Halawi[1]
Stella Biderman[1] Jacob Steinhardt[5]



Signatures of human-like processing in Transformer forward passes

Jennifer Hu
Kempner Institute
Harvard University

Michael A. Lepori
Department of Computer Science
Brown University

Michael Franke
Department of Linguistics
University of Tübingen

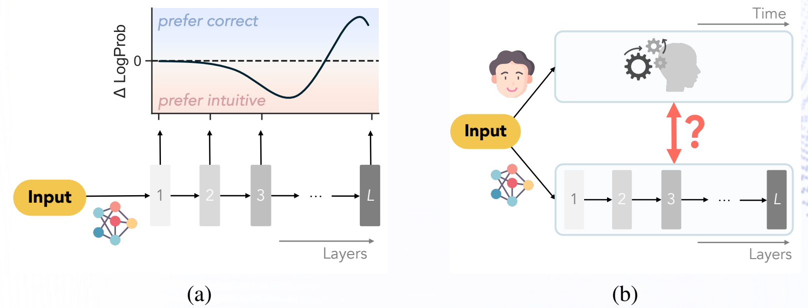
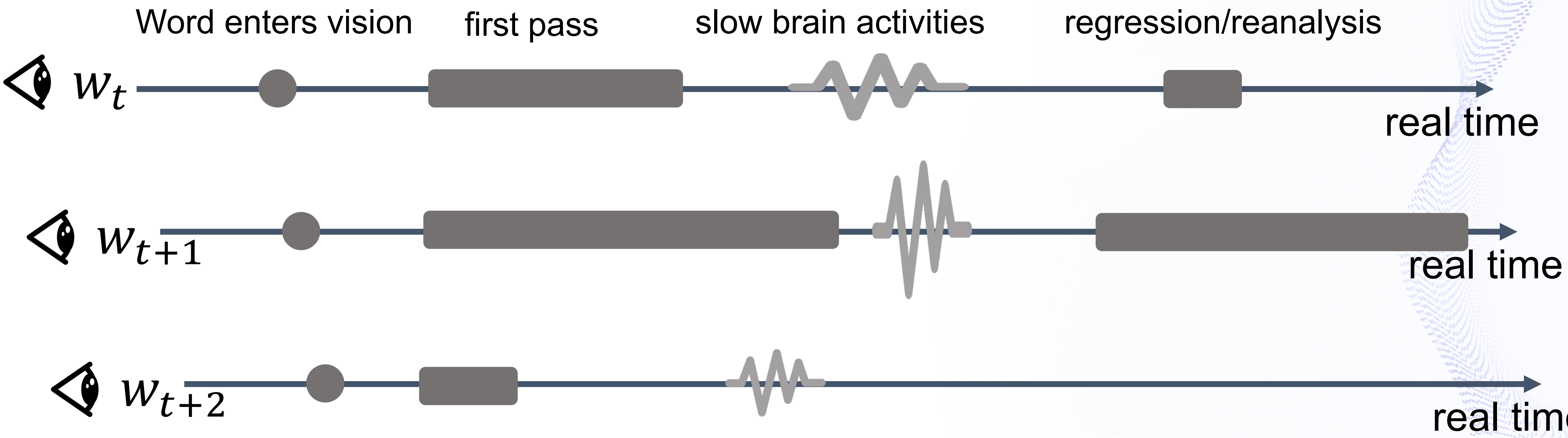


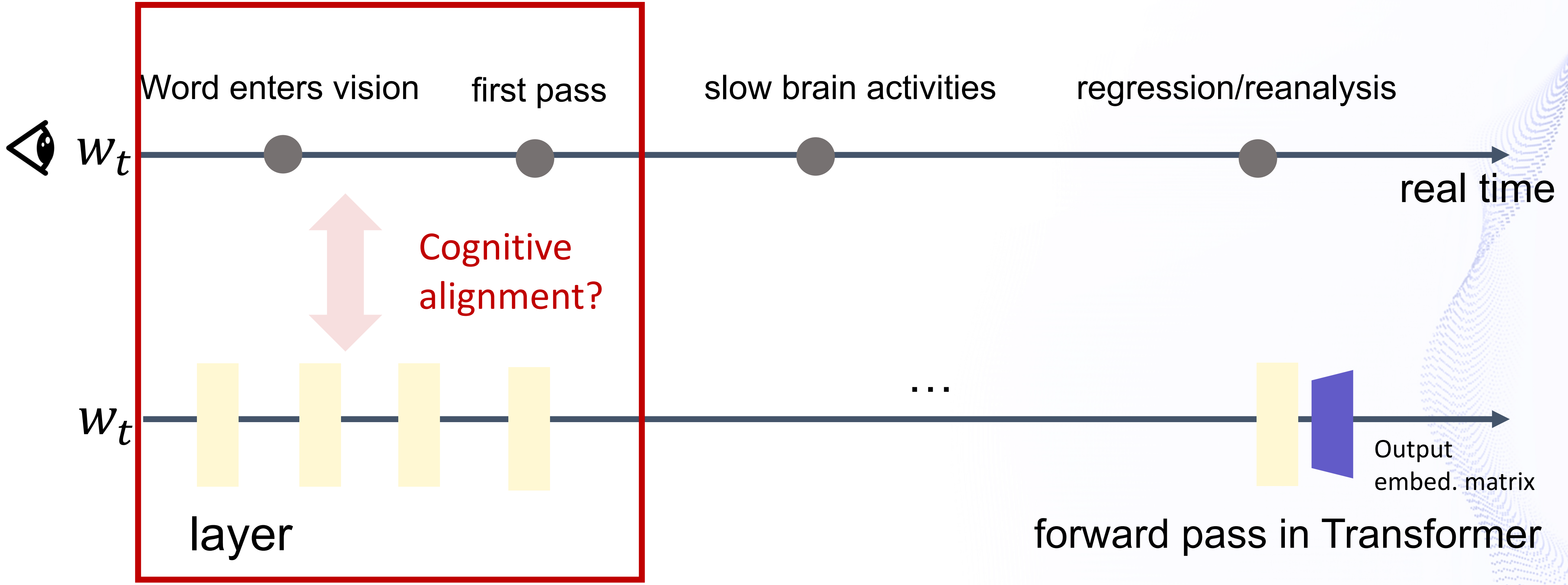
Figure 1: **Overview of our study.** (a) Experiment 1: We explore whether forward passes show mechanistic signatures of competitor interference, first preferring a salient competing intuitive answer before preferring the correct answer. (b) Experiment 2: We systematically investigate the ability of dynamic measures derived from forward passes to predict indicators of processing load in humans.

What happens on a human side

- Humans show behavioral/physiological signals in a different time-scale when processing a word in sentence



General question: Are humans' and LMs' real-time processing aligned?



One layer ≈ One Gamma Cycle…?

- One transformer layer ≈ one gamma cycle’s worth of cortical processing…?

(I only just received this advertisement yesterday 😅)

One Layer ≈ One Gamma Cycle

a Cognitive Clock for Brains and LLMs

SKYLAR DETURE
DEC 01, 2025

1 Share

A Numerical Coincidence

Okay, so here’s a pretty cool numerical coincidence that might be more than a coincidence.

If you estimate how many *gamma cycles a human brain spends per word*, and you look at the *number of layers that an LLM spends to generate a token*, the estimates are pretty close: on the order of tens of “processing steps” per word/token.

Calculated Theoretical γ cycles/word for Speech and Typing

Mode	WPM	Seconds/word	Low-γ (≈40 Hz) cycles/word	Mid-γ (≈70 Hz) cycles/word	High-γ (≈100 Hz) cycles/word	High-γ (≈150 Hz) cycles/word
Typing slow	40	1.50	60	105	150	225
Typing fast	80	0.75	30	52	75	112
Speech slow	100	0.60	24	42	60	90
Speech fast	200	0.30	12	21	30	45

Get the data · Embed · Created with Datawrapper

That suggests a nice heuristic:

> One transformer layer ≈ one gamma cycle’s worth of cortical processing.

It’s handwavey and hypothetical but it’s quite convenient. Actual time comparisons between human and LLM cognition are problematic because you can run a transformer at any speed you want (if you have enough compute). Using *one layer ≈ one gamma cycle* lets us talk about the “speed” of thought in both systems even when wall-

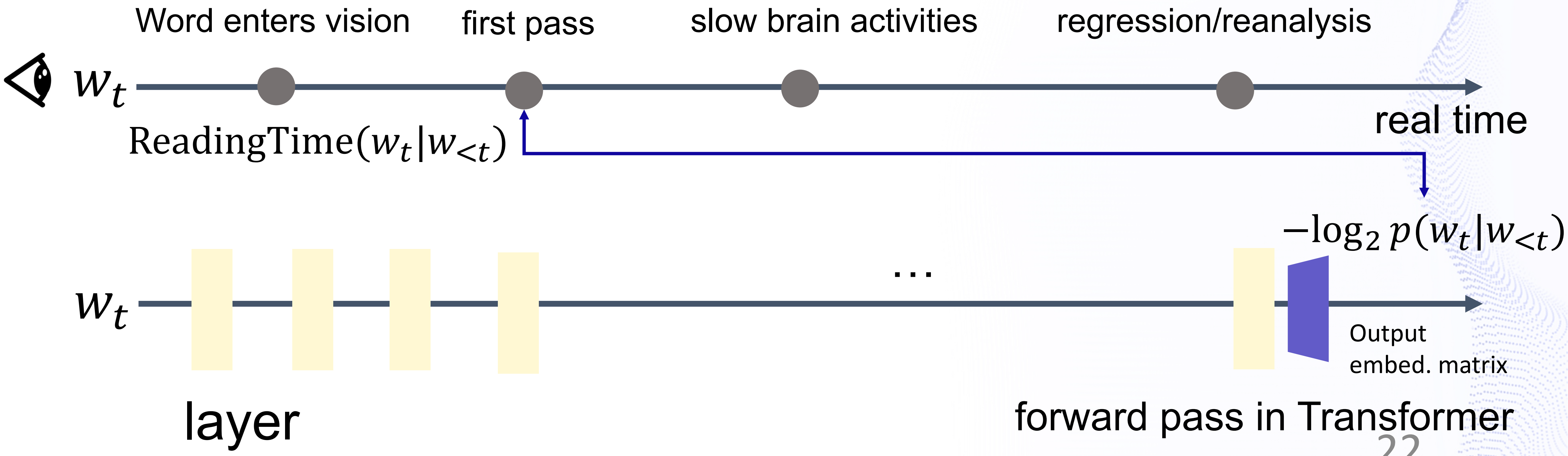
The "gamma cycle time window" in neuroscience refers to a brief temporal window, typically between ***10 to 33 milliseconds (ms)****, during which neurons synchronize their firing to integrate information.*

Neural Integration: *Neurons optimally integrate synaptic inputs from other neurons that arrive within this narrow time frame. Inputs arriving outside this window are effectively "ignored" for that specific processing event, which helps the brain organize information efficiently. (from Gemini)*

https://sdeture.substack.com/p/one-layer-one-gamma-cycle

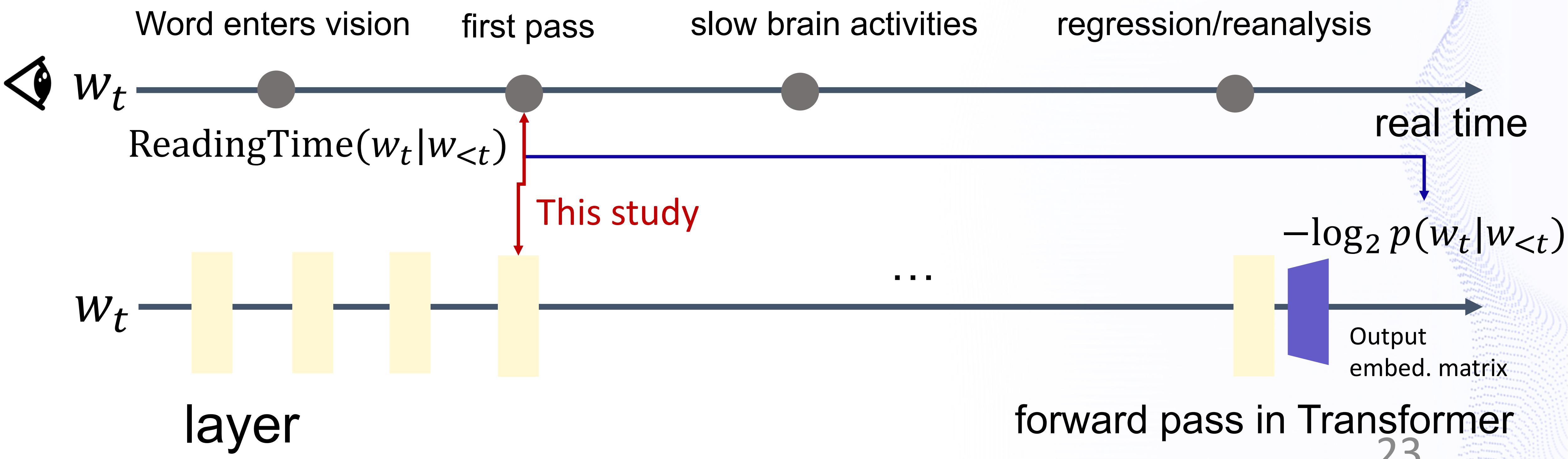
Existing surprisal-based reading time modeling

- Existing reading-time modeling studies only used the probability computed at the last layer
  - C.f. brain alignment studies compares alignment between LM internals and brain images [Schrimpf+, 21]



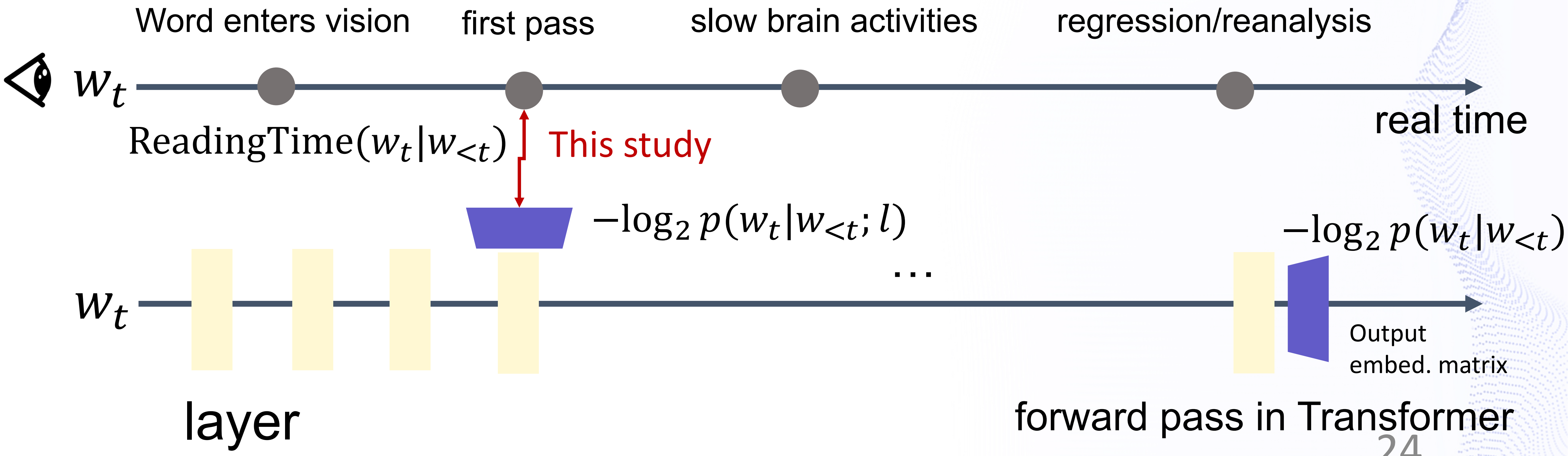
This study: internal alignment

- (Fast) human sentence processing behavior can align with earlier layer of LMs…?



How? logit-lens/tuned-lens

- We need next-word probability from internal layers
- Interpretability techniques are useful
  - Logit-lens [nostalgebraist, 20] extracts probability by directly applying output embedding matrix
  - Tuned-lens [Berlose+,23]



Psychometric predictive power

Word:	CNN	wants	to	...
Human RT	349	217	132	...
LM surprisal	12.4	2.2	0.3	...

fit

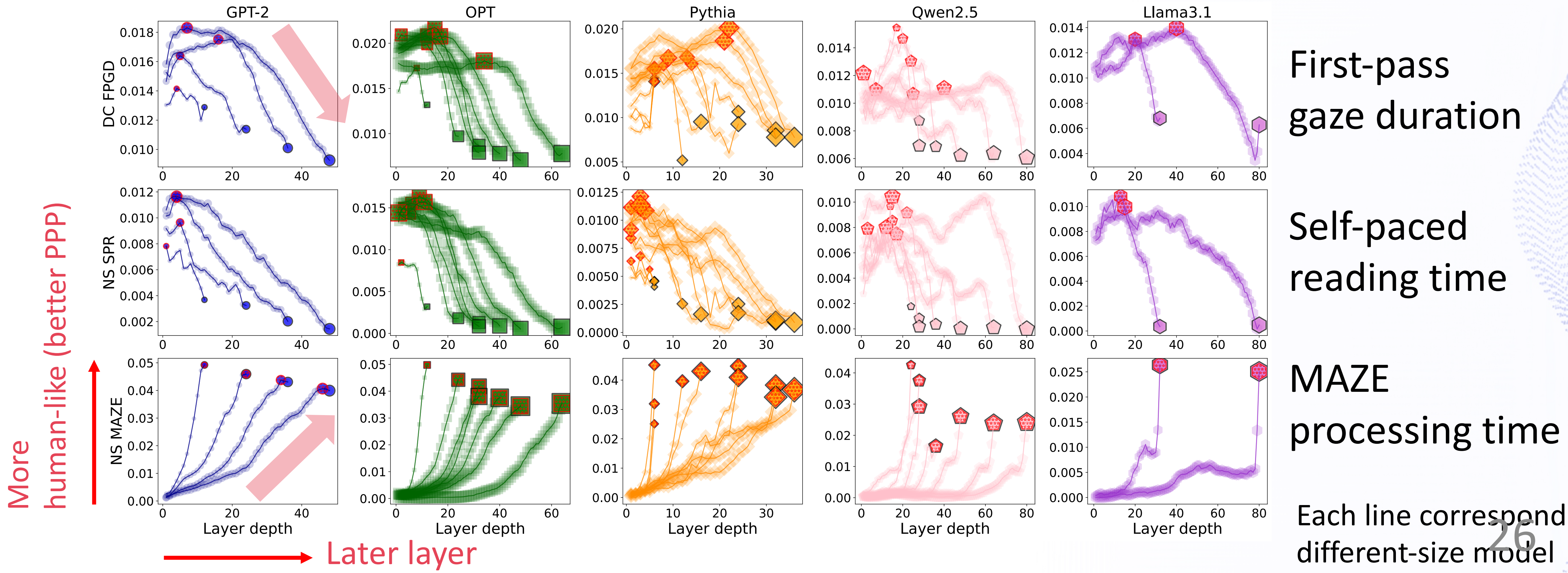
- Psychometric predictive power (PPP)
  - Loglikelihood difference (goodness-of-fit) between the target regression model and baseline regression model
    - Target regression model:
      ReadingTime(w_t) ~ length(w_t) + freq(w_t) + length(w_{t-1}) + freq(w_{t-1}) + length(w_{t-2}) + freq(w_{t-2}) + surprisal(w_t) + surprisal(w_{t-1}) + surprisal(w_{t-2})
    - Baseline regression model
      ReadingTime(w_t) ~ length(w_t) + freq(w_t) + length(w_{t-1}) + freq(w_{t-1}) + length(w_{t-2}) + freq(w_{t-2}) + surprisal(w_{t-1}) + surprisal(w_{t-2})

Handles spillover effects in advance

- We used 30 LMs and 15 datasets on human reading behavior/physiology (e.g., N400 signals)

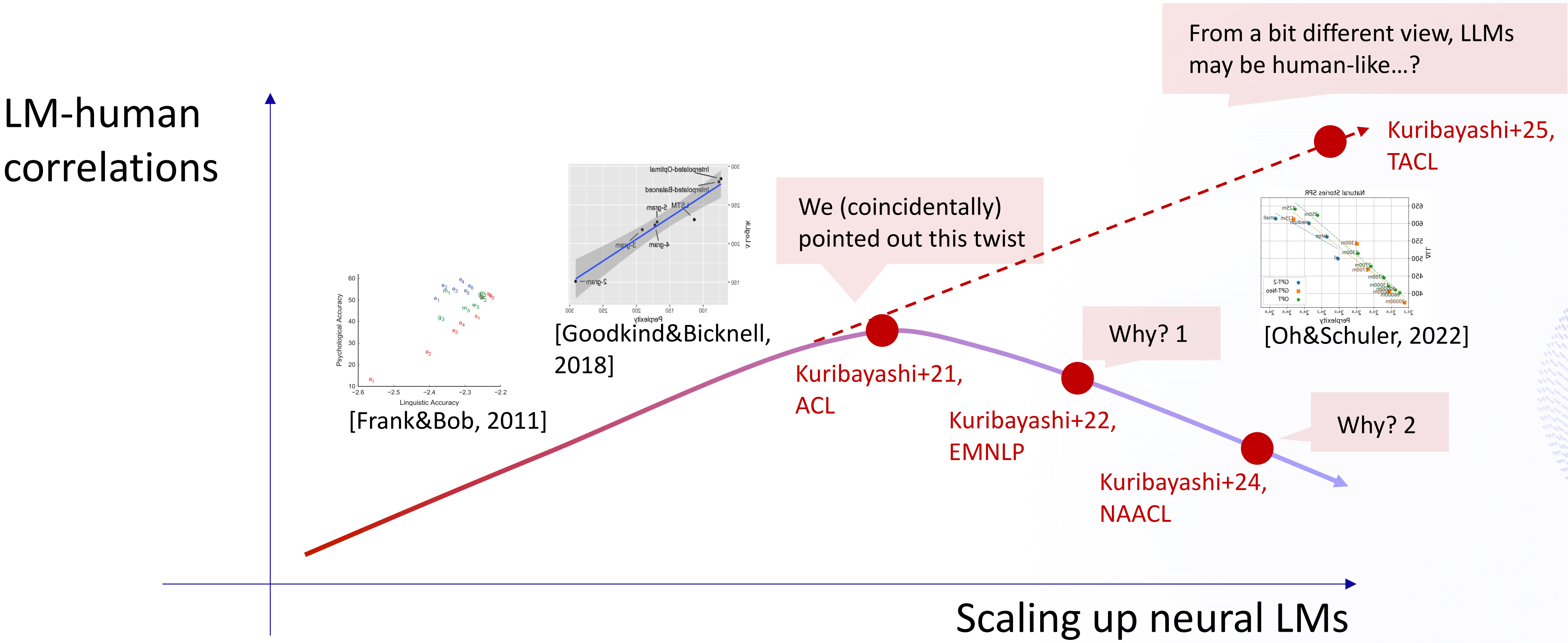
Results (summary)

- Different human measures align with different LM layers
- Fast human response (e.g., first gaze duration) tends to align better with early layers of LMs than slow response (EEG signals)



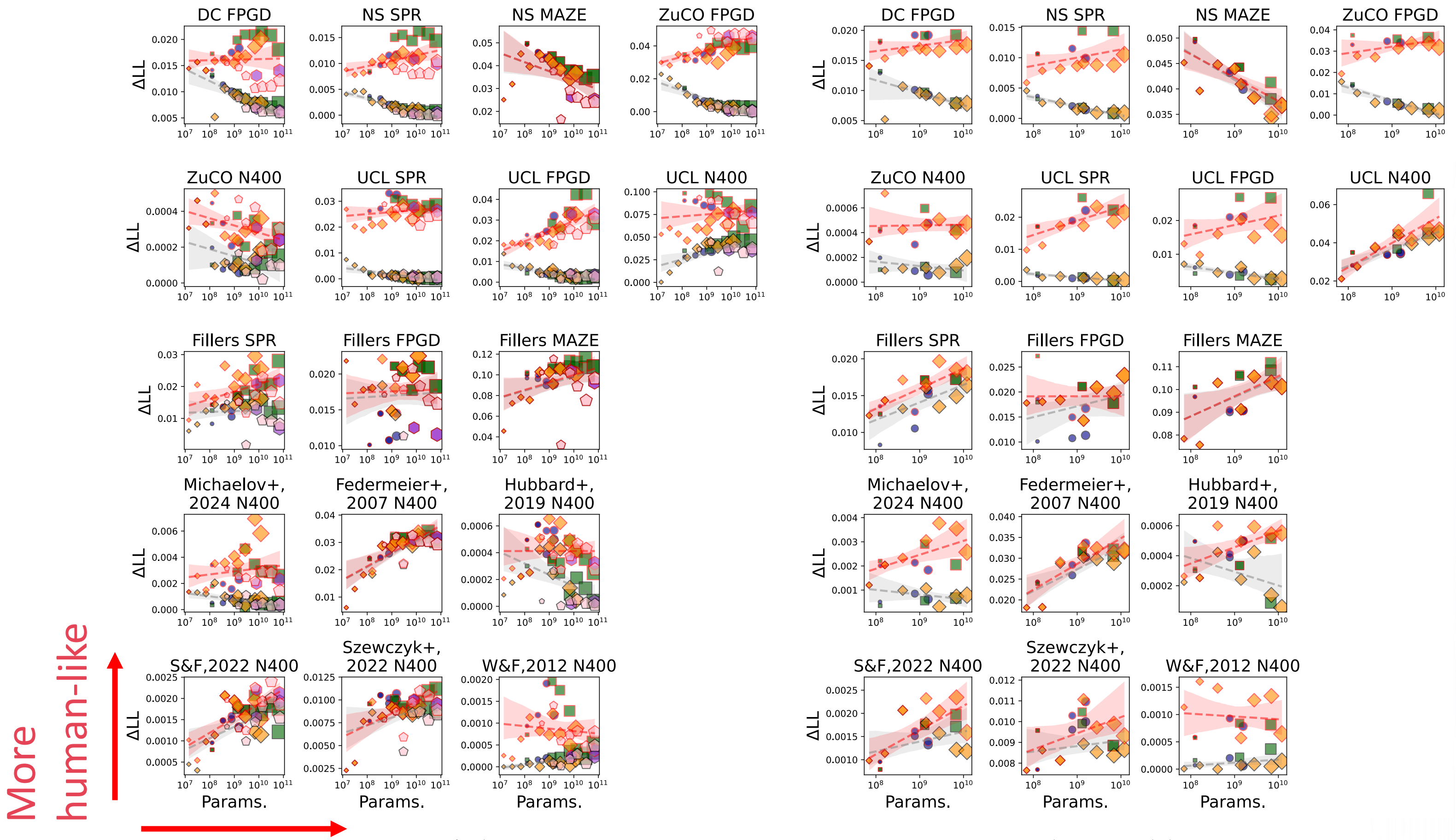
Each line corresponds to different-size model

Are LMs approaching to human sentence processing model? --- scaling law in cognitive modeling



Analysis: results

- Once the scope is extended to LM internals, larger LMs are not always worse (rather better) model of human sentence processing



Bigger model

(a) Logit-lens

(b) Tuned-lens

Relationship between model size and PPP from best layer

Relationship between model size and PPP from last layer

Analysis: connection to working memory limitation

- Transformer's superhuman context access is considered as one reason of human-LLM misalignment in cognitive modeling [Kuribayashi+,22][Oh+,24]

- One interpretation: Surprisal from earlier layer is less contextualized that matches human-like, moderately context-dependent processing

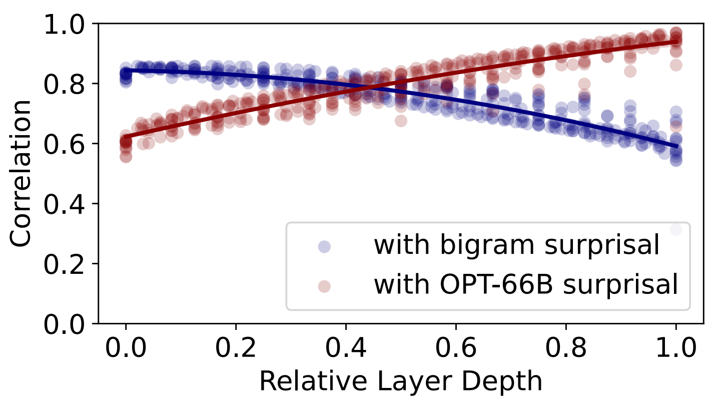
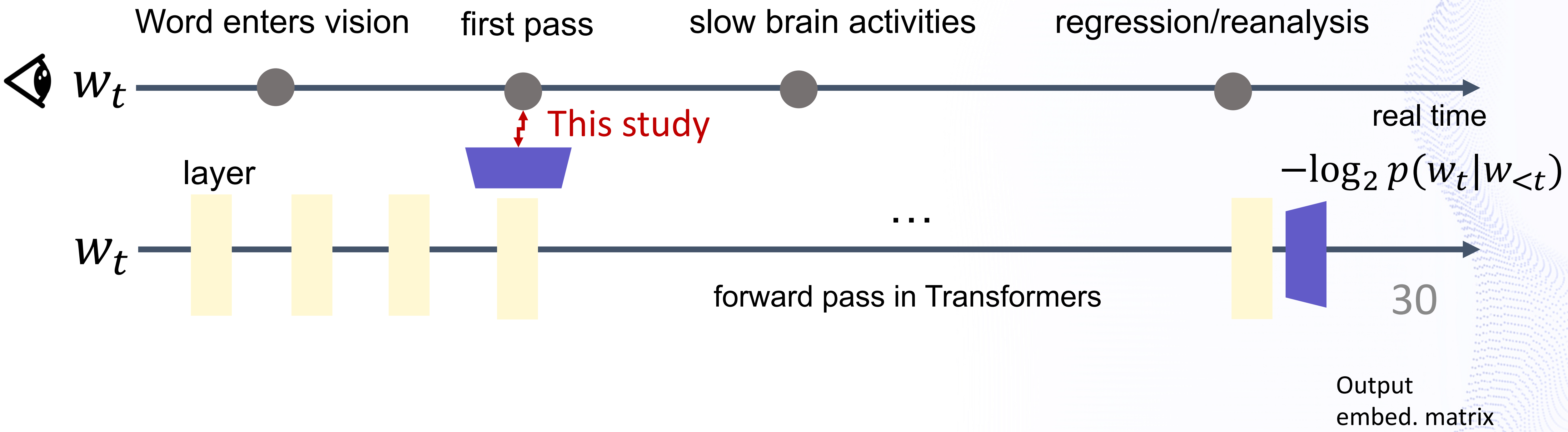


Figure 7: The markers correspond to all the internal layers of our targeted LMs, which are sorted by relative layer depth (x-axis). Two types of scores (y-axis) are plotted: (i) Pearson correlation coefficient between each layer's surprisal vs. less-contextualized bigram surprisal (blue); and (ii) each layer's surprisal vs. well-contextualized LLM surprisal (red). We used tuned-lens results.

Take home messages

- Information-theoretic values from internal layers can also be options to analyze human-LLM cognitive alignment
- Good LM-counterpart to (fast) human word/sentence processing would be an early layer of LLMs



Work in progress

Confidential

Confidential

Confidential

Confidential

Confidential

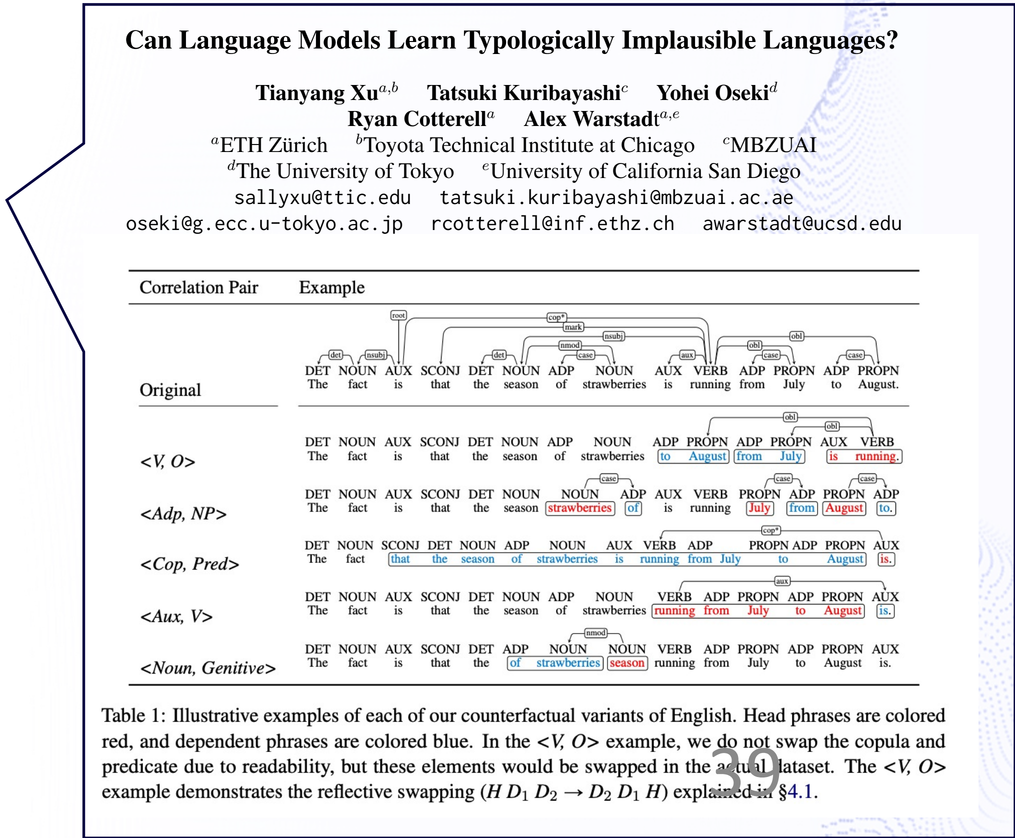
If we have time

What kind of language is easier for LMs to learn?

- How to answer this question?
  - What kind of metric makes a fair comparison (e.g., against different character set)?
  - **How can one isolate a specific linguistic factor (e.g., word order)?**

[Mielke+, 2019]

- We need artificially controlled corpus
  - Corpus-first approach: manipulate existing corpus
    - E.g., Create head-final English and head-initial English
  - **Grammar-first approach: generate fully-artificial but error-free controlled corpora with grammar rules**



Can Language Models Learn Typologically Implausible Languages?

Tianyang Xu[a,b] Tatsuki Kuribayashi[c] Yohei Oseki[d]
Ryan Cotterell[a] Alex Warstadt[a,e]
[a]ETH Zürich [b]Toyota Technical Institute at Chicago [c]MBZUAI
[d]The University of Tokyo [e]University of California San Diego
sallyxu@ttic.edu tatsuki.kuribayashi@mbzuai.ac.ae
oseki@g.ecc.u-tokyo.ac.jp rcotterell@inf.ethz.ch awarstadt@ucsd.edu

Correlation Pair	Example
Original	DET NOUN AUX SCONJ DET NOUN ADP NOUN AUX VERB ADP PROPN ADP PROPN The fact is that the season of strawberries is running from July to August.
<V, O>	DET NOUN AUX SCONJ DET NOUN ADP NOUN ADP PROPN ADP PROPN AUX VERB The fact is that the season of strawberries to August from July is running.
<Adp, NP>	DET NOUN AUX SCONJ DET NOUN NOUN ADP AUX VERB PROPN ADP PROPN ADP The fact is that the season strawberries of is running July from August to.
<Cop, Pred>	DET NOUN SCONJ DET NOUN ADP NOUN AUX VERB ADP PROPN ADP PROPN AUX The fact that the season of strawberries is running from July to August is.
<Aux, V>	DET NOUN AUX SCONJ DET NOUN ADP NOUN VERB ADP PROPN ADP PROPN AUX The fact is that the season of strawberries running from July to August is.
<Noun, Genitive>	DET NOUN AUX SCONJ DET ADP NOUN NOUN VERB ADP PROPN ADP PROPN AUX The fact is that the of strawberries season running from July to August is.

Table 1: Illustrative examples of each of our counterfactual variants of English. Head phrases are colored red, and dependent phrases are colored blue. In the <V, O> example, we do not swap the copula and predicate due to readability, but these elements would be swapped in the actual dataset. The <V, O> example demonstrates the reflective swapping ($H\ D_1\ D_2 \rightarrow D_2\ D_1\ H$) explained in §4.1.

Related Work: Artificial Languages

- White and Cotterell (2021) used **PCFGs** to generate 64 ALs and investigate which word order leads to lower perplexity

	Japanese		English		Spanish	
Switch	**Value**	**Example**	**Value**	**Example**	**Value**	**Example**
S	0	猫が食べる。	0	The cat eats.	0	El gato come.
VP	0	猫がネズミを食べる。	1	The cat eats the mouse.	1	El gato come el ratón.
Comp	0	猫が食べると思う。	1	I think that the cat eats.	1	Pienso que el gato come.
PP	0	テーブルの上の猫が食べる。	1	The cat on the table eats.	1	El gato sobre la mesa come.
NP	0	小さな猫が食べる。	0	The small cat eats.	1	El gato pequeño come.
Rel	0	ミルクを飲む猫が食べる。	1	The cat that drinks milk eats.	1	El gato que bebe leche come.

Table 2: Demonstration of the orders of the switch constituents in Japanese, English and Spanish

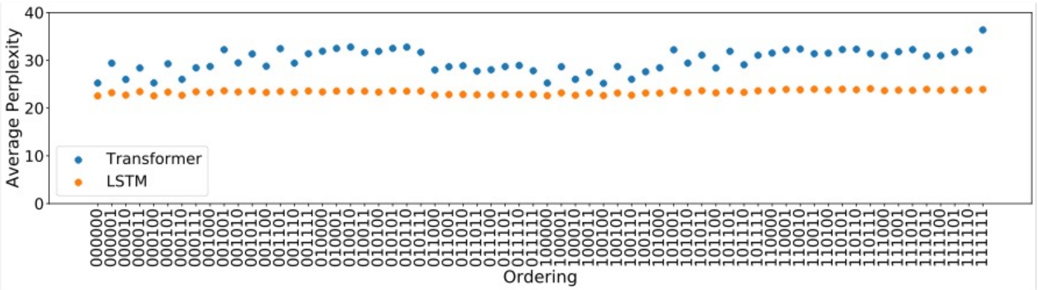
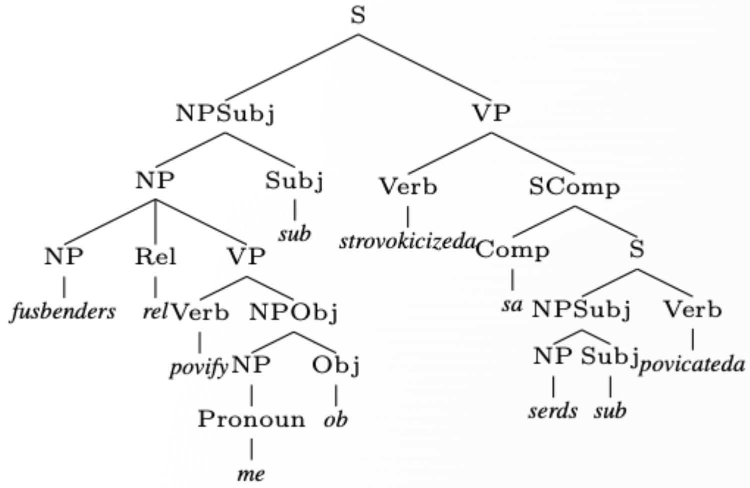


Figure 3: All scores achieved by LSTM- and transformer-based models



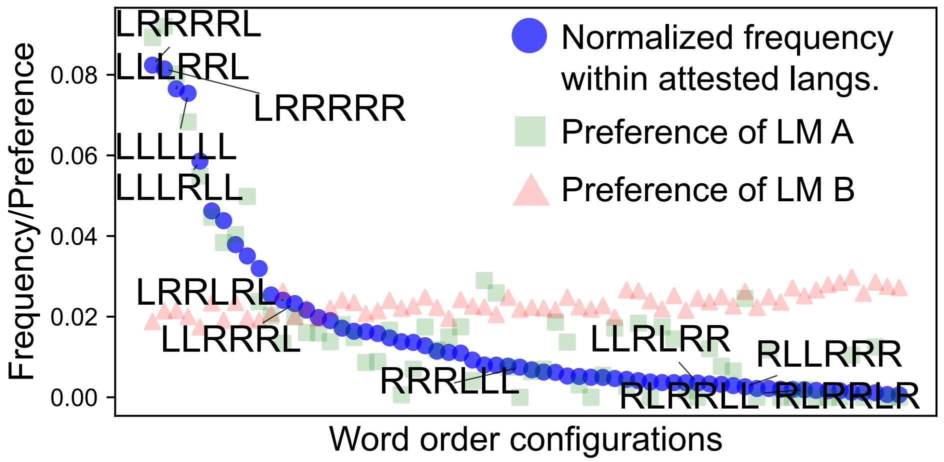
(b) Grammar 011101: fusbenders rel povify me ob sub strovokicizeda sa serds sub povicateda .

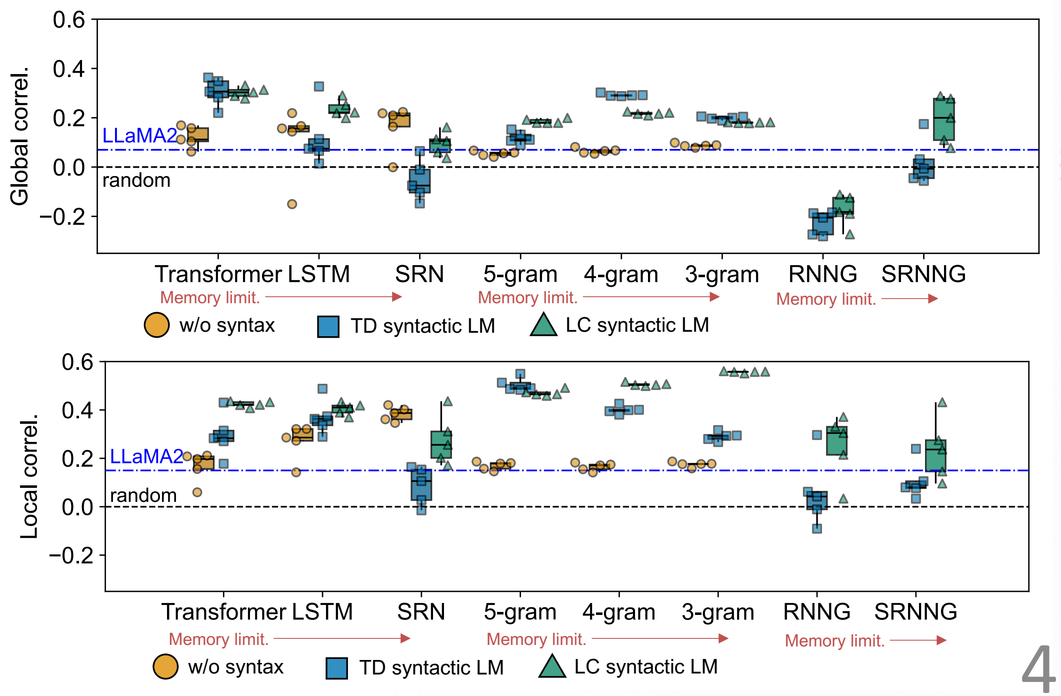
Figures from White and Cotterell 2021

- Their PCFGs **did not cover** complex constructions such as **unbounded dependencies, or VSO and OSV** orders, resulting in **limited** coverage and reality of ALs

Systematic comparison of typological alignment

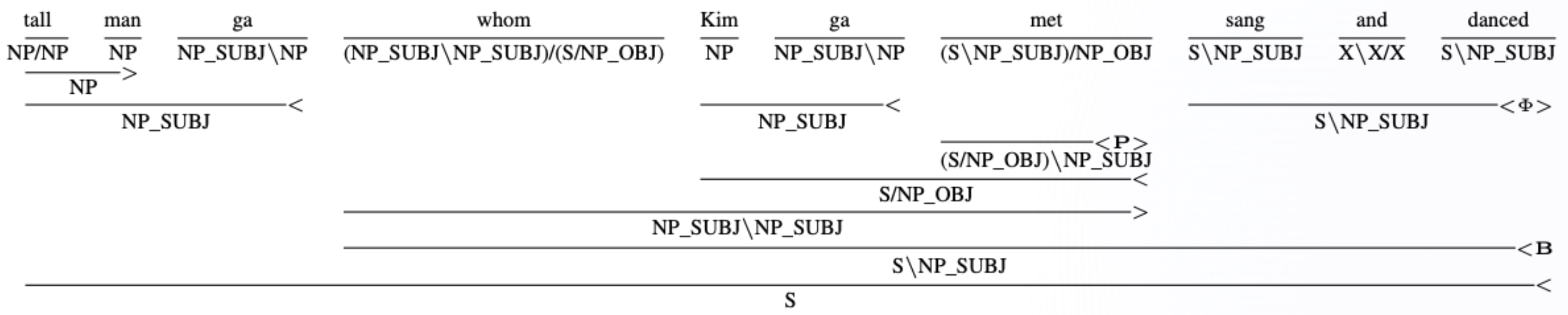
- Which LMs' inductive bias is most aligned with typological frequency of word order?
  - Typological frequency from WALS [Dryer&Haspelmath, 2013]
  - Cognitively-motivated one (memory limitation & left-corner parsing) is relatively well





PCFG to CCG (CoNLL 2025)

- Propose a general direction to use **Generalized Categorial Grammars (GCGs)** for AL creation
  - Naturally include mildly context-sensitive constructions
  - Create **96 ALs**, including VSO and OSV variations (8% of NLs) missed in existing works



- **Re-evaluate** the word order preference of LSTMs and Transformers on these **new ALs** and also extend analyses to **learning (preference) trajectory**

Exemplifying Different Word Orders

Param.	Description	0 (head-final)	1 (head-initial)
S	Order of subject and verb	$VI \to S\backslash NP_{SUBJ}$ $VT \to (S\backslash NP_{SUBJ})\|NP_{OBJ}$ $VCOMP \to (S\backslash NP_{SUBJ})\|SCOMP$	$VI \to S/NP_{SUBJ}$ $VT \to (S/NP_{SUBJ})\|NP_{OBJ}$ $VCOMP \to (S/NP_{SUBJ})\|SCOMP$
VP	Order of object and verb	$VT \to (S\|NP_{SUBJ})\backslash NP_{OBJ}$ $VCOMP \to (S\|NP_{SUBJ})\backslash SCOMP$ $REL \to (NP_{SUBJ}\|NP_{SUBJ})\|(S\backslash NP_{OBJ})$	$VT \to (S\|NP_{SUBJ})/NP_{OBJ}$ $VCOMP \to (S\|NP_{SUBJ})/SCOMP$ $REL \to (NP_{SUBJ}\|NP_{SUBJ})\|(S/NP_{OBJ})$
O	Order of subject and object	Restriction to make an S precede O as canonical word order	Restriction to make an O precede S as canonical word order
COMP	Position of complementizer	$COMP \to SCOMP\backslash S$	$COMP \to SCOMP/S$
PP	Postposition or preposition	$PREP \to (NP\backslash NP)/NP$	$PREP \to (NP/NP)\backslash NP$
ADJ	Order of adjective and noun	$ADJ \to NP/NP$	$ADJ \to NP\backslash NP$
REL	Position of relativizer	$REL \to (NP_{SUBJ}/NP_{SUBJ})\backslash(S\|NP_{OBJ})$	$REL \to (NP_{SUBJ}\backslash NP_{SUBJ})/(S\|NP_{OBJ})$

Table 2: Word order parameters and their associated GCG categories. "A→B" indicates A|B (A is expanded to B) in the GCG derivation.

Now word order switch is translated to the directionality of "/" or "\"

(Re)defining word order switches with GCG notation

Parameter value	Similar Language	Sentence
0000000	Japanese	Tall man ga and small child ga grandmother o visited
0101101	English	Tall man ga and small child ga visited grandmother o
0101111	Spanish	Man tall ga and child small ga visited grandmother o
0000010	Hmong Daw	Man tall ga and child small ga grandmother o visited

Dataset Generation

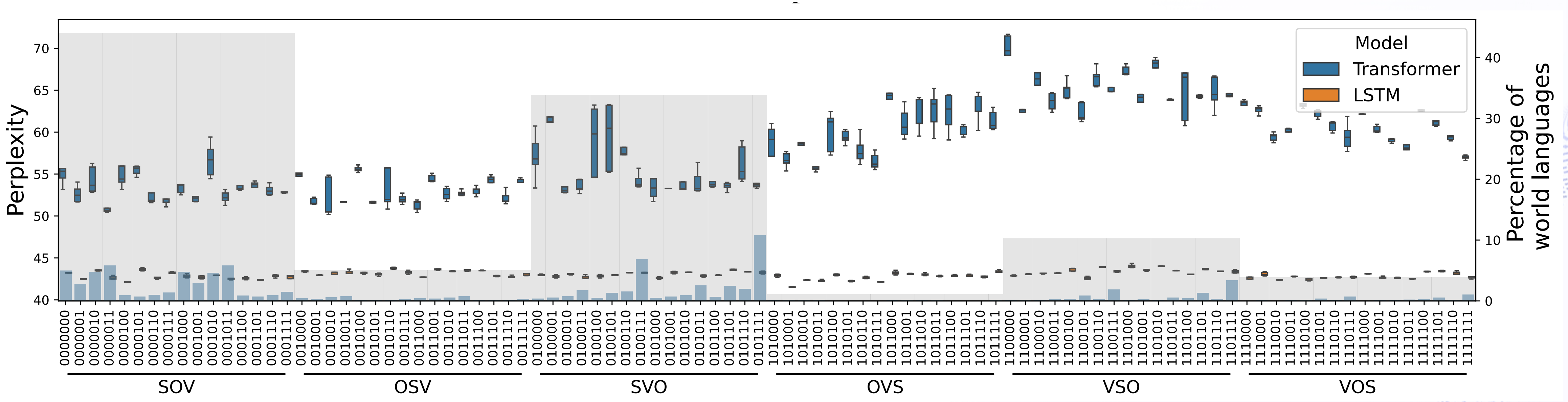
Defining 96 GCGs using 7 binary word order parameters

Randomly generating category sequences, i.e., templates (like PoS tag seq.)

Filtering the templates based on their parsability w/ each GCG parser

Sampling lexicons for each category in each template, creating 50K sentences per grammar

Replication of White&Cotterell 2021



- Mostly reproduced :
  - Transformer exhibits more varied preferences etc.
  - Seemingly, slightly better typological alignment of Transformer, compared to W&C 2021 (but not exactly compared)

Length generalization (EMNLP 2025)

- Significant correlation between simple RNN's inductive bias and typological frequency **when length generalization is evaluated**

Which Word Orders Facilitate Length Generalization in LMs?
An Investigation with GCG-Based Artificial Languages

Nadine El-Naggar* **Tatsuki Kuribayashi*** **Ted Briscoe**
Mohamed bin Zayed University of Artificial Intelligence
{nadine.naggar, tatsuki.kuribayashi, ted.briscoe}@mbzuai.ac.ae

	Short							Medium							Long						
Model	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓
Transformer (PPL ↓)	41.8	41.6	42.3	42.6	42.7	43.3	**−27.7**†	65.2	63.5	64.2	65.9	66.1	65.0	**−10.4**	102.3	99.4	97.9	104.0	107.6	97.9	**−19.2**
LSTM (PPL ↓)	38.7	38.8	38.7	38.4	39.1	38.5	**−14.2**	85.9	91.7	88.0	97.5	92.9	97.9	**−31.0**†	131.9	141.5	160.7	205.5	180.9	207.5	**−33.4**†
RNN (PPL ↓)	40.4	41.0	40.6	39.7	40.1	39.7	13.0	67.8	67.9	66.7	69.6	69.0	69.4	**−17.4**	91.8	94.6	93.2	118.0	109.0	114.2	**−43.1**†
Natural Lang. (Prob. ↑)	0.54	0.04	0.23	0.01	0.12	0.05	-	0.54	0.04	0.23	0.01	0.12	0.05	-	0.54	0.04	0.23	0.01	0.12	0.05	-

Table 3: Average PPLs within each base word order group as well as Pearson's correlation coefficient between PPL and the frequency of respective word order in the world. Negative TA (typological alignment) scores are highlighted in bold. Statistical significance of correlation coefficient (p<0.05) is marked with †.

Curriculum learning effect (under review)

- Which language is easier to learn in cognitively more plausible learning scenario?
  - The importance of “starting small” [Elman, 1993]

- Length-based curriculum learning as additional environmental bias
  - Is there interaction effect between model’s inductive bias and curriculum learning bias? --- **Yes**
  - Under curriculum learning, less aligned with typological tendencies…

		Short							Medium							Long						
Model	CL	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓
Transformer		41.8	41.6	42.3	42.6	42.7	43.3	**−27.7**†	65.2	63.5	64.2	65.9	66.1	65.0	**−10.4**	102.3	99.4	97.9	104.0	107.6	97.9	**−19.2**
Transformer	✓	50.2	48.8	49.3	52.6	53.3	53.4	**−22.3**†	63.2	61.8	62.9	62.8	68.9	64.7	**−5.9**	95.1	112.9	83.1	89.9	106.2	96.2	**−18.0**†
LSTM		38.7	38.8	38.7	38.4	39.1	38.5	**−14.2**	85.9	91.7	88.0	97.5	92.9	97.9	**−31.0**†	131.9	141.5	160.7	205.5	180.9	207.5	**−33.4**†
LSTM	✓	44.4	44.9	44.5	43.9	44.2	44.4	16.1	95.6	102.3	101.9	112.4	119.8	117.7	**−20.1**†	113.8	122.0	119.3	153.6	151.9	163.7	**−32.3**†
RNN		40.4	41.0	40.6	39.7	40.1	39.7	13.0	67.8	67.9	66.7	69.6	69.0	69.4	**−17.4**	91.8	94.6	93.2	118.0	109.0	114.2	**−43.1**†
RNN	✓	45.9	47.4	45.1	44.5	45.0	44.8	14.2	80.3	84.5	89.7	76.6	91.7	78.1	21.2	102.9	107.2	113.0	117.6	113.7	117.8	**−20.2**†
NL (Prob. ↑)		0.54	0.04	0.23	0.01	0.12	0.05	-	0.54	0.04	0.23	0.01	0.12	0.05	-	0.54	0.04	0.23	0.01	0.12	0.05	-

Table 1: Average PPLs within each base word order group as well as Pearson’s correlation coefficient between PPL and the frequency of respective word orders in the world. Negative TA (typological alignment) scores are highlighted in bold. Statistical significance of correlation coefficient ($p<0.05$) is marked with †.

Confidential

RE: My research topics

- **Cognitive modeling**
  - **Are larger language models cognitively plausible?**
- Interpretability
  - What information do LMs truly pay attention to?
- Linguistic typology and language acquisition
  - What kind of language design is easy for LMs to learn?
    - collaborated with Alex as well!
- Past: Automated writing assistance

Lower Perplexity is Not Always Human-Like

Tatsuki Kuribayashi[1,2], Yohei Oseki[3,4], Takumi Ito[1,2],
Ryo Yoshida[3], Masayuki Asahara[5], Kentaro Inui[1,4]
[1]Tohoku University [2]Langsmith Inc. [3]University of Tokyo [4]RIKEN [5]NINJAL
{kuribayashi, takumi.ito.c4, inui}@tohoku.ac.jp ,
{oseki, yoshiryo0617}@g.ecc.u-tokyo.ac.jp ,masayu-a@ninjal.ac.jp

Context Limitations Make Neural Language Models More Human-Like

Tatsuki Kuribayashi[1,2] Yohei Oseki[3,4] Ana Brassard[1,4] Kentaro Inui[1,4]
[1]Tohoku University [2]Langsmith Inc. [3]University of Tokyo [4]RIKEN
{kuribayashi, inui}@tohoku.ac.jp
oseki@g.ecc.u-tokyo.ac.jp ana.brassard@riken.jp

Attention is Not Only a Weight:
Analyzing Transformers with Vector Norms

Goro Kobayashi[1] Tatsuki Kuribayashi[1,2] Sho Yokoi[1,3] Kentaro Inui[1,3]
[1] Tohoku University [2] Langsmith Inc. [3] RIKEN
{goro.koba, kuribayashi, yokoi, inui}@ecei.tohoku.ac.jp

Which Word Orders Facilitate Length Generalization in LMs?
An Investigation with GCG-Based Artificial Languages

Nadine El-Naggar* Tatsuki Kuribayashi* Ted Briscoe
Mohamed bin Zayed University of Artificial Intelligence
{nadine.naggar, tatsuki.kuribayashi, ted.briscoe}@mbzuai.ac.ae