

The Quest for A Science of Language Models

AAAI 2025 Tutorial

February 26, 2025, 2:00pm-3:45pm

Pennsylvania Convention Center Room 113A, Philadelphia, PA, U.S.A.



Tutorial Homepage: <https://glaciohound.github.io/Science-of-LLMs-Tutorial/>
Chi Han, Ph.D. Student @ UIUC, <https://glaciohound.github.io/>
Heng Ji, Professor @ UIUC, <http://blender.cs.illinois.edu/hengji.html>

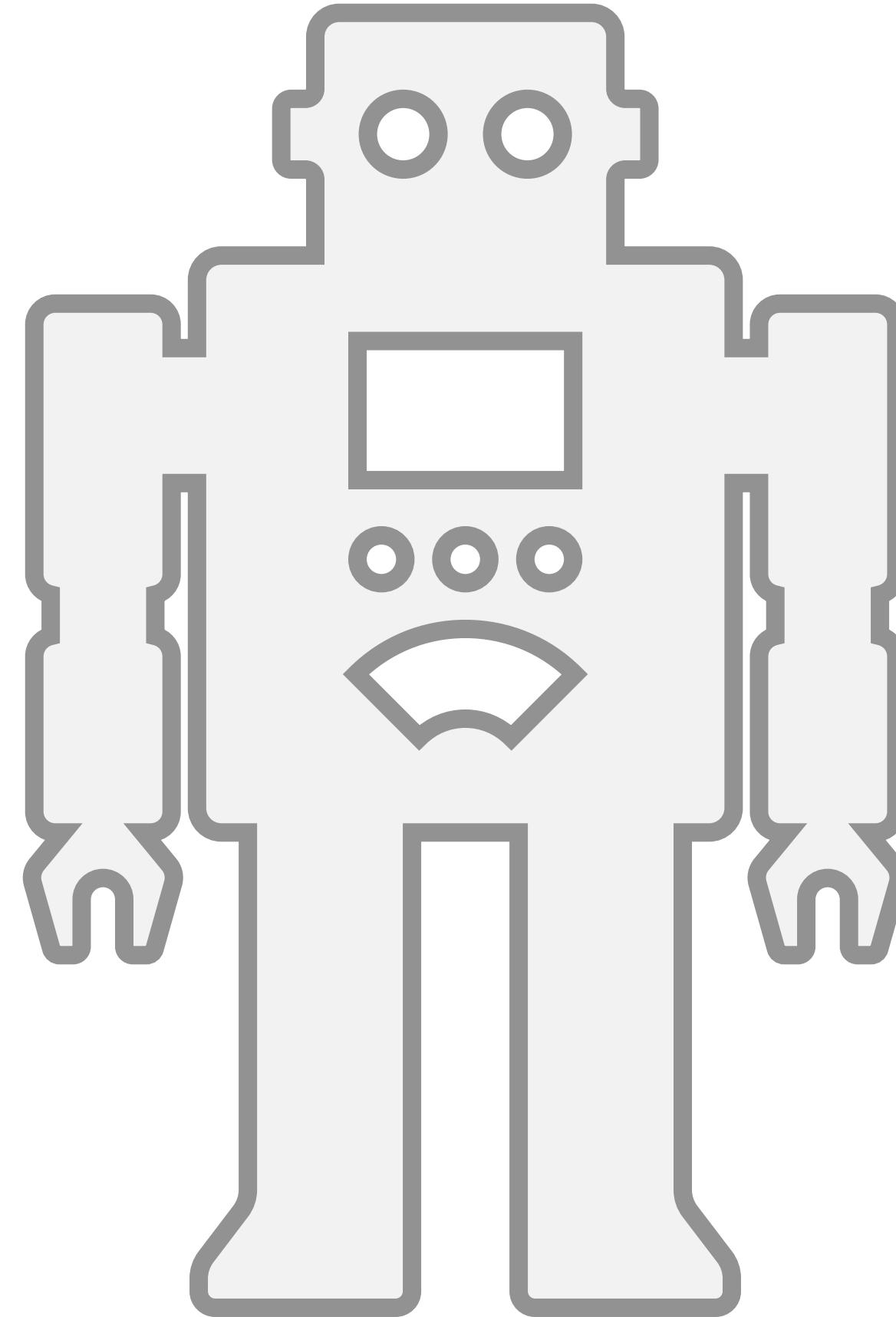


UNIVERSITY OF
ILLINOIS
URBANA - CHAMPAIGN

(LMs: abbreviation for “language models”)

We Increasingly Rely on LMs

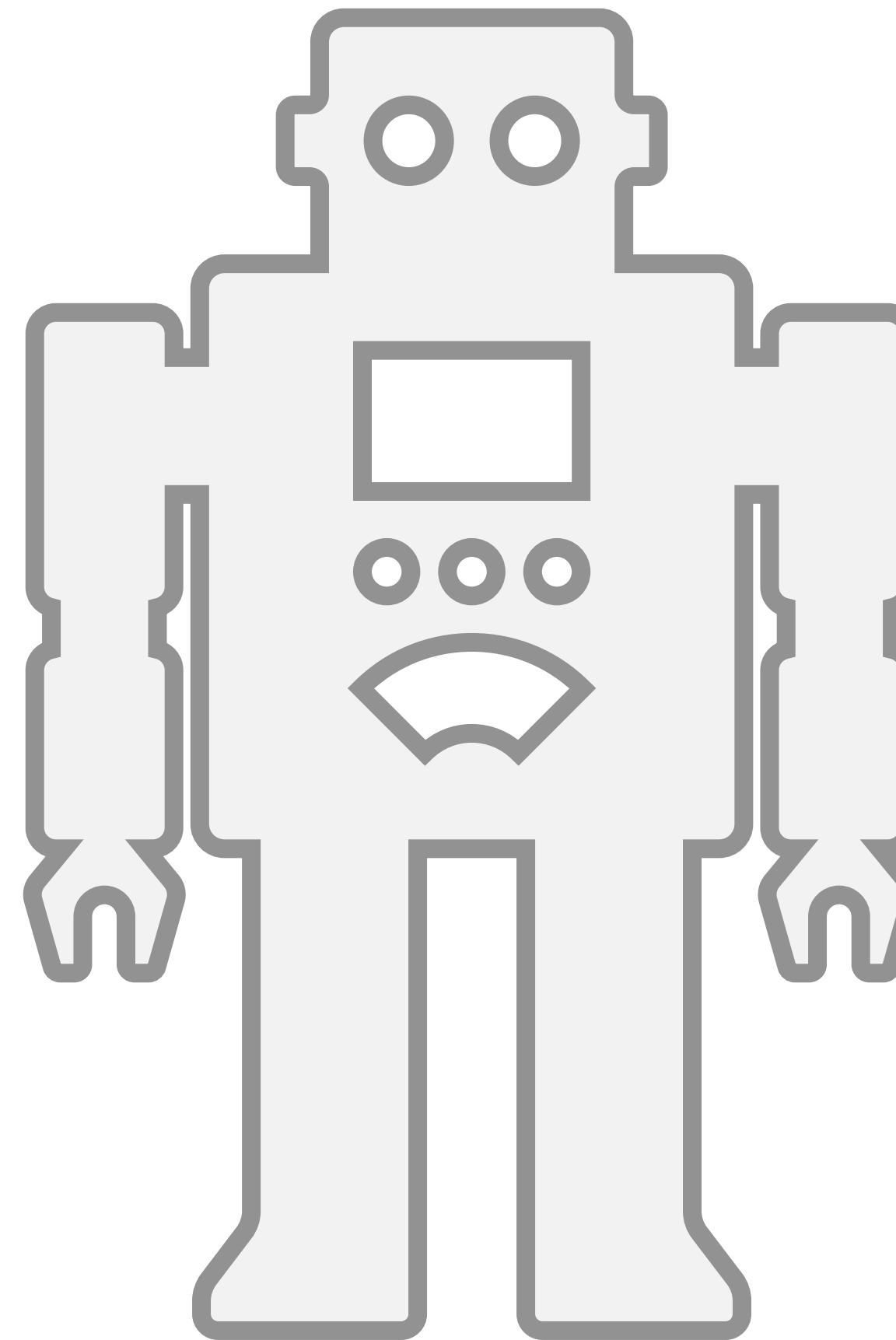
yet we still do not fully understand them.



(LMs: abbreviation for “language models”)

We Increasingly Rely on LMs

yet we still do not fully understand them.



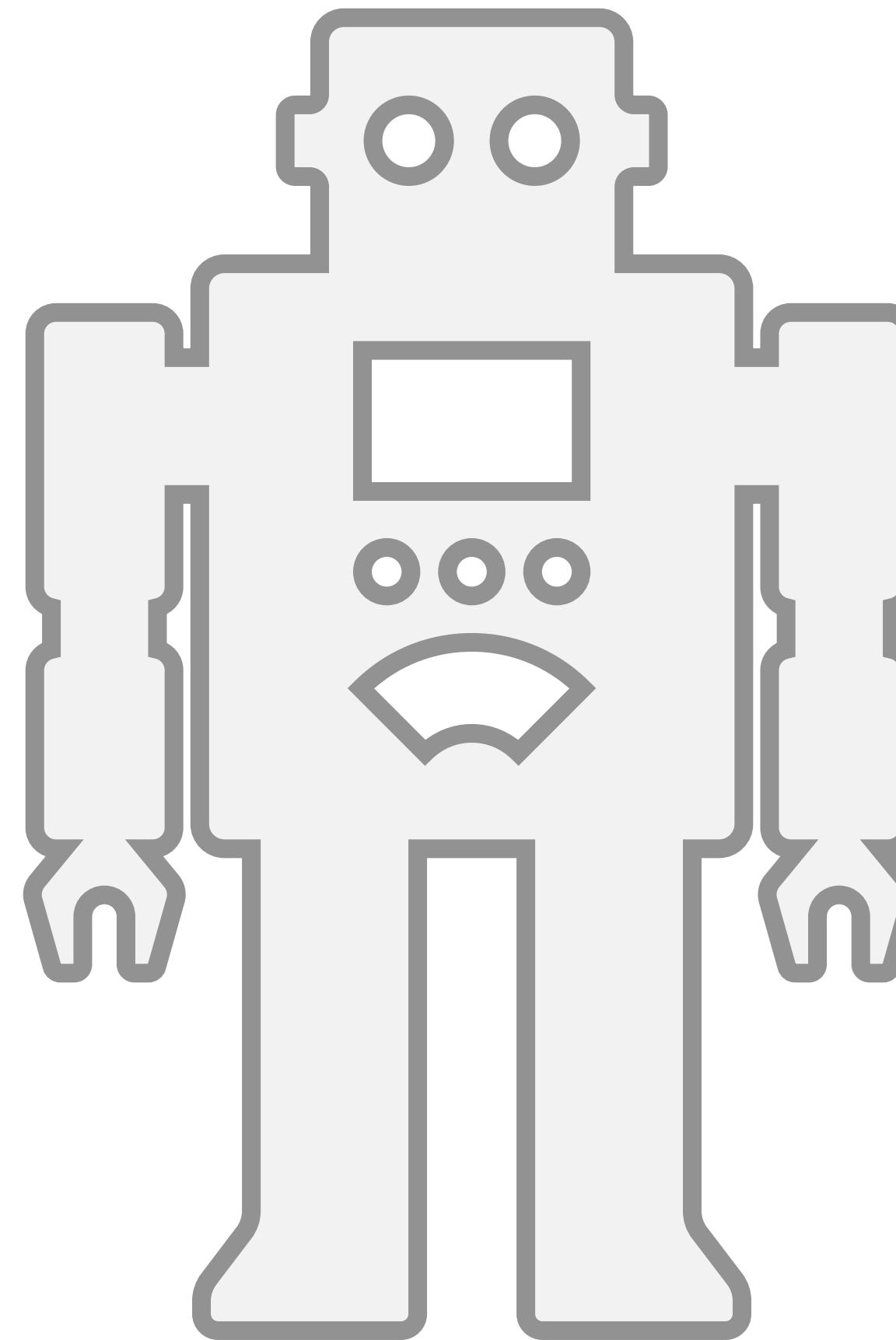
“Can you solve this problem for me? It is too hard for me”

(Are they intelligent enough to solve it, or do they pretend to be doing so?)

(LMs: abbreviation for “language models”)

We Increasingly Rely on LMs

yet we still do not fully understand them.



“Can you solve this problem for me? It is too hard for me”

(Are they intelligent enough to solve it, or do they pretend to be doing so?)

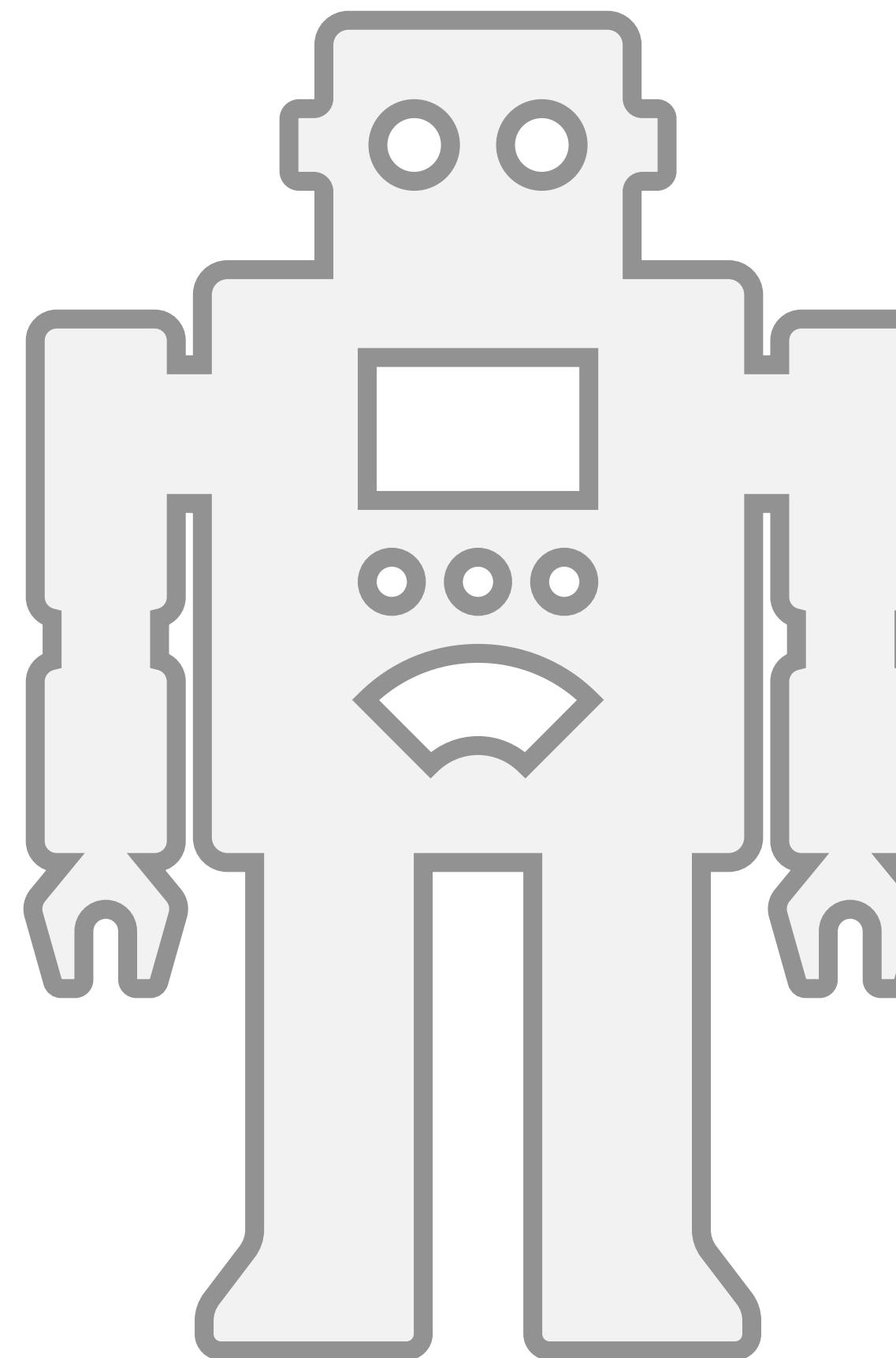
“Could you tell me of all the kings who have ruled over Europe?”

(do they know all these knowledge, or are they sometimes guessing?)

(LMs: abbreviation for “language models”)

We Increasingly Rely on LMs

yet we still do not fully understand them.



“Can you solve this problem for me? It is too hard for me”

(Are they intelligent enough to solve it, or do they pretend to be doing so?)

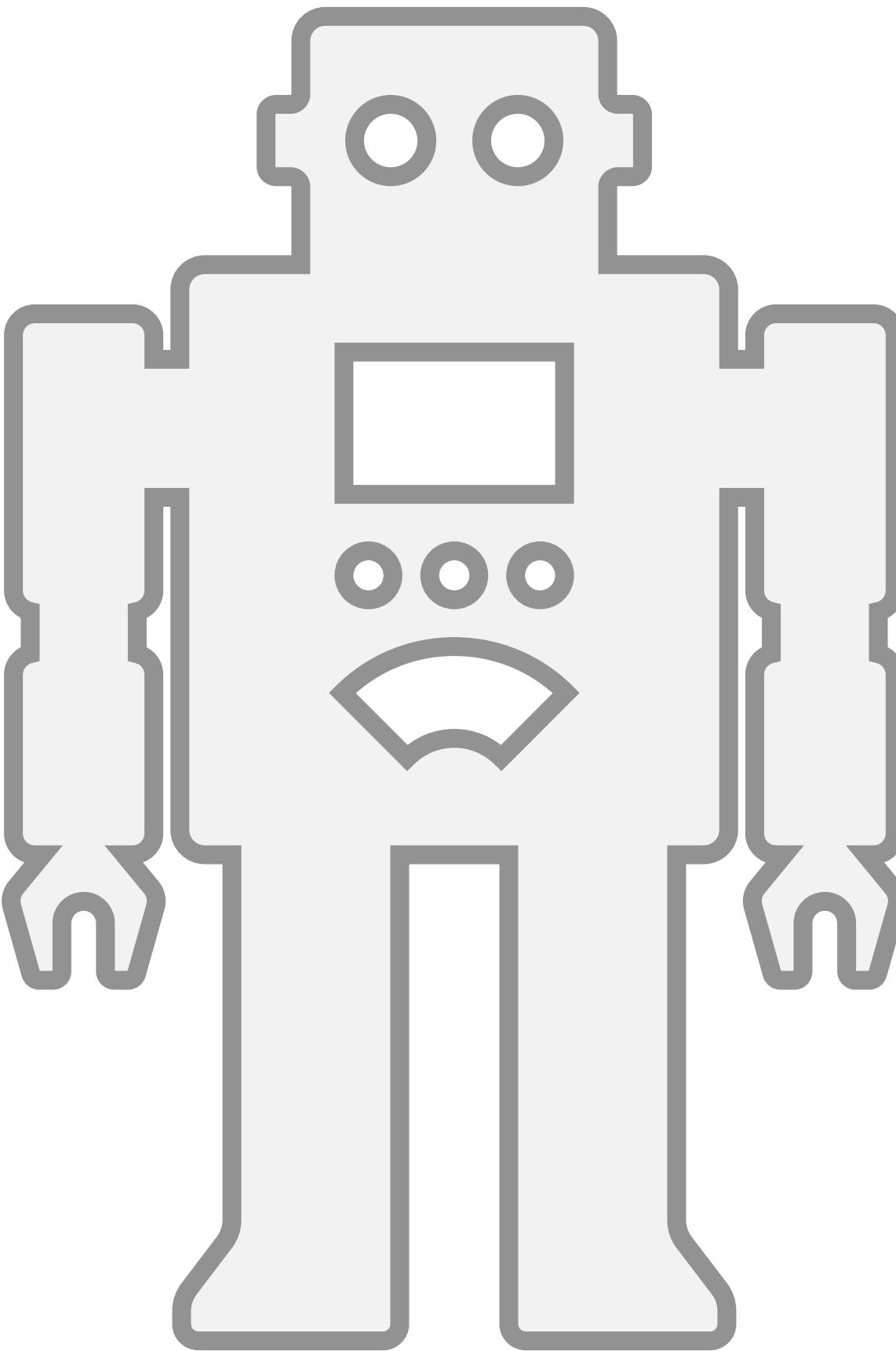
“Could you tell me of all the kings who have ruled over Europe?”

(do they know all these knowledge, or are they sometimes guessing?)

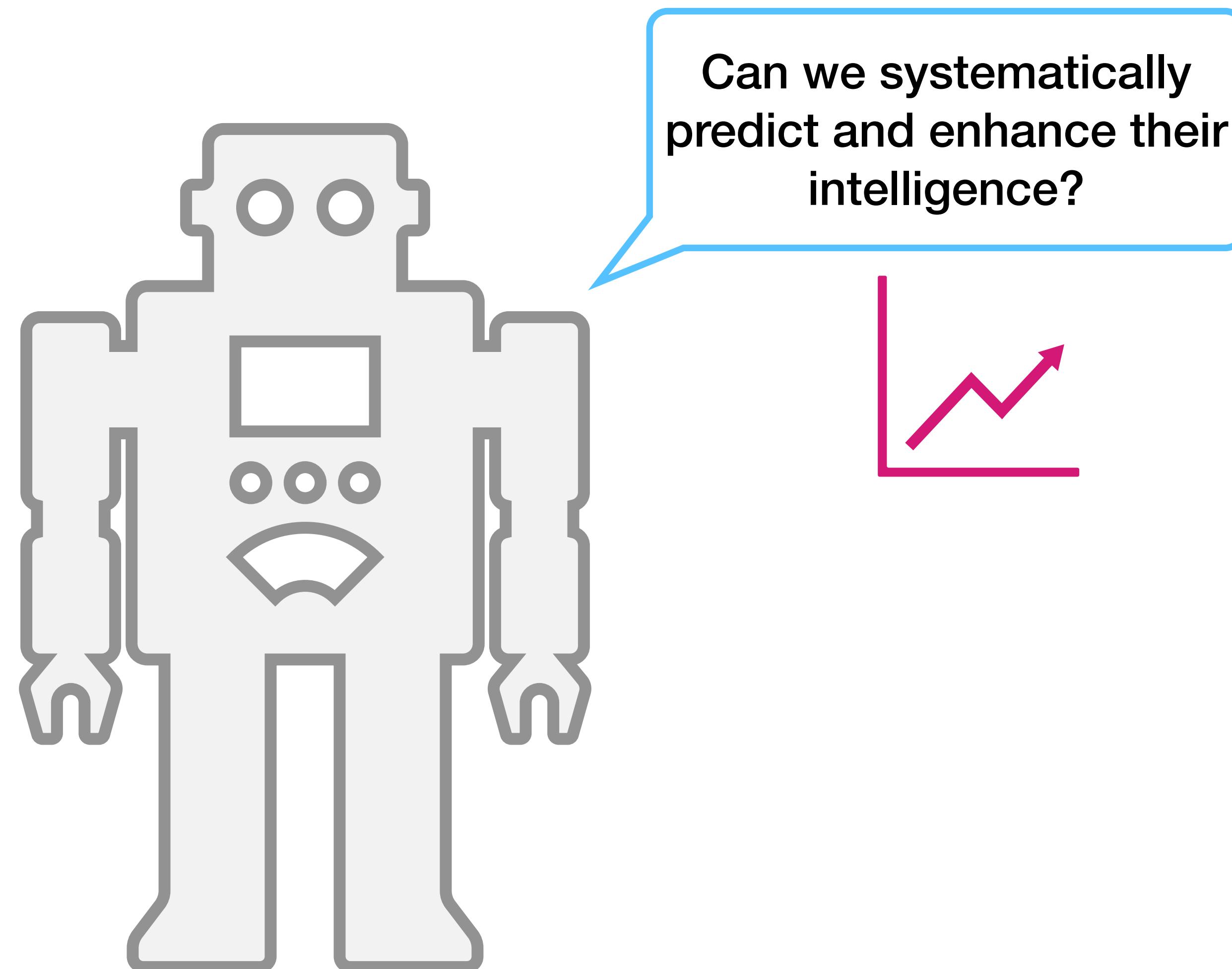
“Why did you perform bad on this task?”

(What are the causes of their drawbacks?)

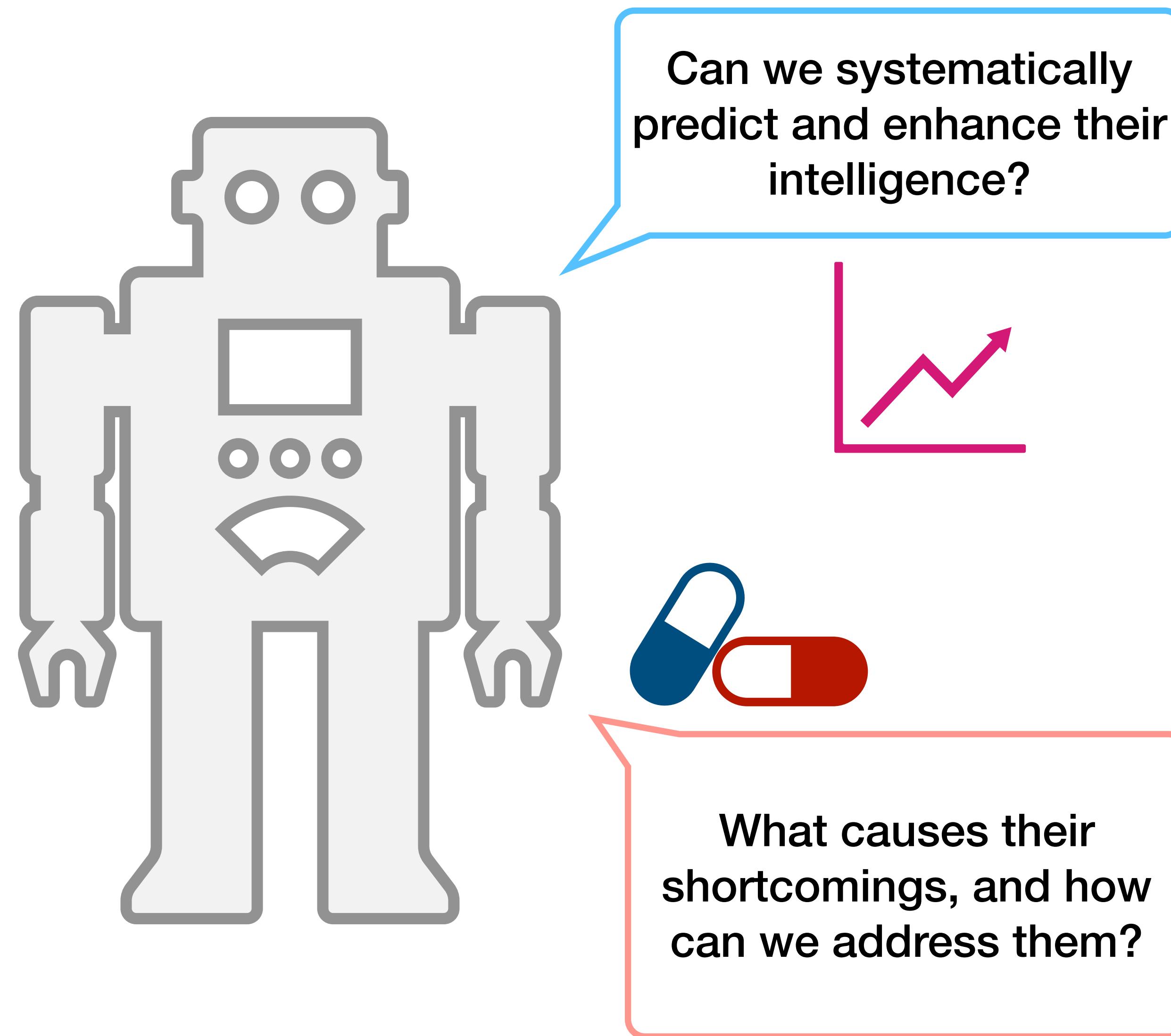
Can We Systematically Describe How LMs Behave?



Can We Systematically Describe How LMs Behave?

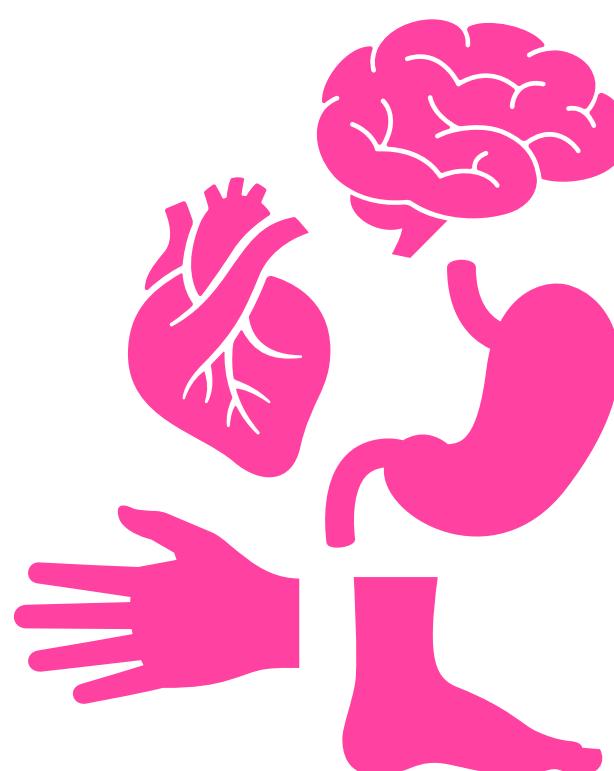


Can We Systematically Describe How LMs Behave?

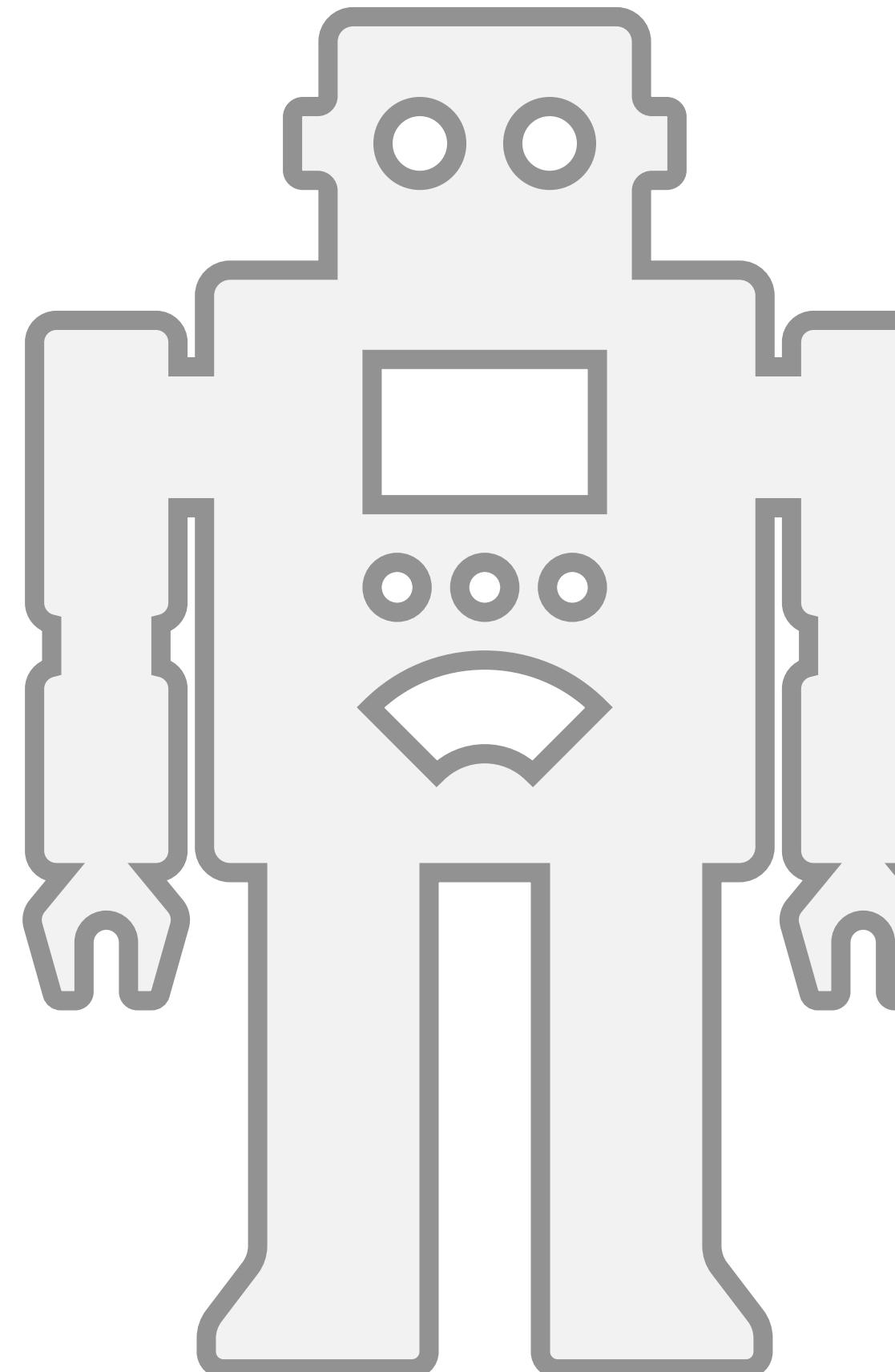


Can We Systematically Describe How LMs Behave?

How do their components function?



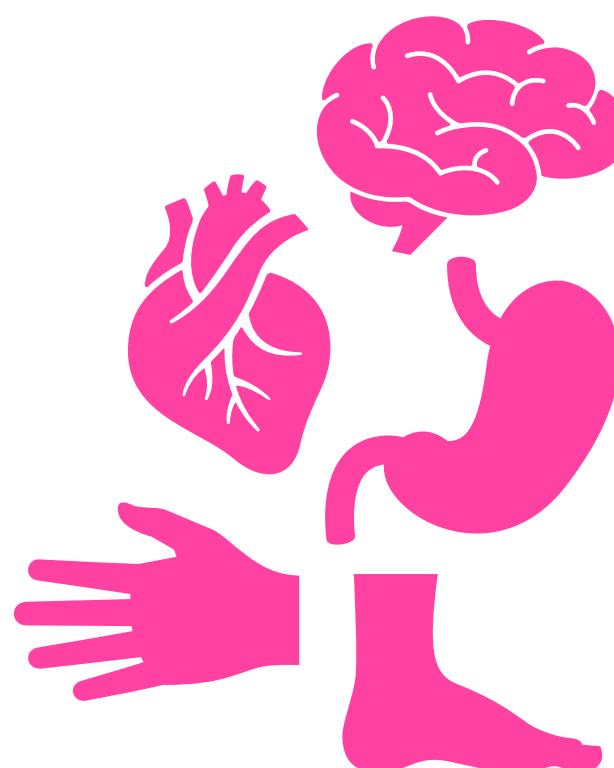
Can we systematically predict and enhance their intelligence?



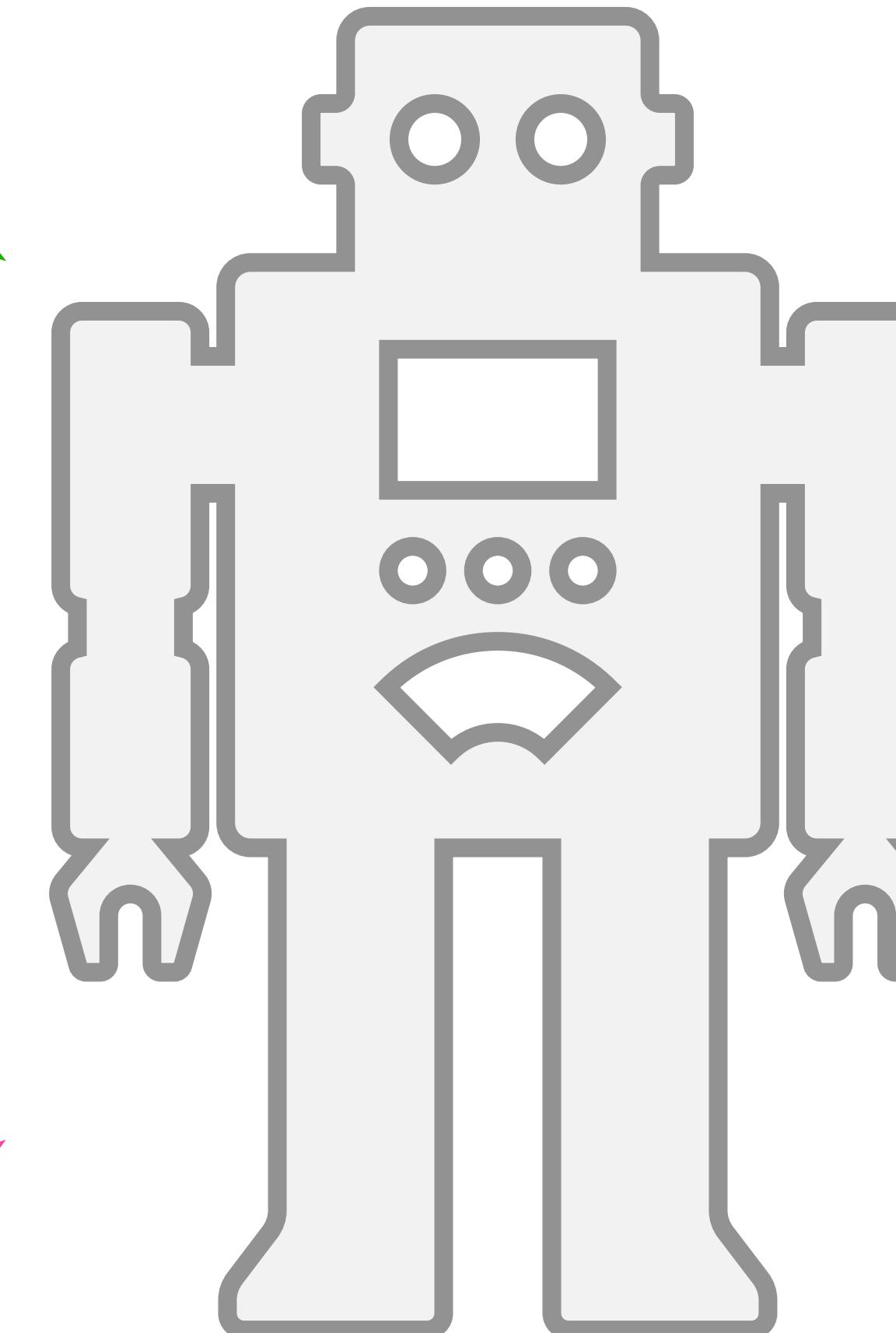
What causes their shortcomings, and how can we address them?

Can We Systematically Describe How LMs Behave?

How do their components function?



How do LMs reason and utilize knowledge?



Can we systematically predict and enhance their intelligence?



What causes their shortcomings, and how can we address them?

Why Do We Need A New Science?

New sciences often emerge as a result of scaling up old sciences

Machine Learning —→ **Deep Learning** —→ **Language Models**

PAC theory,
optimization,
...

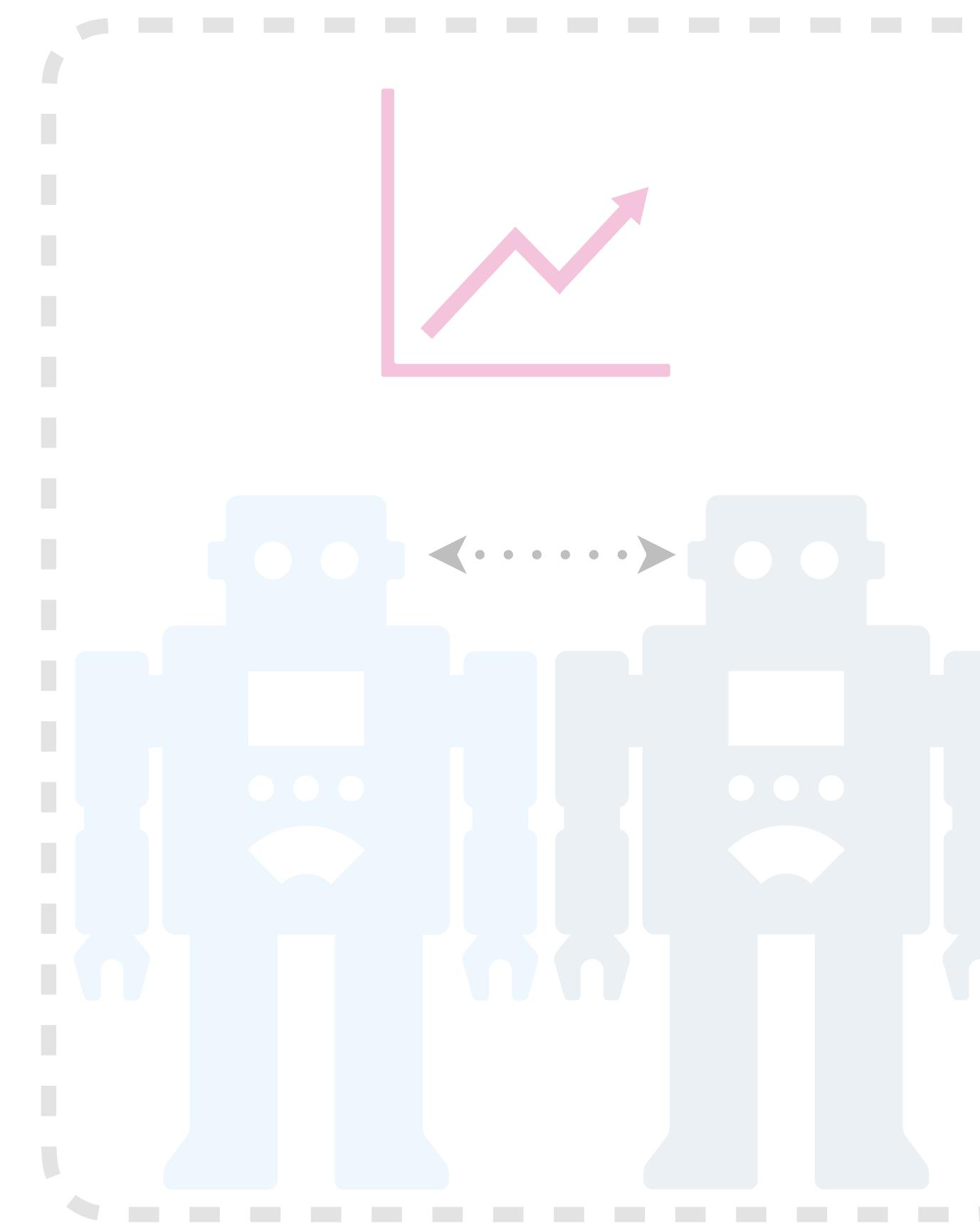
Gradient Descent,
Neural Tangent
Kernel, ...

A Sciences of LMs

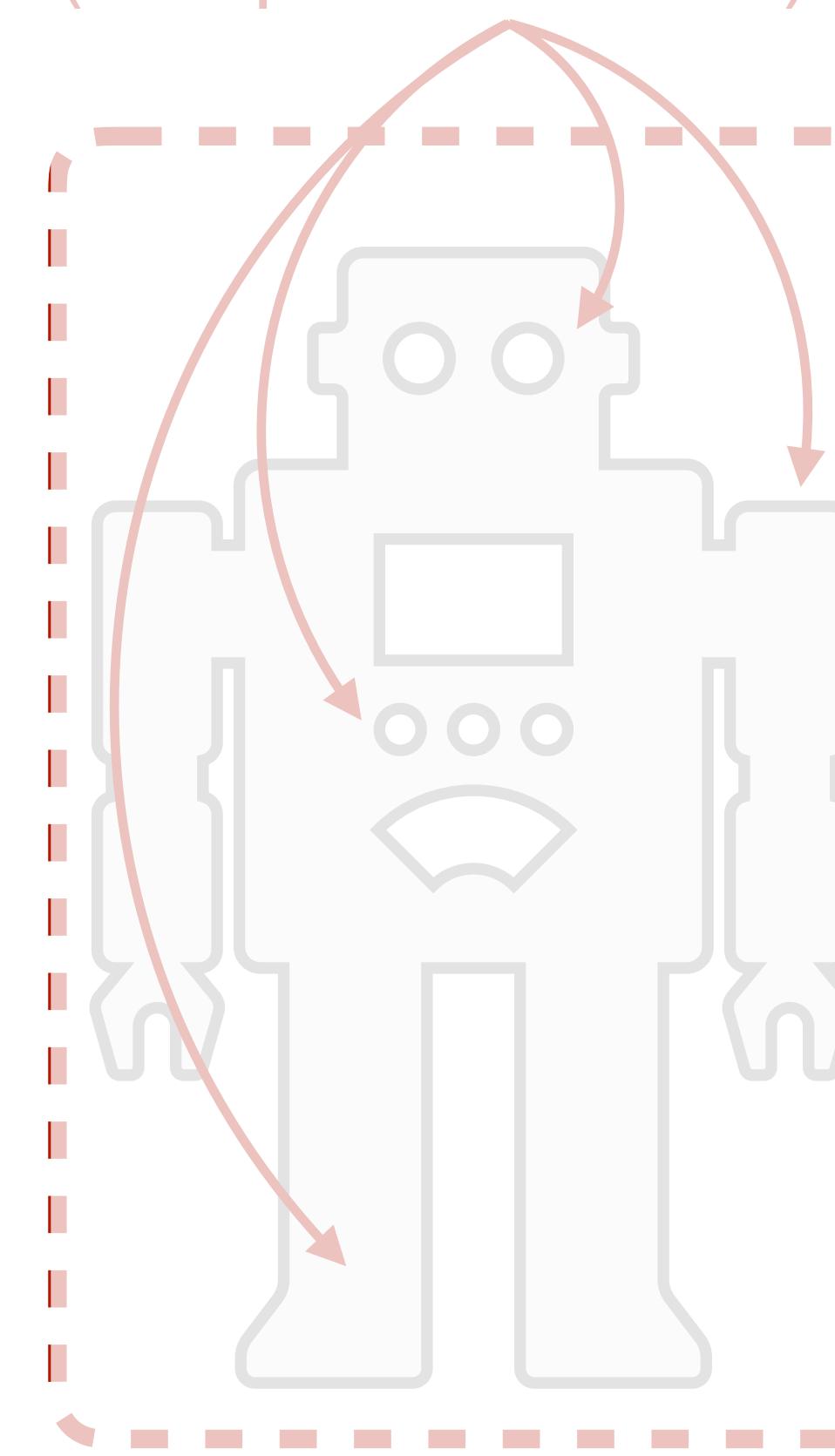
Spectrum of Sciences of LMs



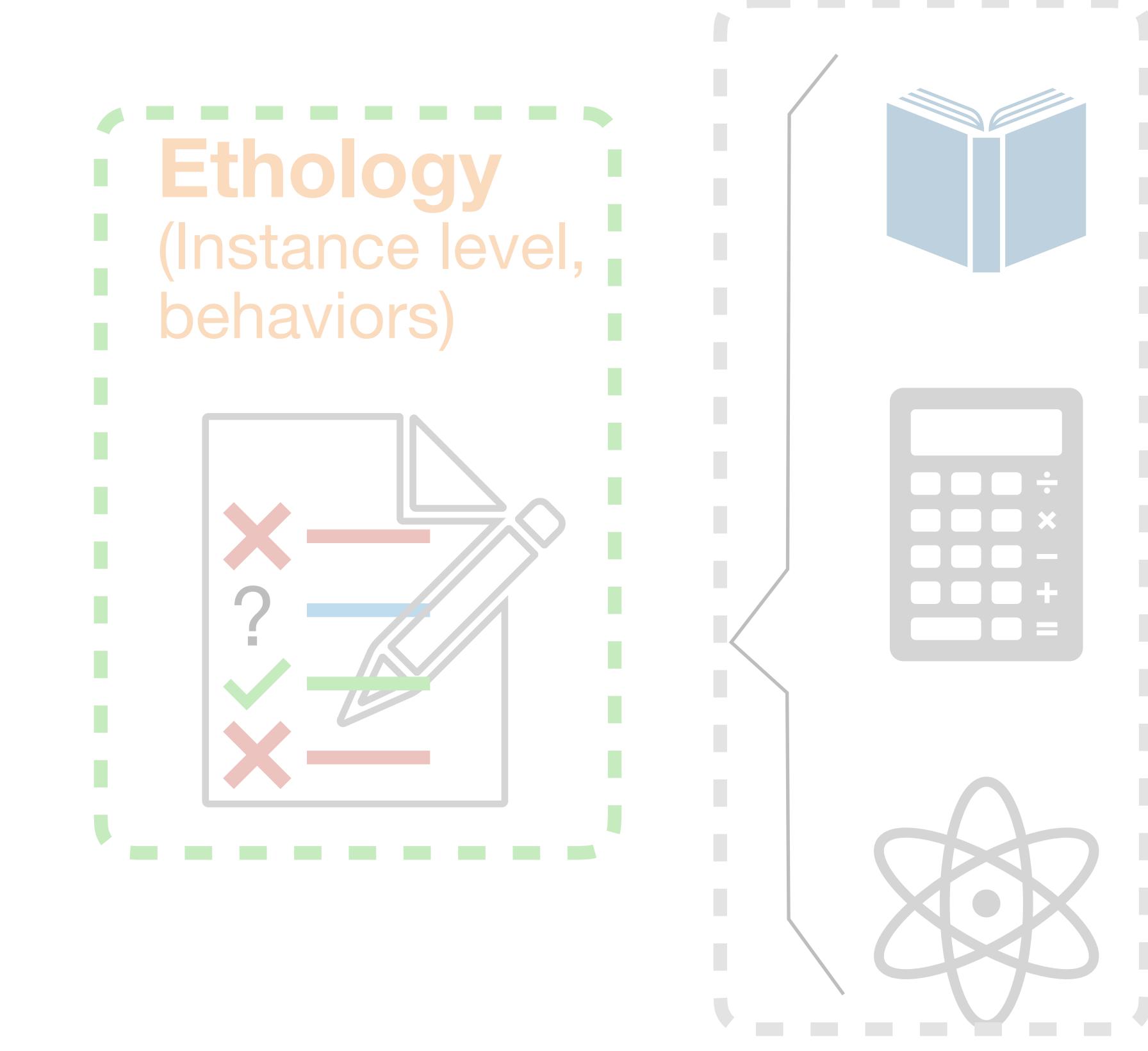
Physics of LMs
(laws at population level)



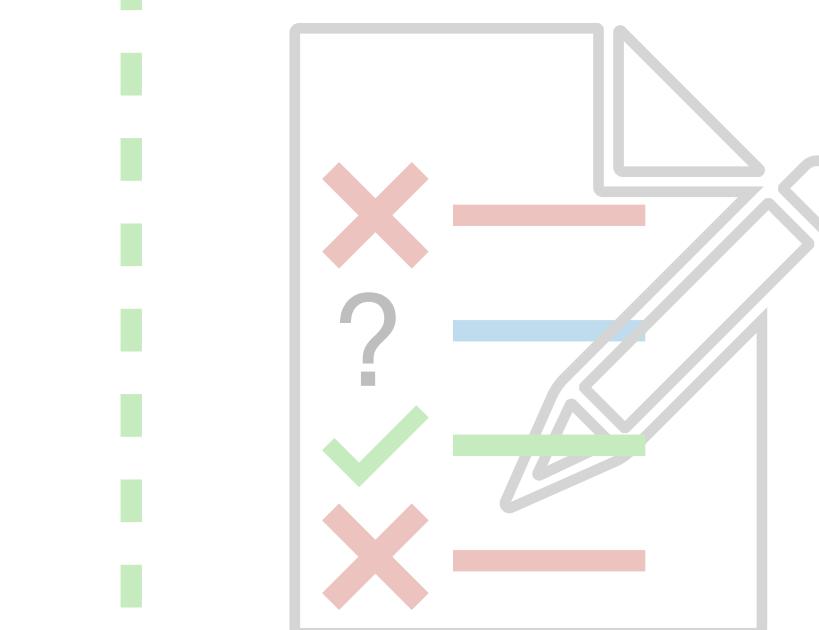
Physiology of LMs
(components-level)



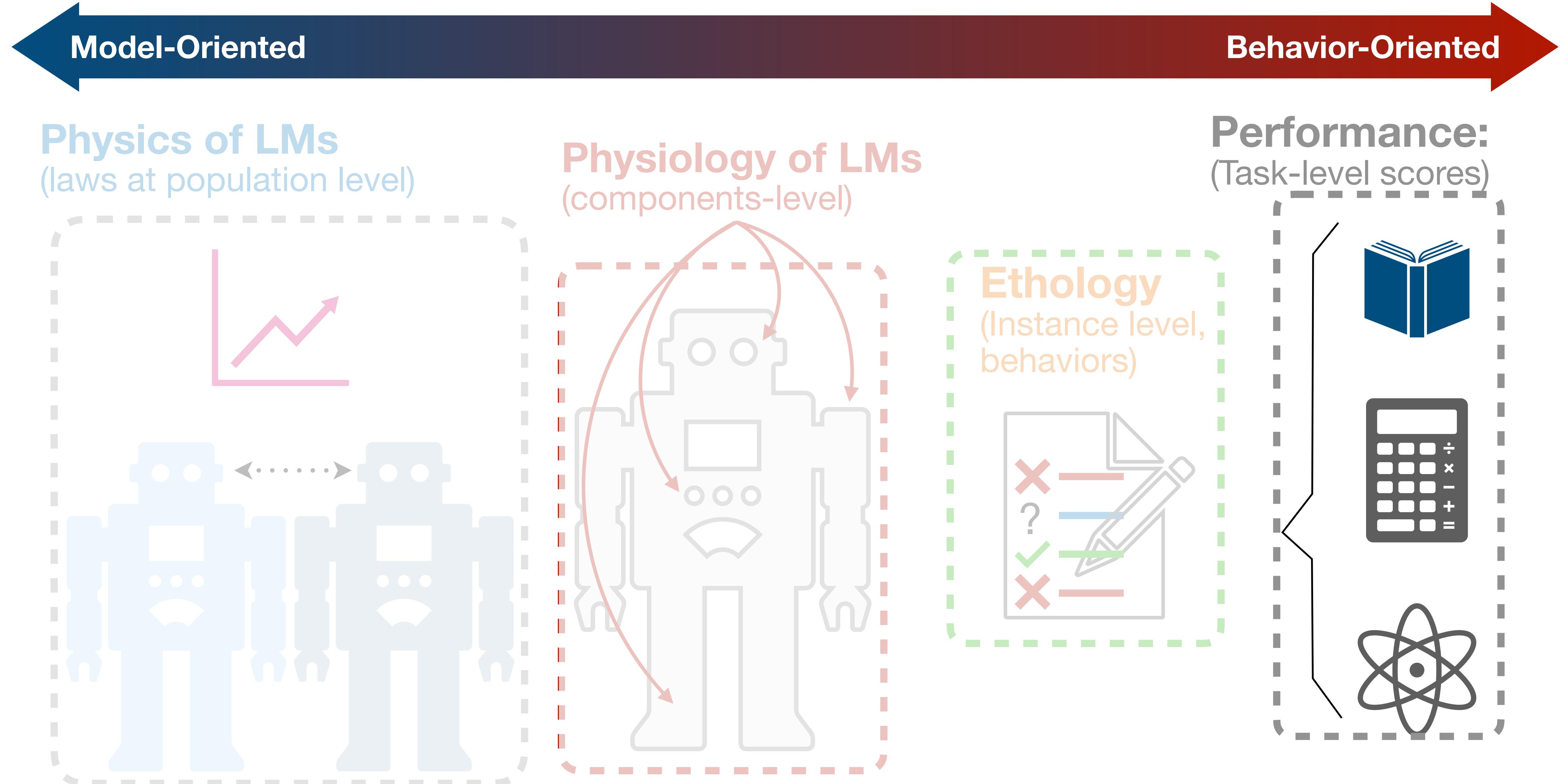
Performance:
(Task-level scores)



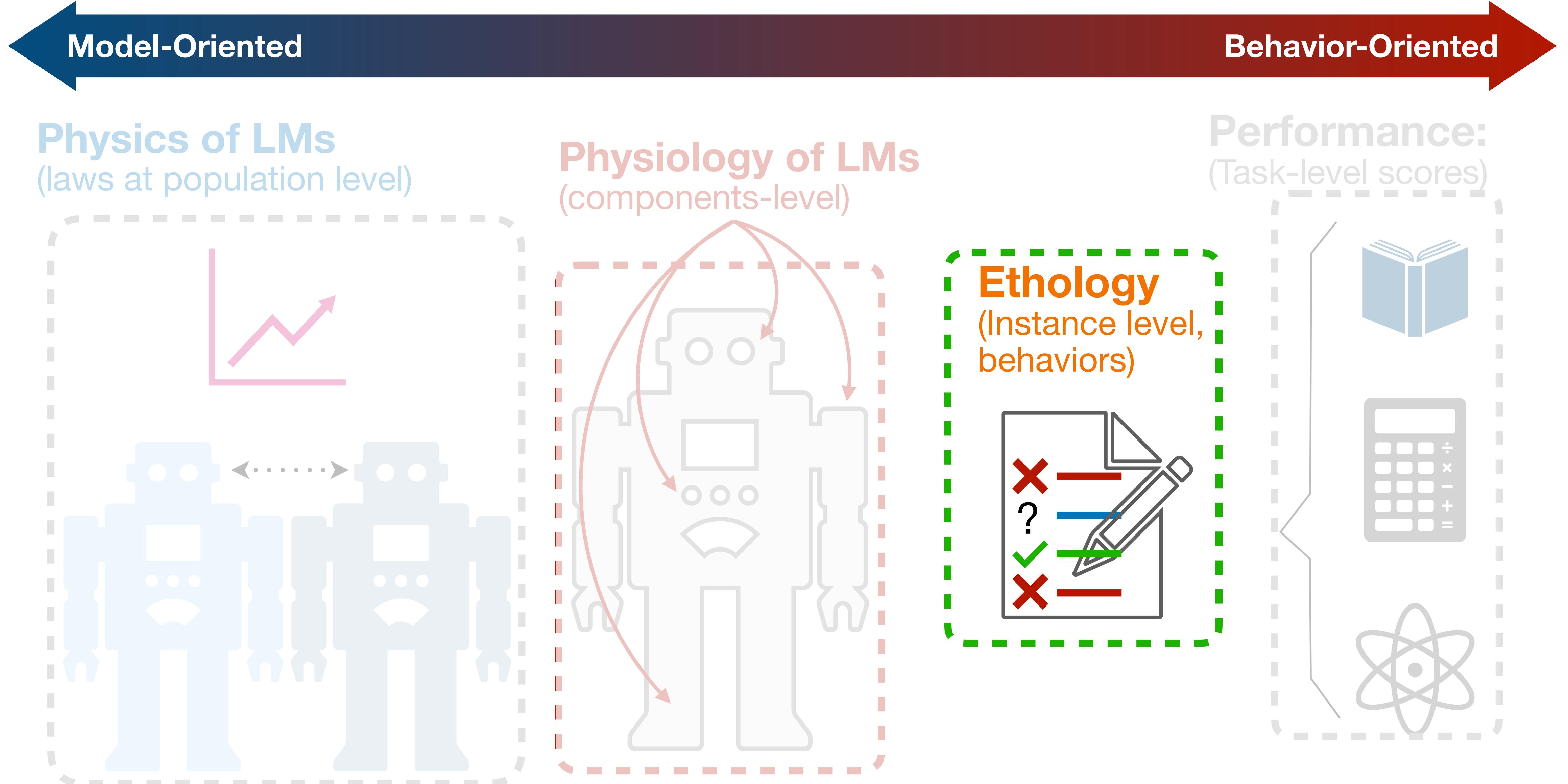
Ethology
(Instance level,
behaviors)



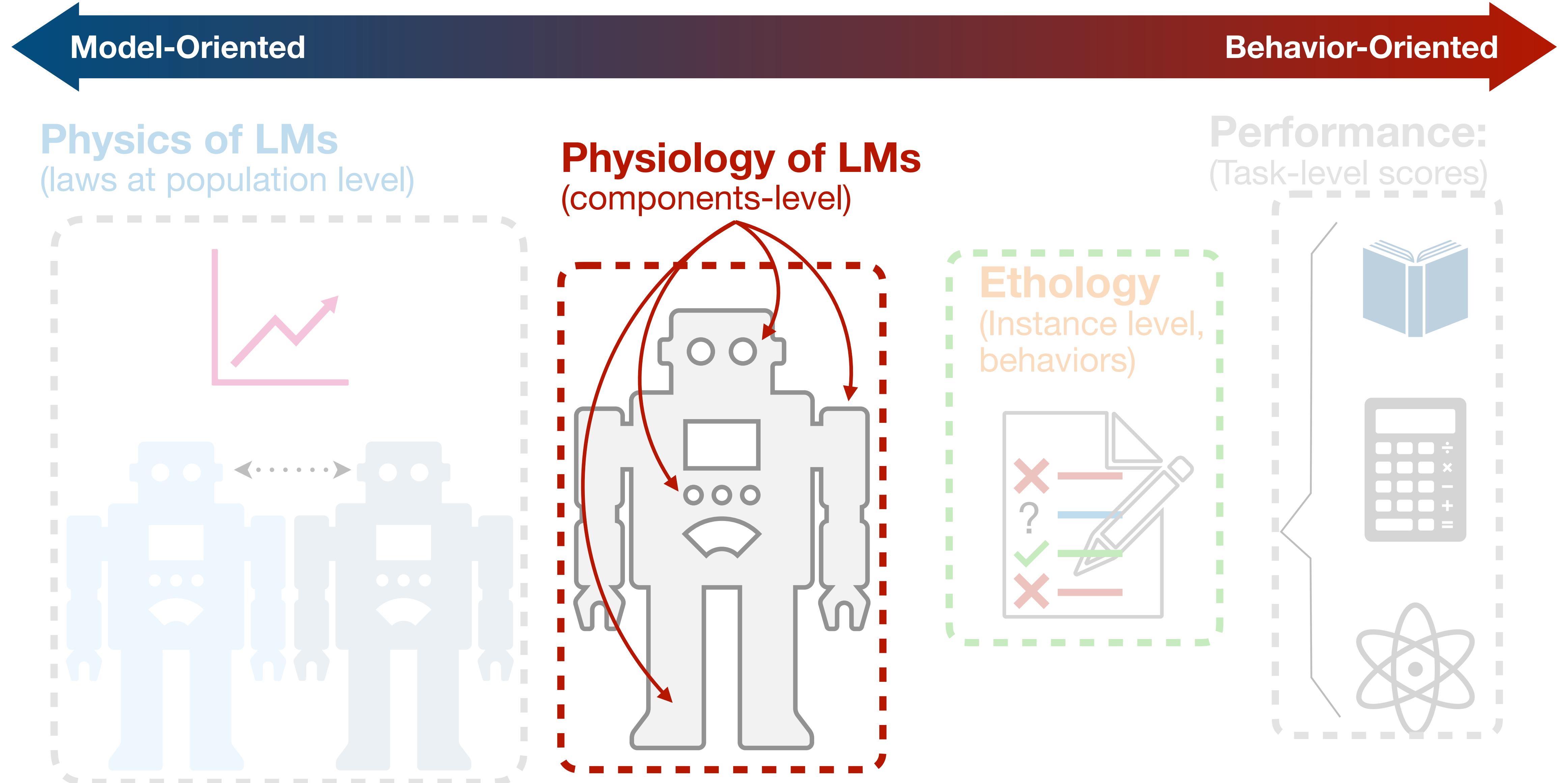
Spectrum of Sciences of LMs



Spectrum of Sciences of LMs



Spectrum of Sciences of LMs



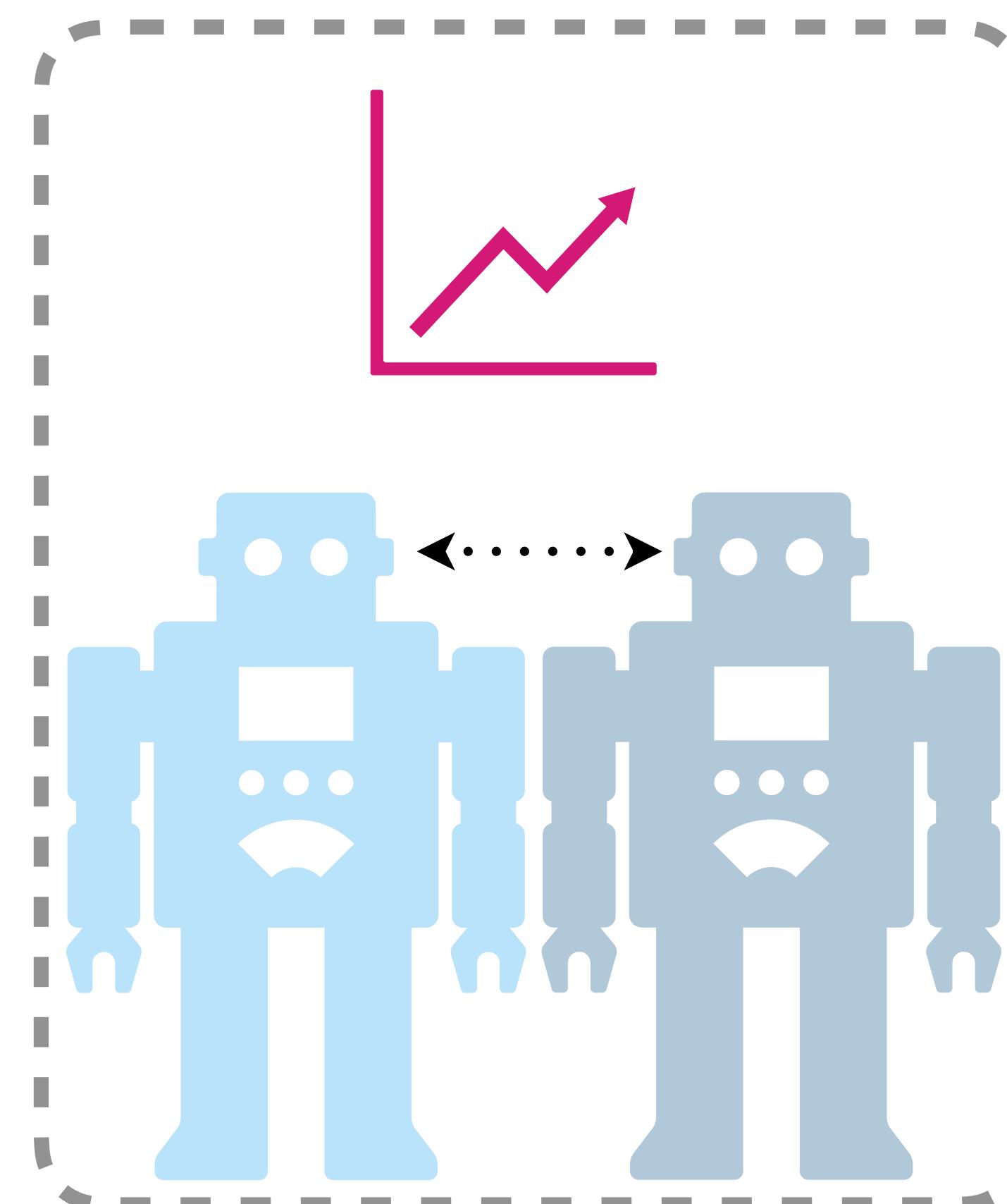
Spectrum of Sciences of LMs

Model-Oriented

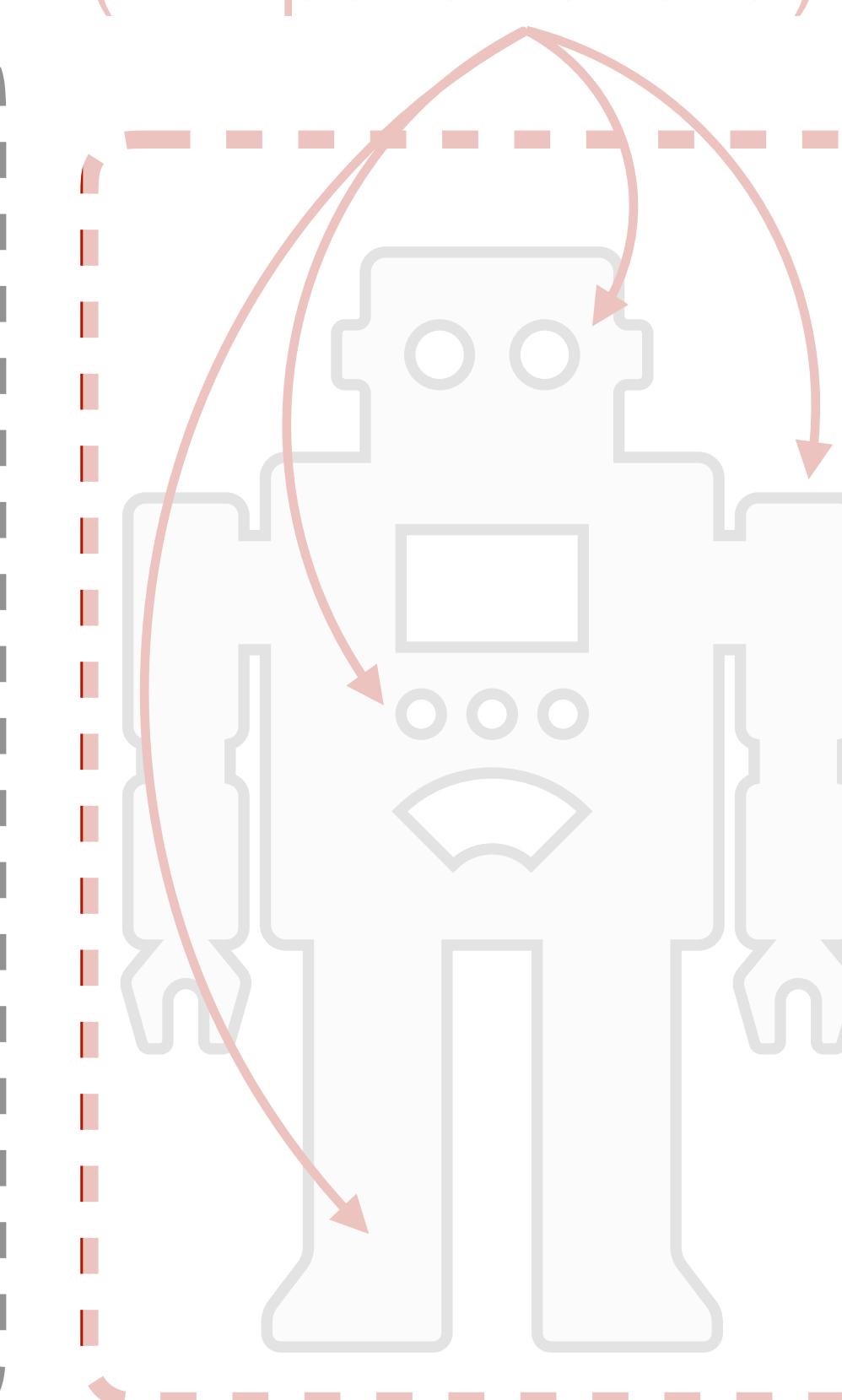
Behavior-Oriented

Physics of LMs

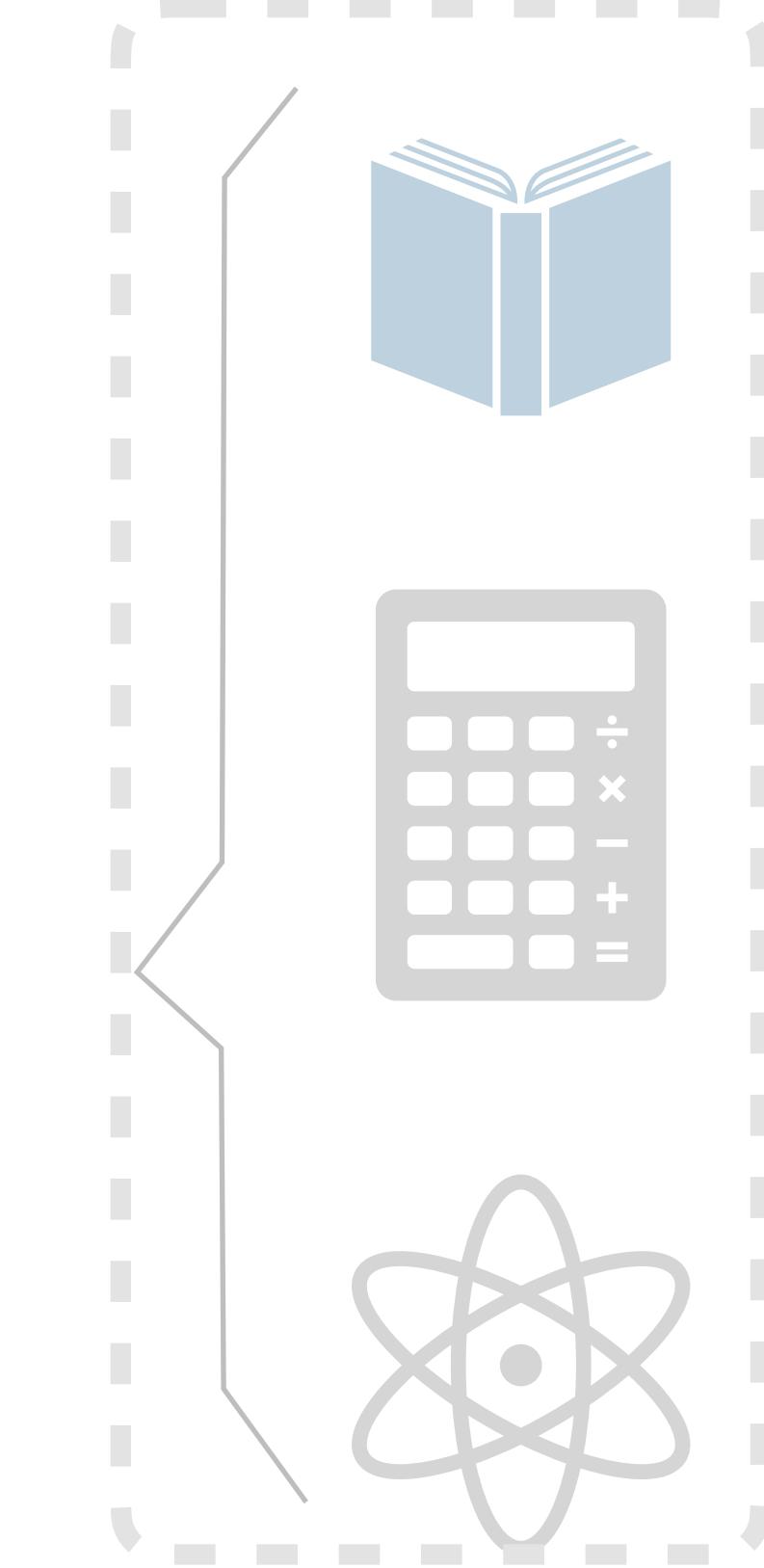
(laws at population level)



Physiology of LMs (components-level)

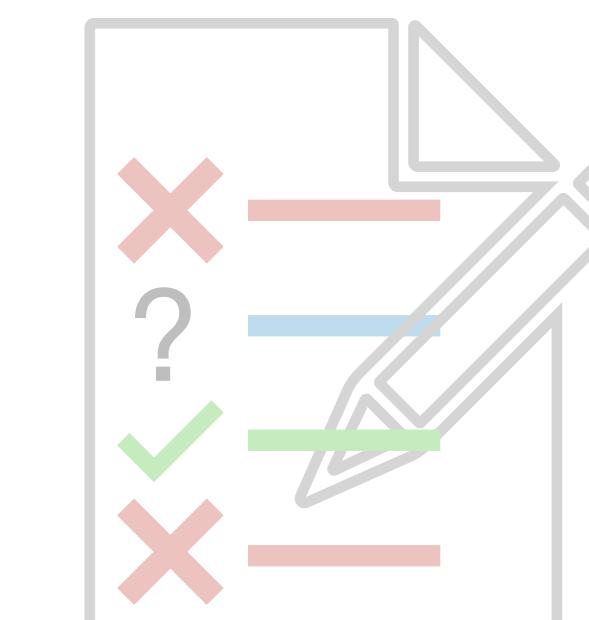


Performance: (Task-level scores)



Ethology

(Instance level
behaviors)



Scientific Principles for a Science of LMs

Scientific Principles for a Science of LMs

- **Generality** across model sizes, architectures, training details, and randomness.

Scientific Principles for a Science of LMs

- **Generality** across model sizes, architectures, training details, and randomness.
- **Simplicity** to avoid meta-overfitting

Scientific Principles for a Science of LMs

- **Generality** across model sizes, architectures, training details, and randomness.
- **Simplicity** to avoid meta-overfitting
- **Interpretability** with consistent mechanistic insight

Scientific Principles for a Science of LMs

- **Generality** across model sizes, architectures, training details, and randomness.
- **Simplicity** to avoid meta-overfitting
- **Interpretability** with consistent mechanistic insight
- **Predictive power** on new phenomena

Scientific Principles for a Science of LMs

- **Generality** across model sizes, architectures, training details, and randomness.
- **Simplicity** to avoid meta-overfitting
- **Interpretability** with consistent mechanistic insight
- **Predictive power** on new phenomena
- **Deriving solutions** for LM-related challenges

Scientific Principles for a Science of LMs

- **Generality** across model sizes, architectures, training details, and randomness.
- **Simplicity** to avoid meta-overfitting
- **Interpretability** with consistent mechanistic insight
- **Predictive power** on new phenomena
- **Deriving solutions** for LM-related challenges

In practice, achieving all these principles is challenging, but the more we achieve, the better!

Tutorial Outline

- **Part 1 (Ethology): How Do LMs Behave?**

- **Syntax:** How do LMs work with syntax
- **Knowledge:** Where is knowledge stored
- **Reasoning:** How is reasoning conducted

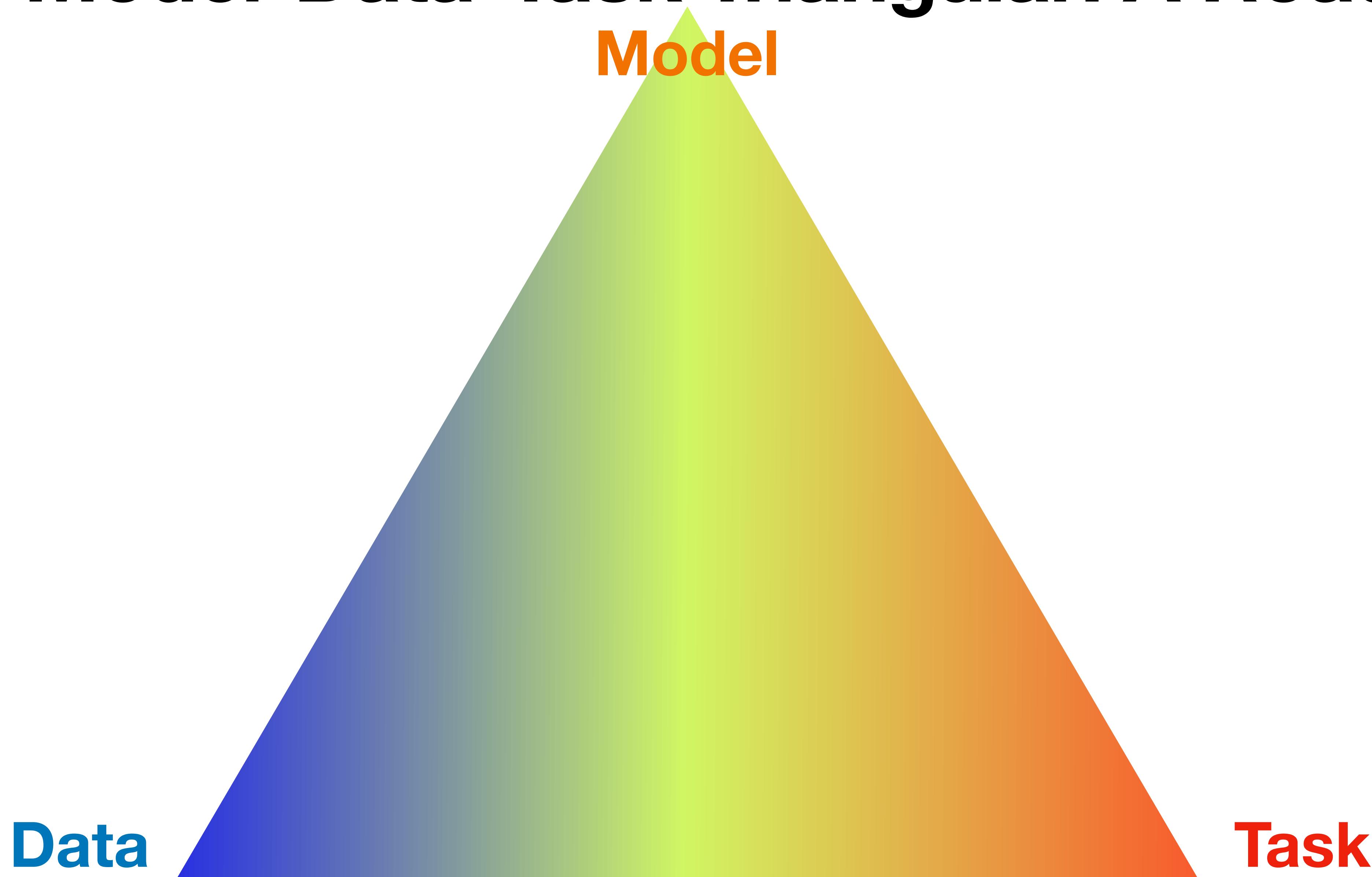
- **Part 2 (Physiology): What Roles Do Components Play?**

- **Attention:** Attention, position and context
- **Embeddings:** What is the function of word embeddings

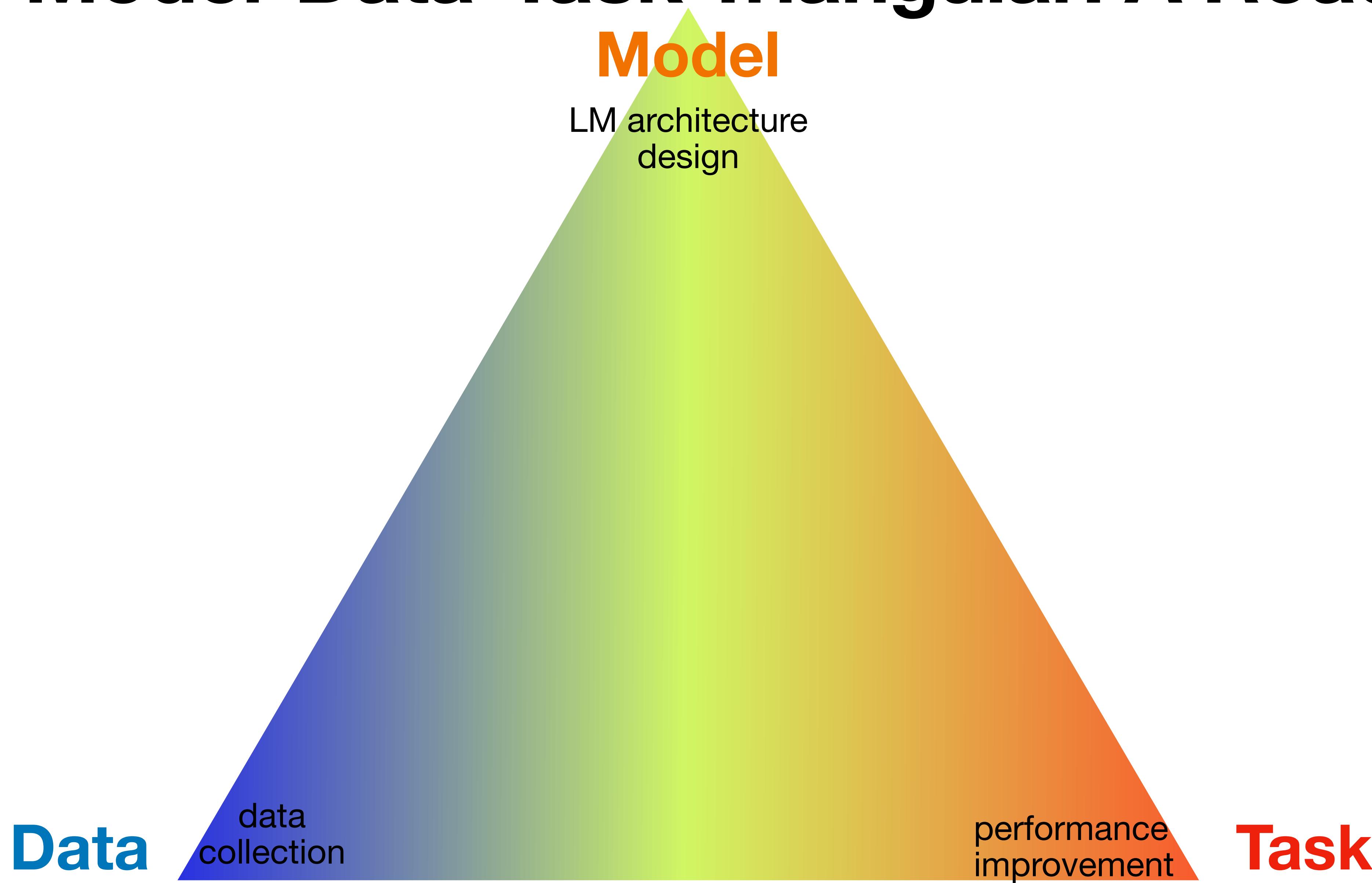
- **Part 3 (Physics): Rules and Laws of LMs**

- **Scaling:** How performance scales
- **Impossibilities:** What LMs cannot do fundamentally

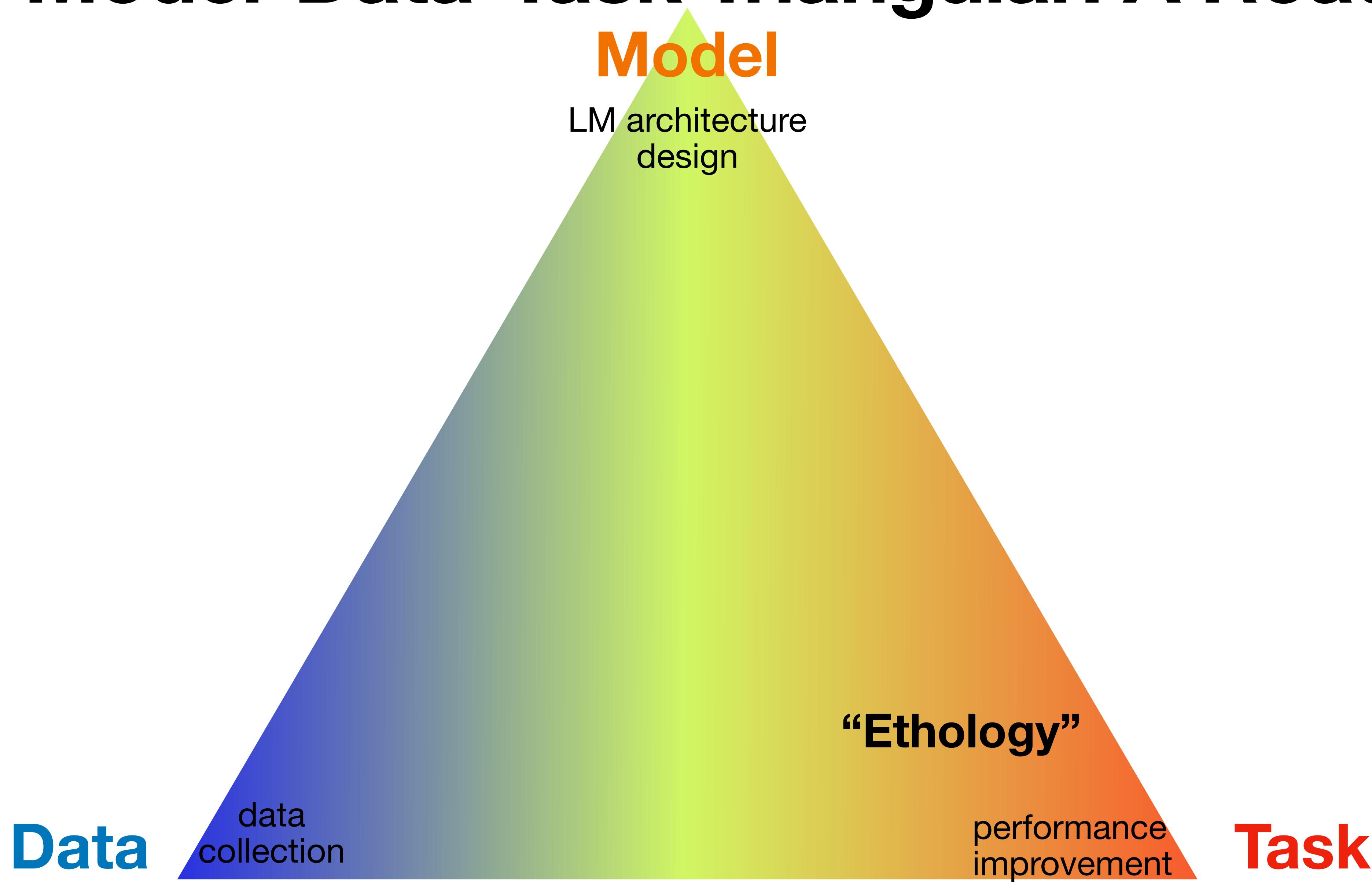
Model-Data-Task Triangular: A Roadmap



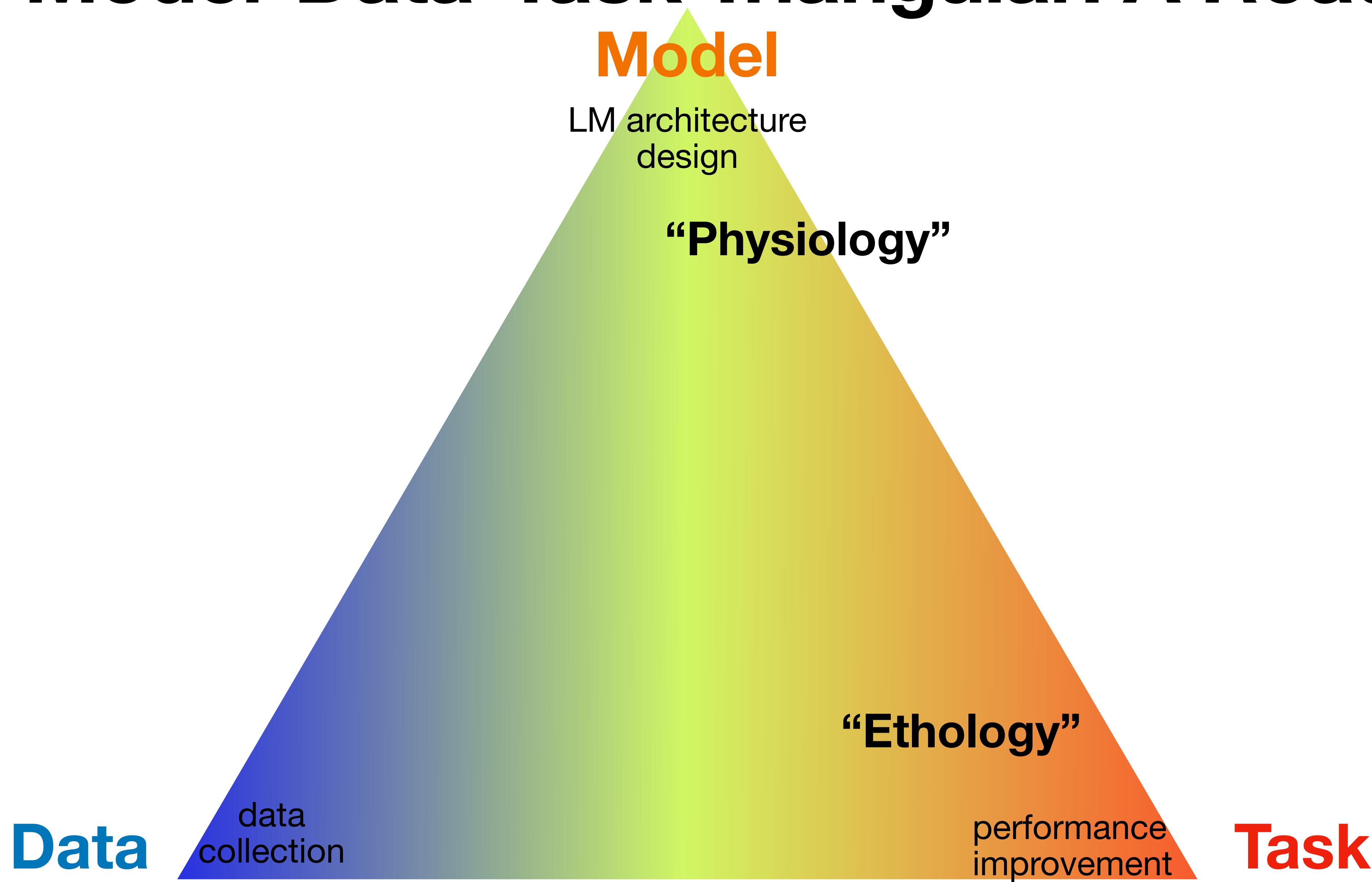
Model-Data-Task Triangular: A Roadmap



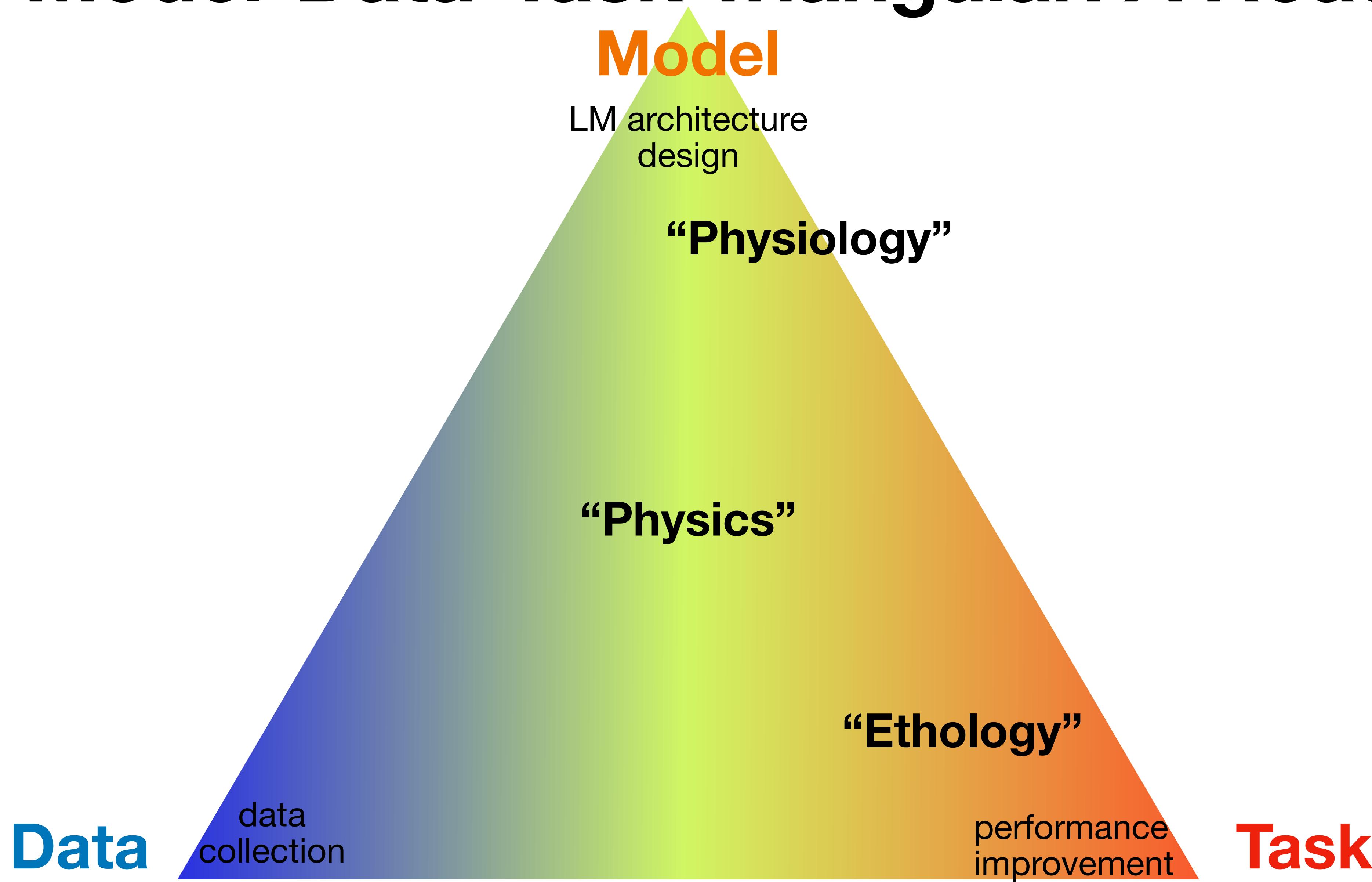
Model-Data-Task Triangular: A Roadmap



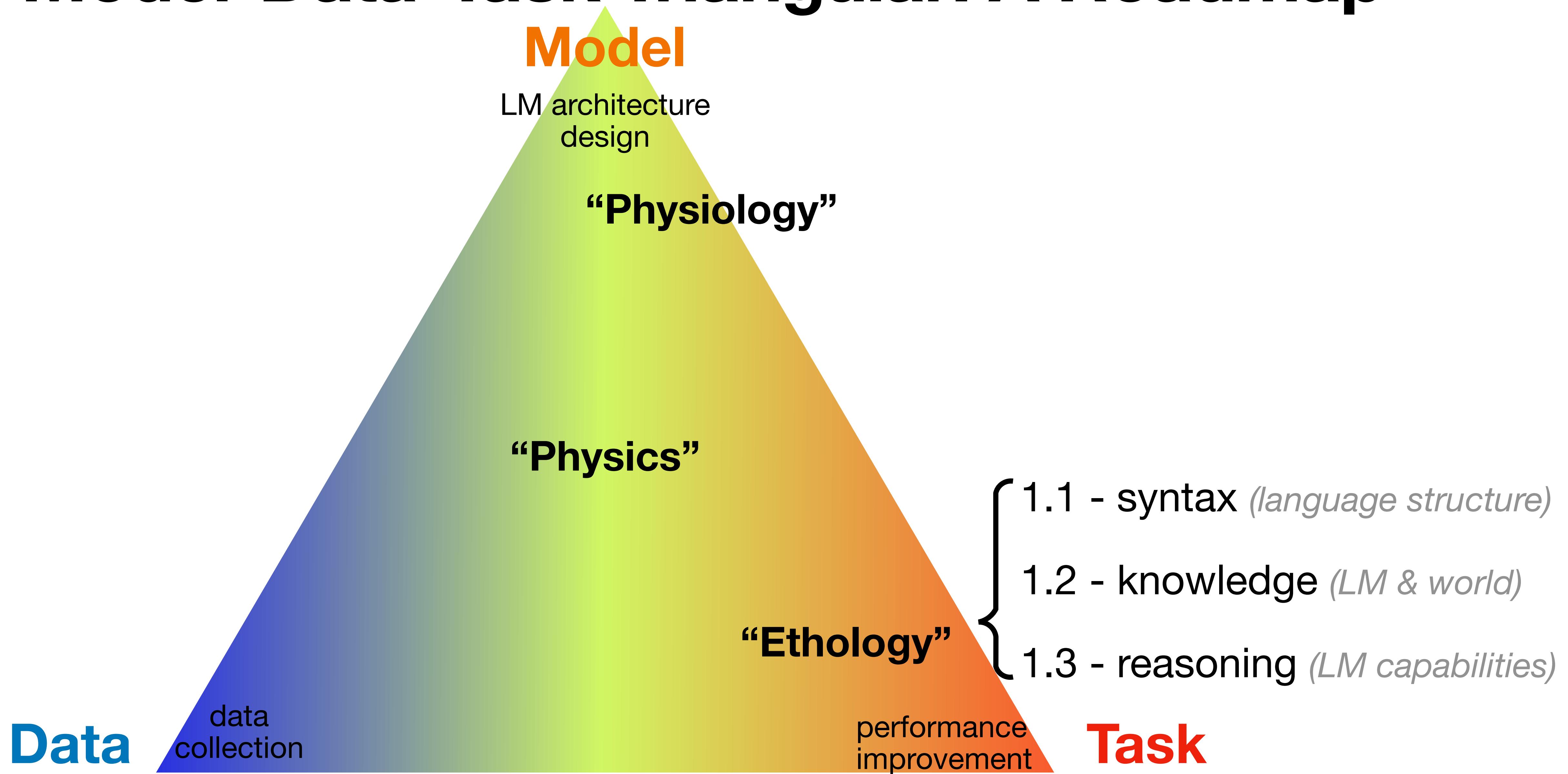
Model-Data-Task Triangular: A Roadmap



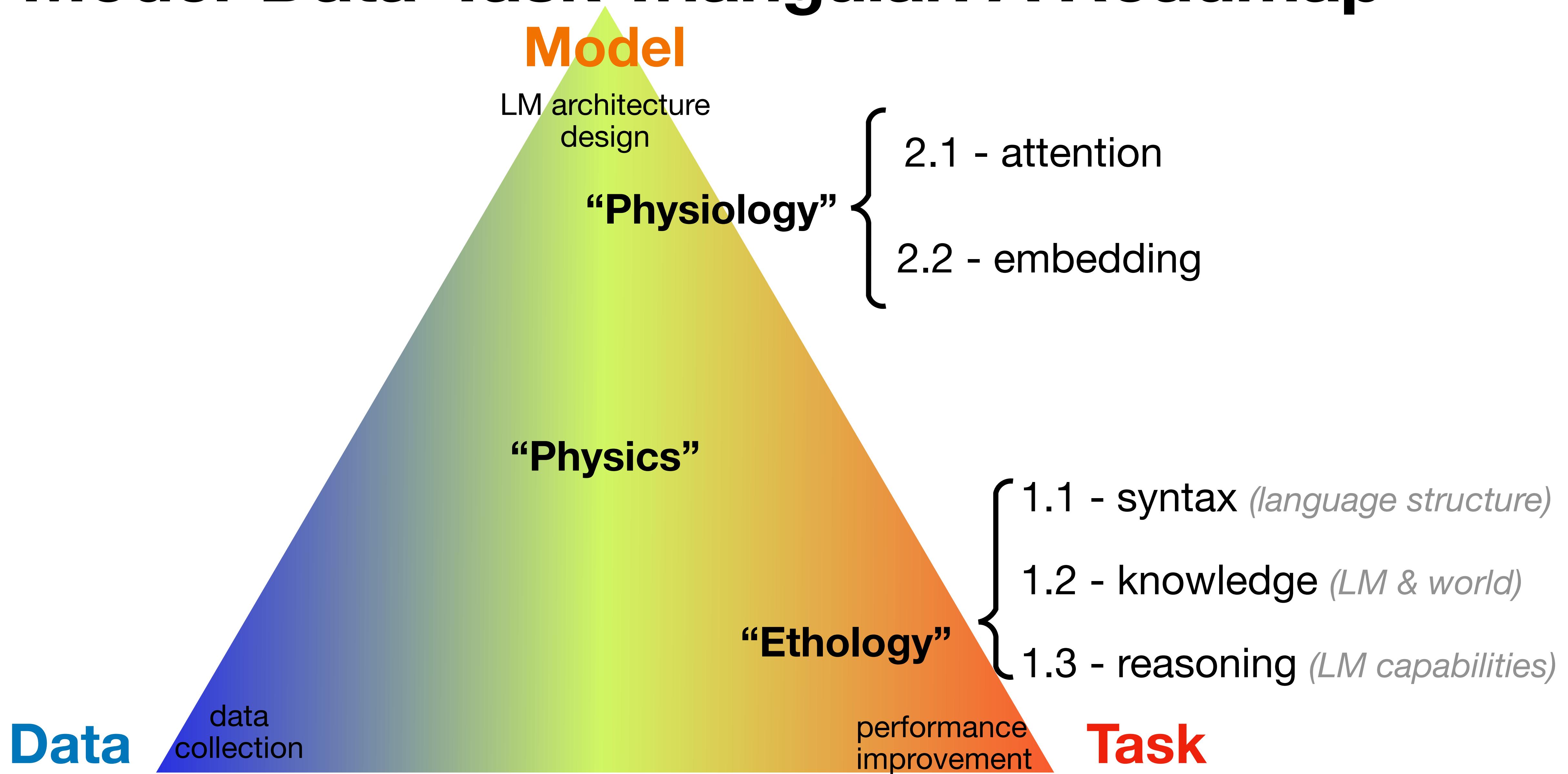
Model-Data-Task Triangular: A Roadmap



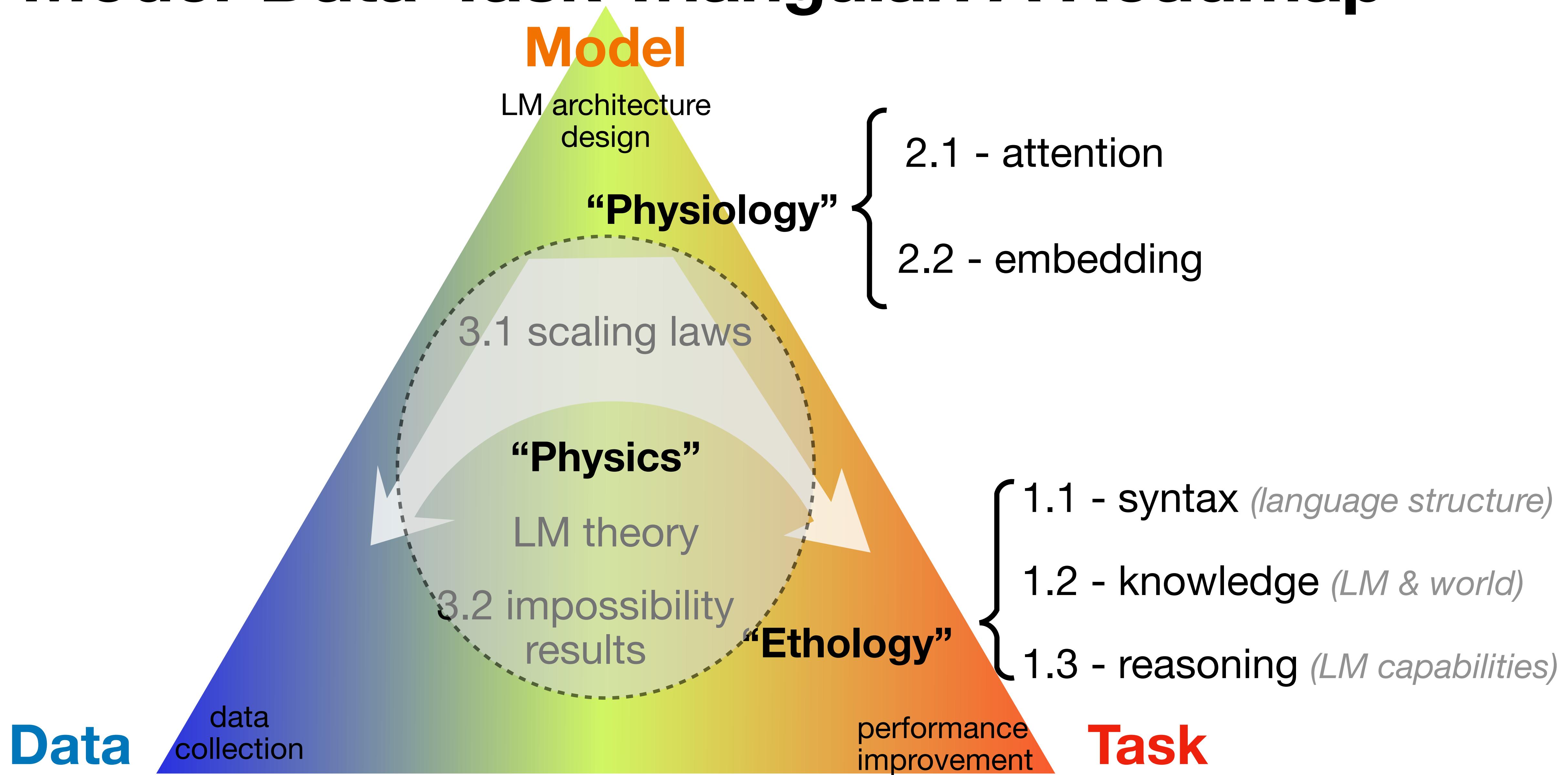
Model-Data-Task Triangular: A Roadmap



Model-Data-Task Triangular: A Roadmap

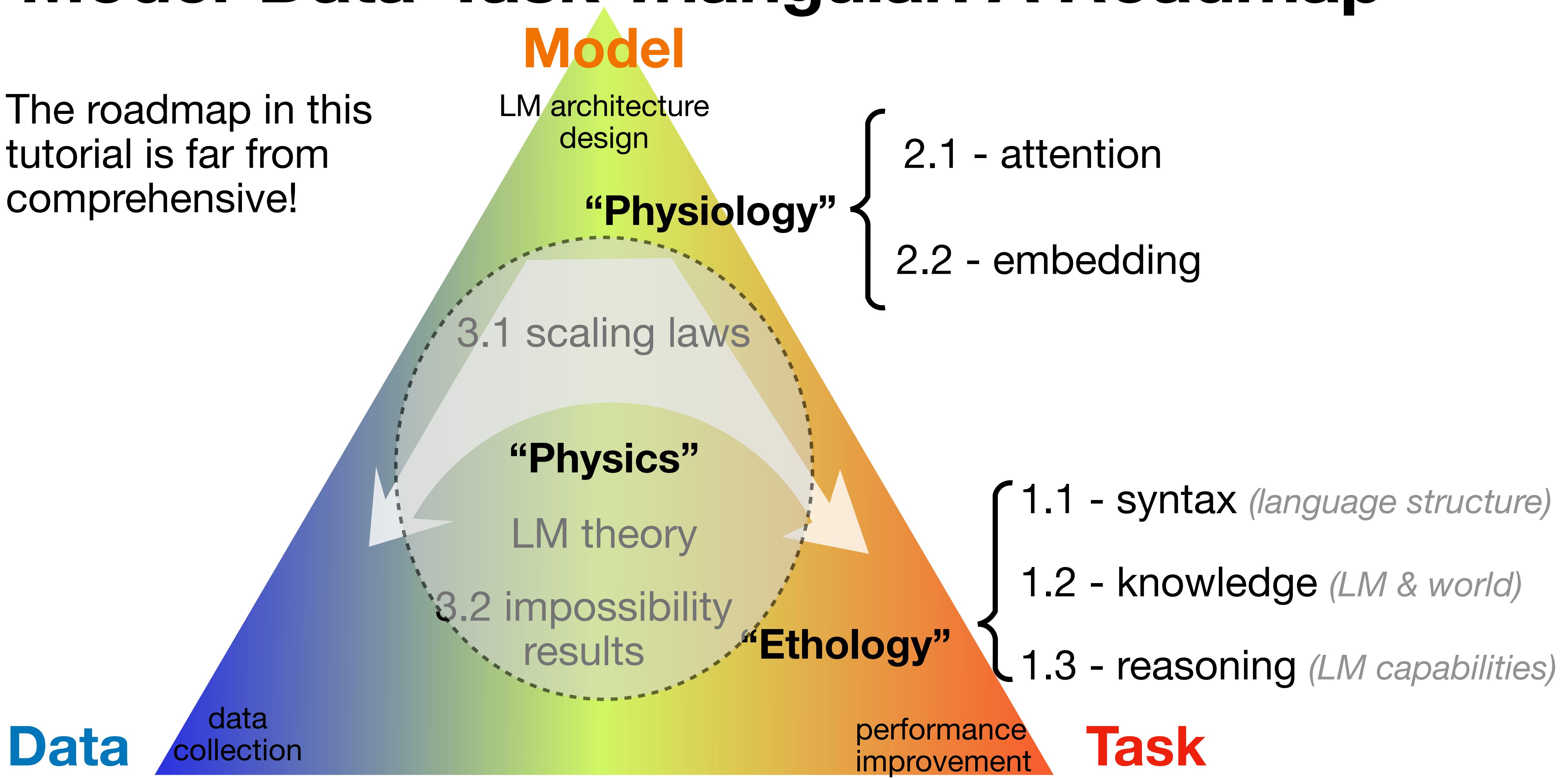


Model-Data-Task Triangular: A Roadmap

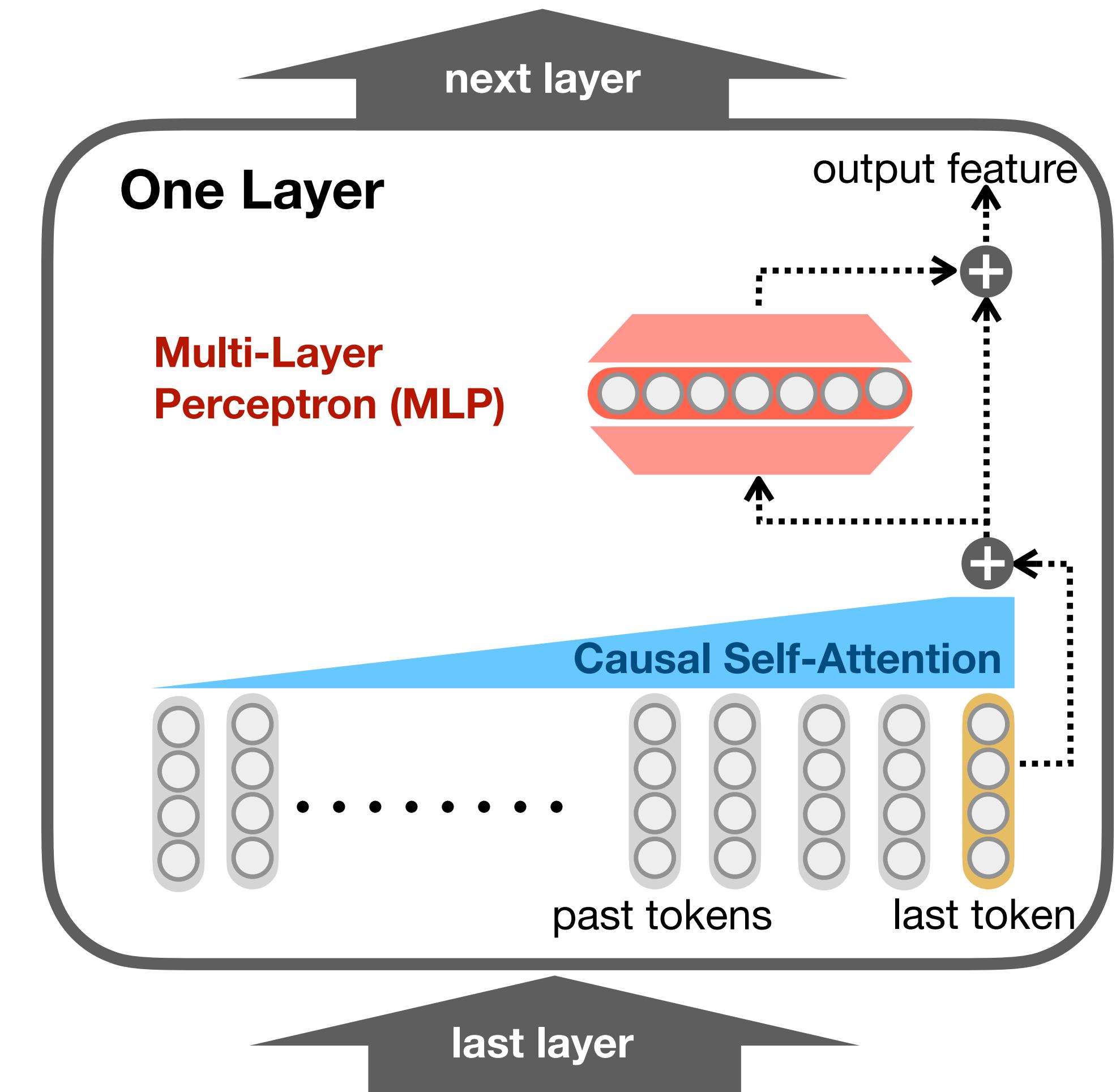
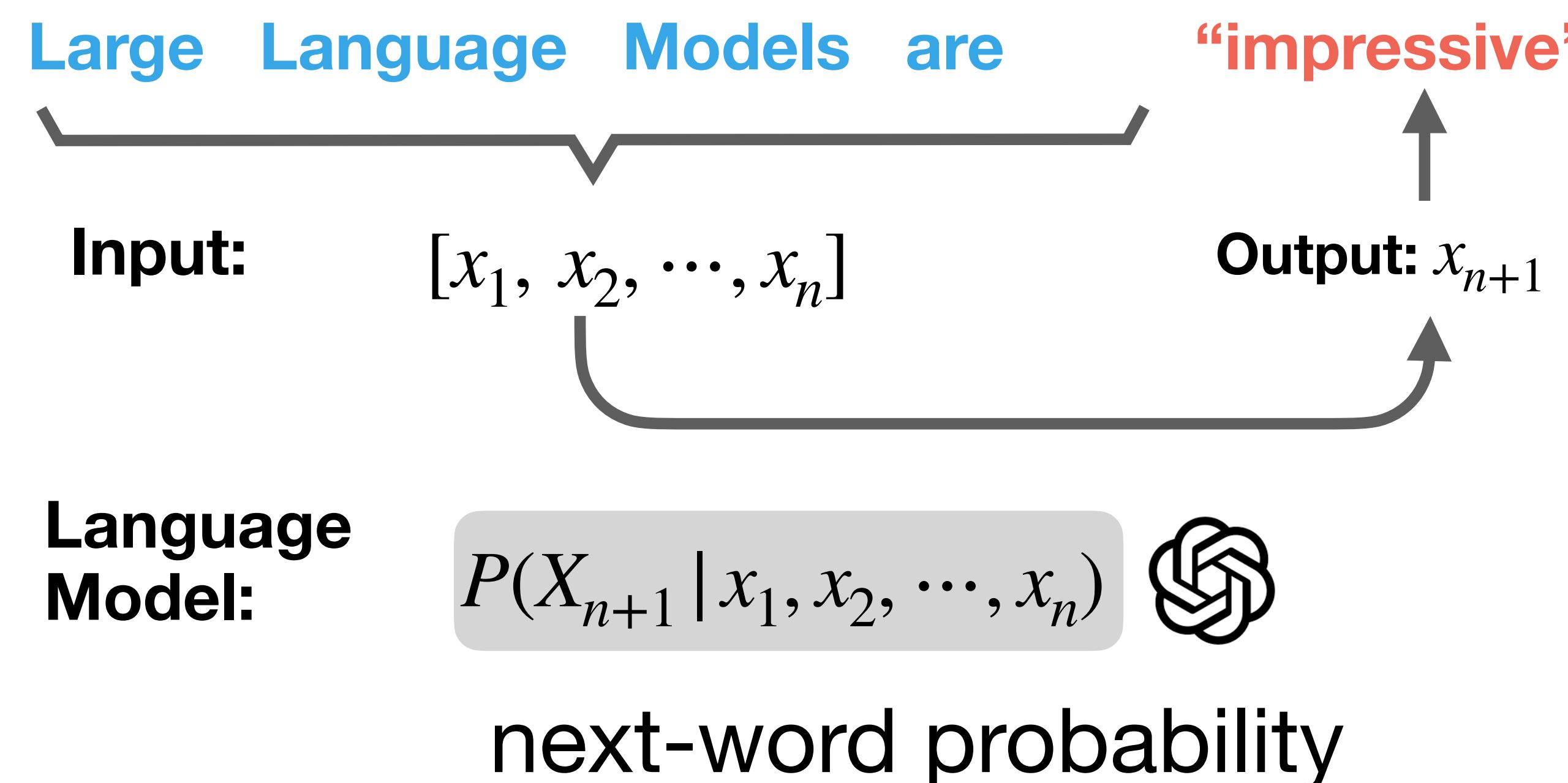


Model-Data-Task Triangular: A Roadmap

The roadmap in this tutorial is far from comprehensive!



Prerequisites: Language Modeling



Language Modeling

A Transformer-Based Architecture

Part 1: Ethology

How Do LMs Behave?

Topics:

- **Syntax:** How do LMs work with syntax
- **Knowledge:** Where is knowledge stored
- **Reasoning:** How is reasoning conducted

How Do LMs Work on Syntax?

LMs Are Robust to “Unnatural Language”

Task: natural language inference

determining if “premise” sentence can infer the “hypothesis” sentence

normally ordered text

texts with shuffled words

	Premise	Hypothesis	Predicted Label
	Boats in daily use lie within feet of the fashionable bars and restaurants.	There are boats close to bars and restaurants.	E
	restaurants and use feet of fashionable lie the in Boats within bars daily .	bars restaurants are There and to close boats .	E
	He and his associates weren't operating at the level of metaphor.	He and his associates were operating at the level of the metaphor.	C
	his at and metaphor the of were He operating associates n't level .	his the and metaphor level the were He at associates operating of .	C

robust answer

LMs Are Robust to “Unnatural Language”

Task: paraphrase
if two sentences are duplicate

Q₁ Does marijuana cause cancer?

Q₂ How can smoking marijuana give you lung cancer?

(a) Prediction: “duplicate” 0.96

Q₁ Does marijuana cause cancer?

Q_{2'} you smoking cancer How marijuana lung can give?

(b) Prediction: “duplicate” 0.98

Task: sentiment classification
if the sentiment is positive or negative

S	the film ’s performances are thrilling .	1.00
---	--	------

S ₁	the film thrilling performances are ’s .	1.00
----------------	--	------

S ₂	’s thrilling film are performances the .	1.00
----------------	--	------

S ₃	’s thrilling are the performances film .	1.00
----------------	--	------

Task: entailment
if the sentence A contains the answer to question Q

QNLI sentence-pair inputs and their LIME attributions (negative -1, neutral 0, positive +1)		Confidence score
Q	How long did Phillips manage the Apollo missions?	1.00
A	Mueller agreed, and Phillips managed Apollo from January 1964, until it achieved the first manned landing in July 1969, after which he returned to Air Force duty.	
Q ₁	Apollo the Phillips How missions long did manage?	0.96
A	Mueller agreed, and Phillips managed Apollo from January 1964, until it achieved the first manned landing in July 1969, after which he returned to Air Force duty.	

LMs Are Robust to “Unnatural Language”

- Different capabilities have different sensitivity to syntax corruption
 - in *sentence acceptability*, naturally requires integrate syntax
 - or when the meaning is reversed (A cause B v.s. B cause A)

Q₁ Does marijuana cause cancer?

Q_{2''} lung can give marijuana smoking How you cancer?

(c) Prediction: “duplicate” 0.99

Q₁ Does marijuana cause cancer?

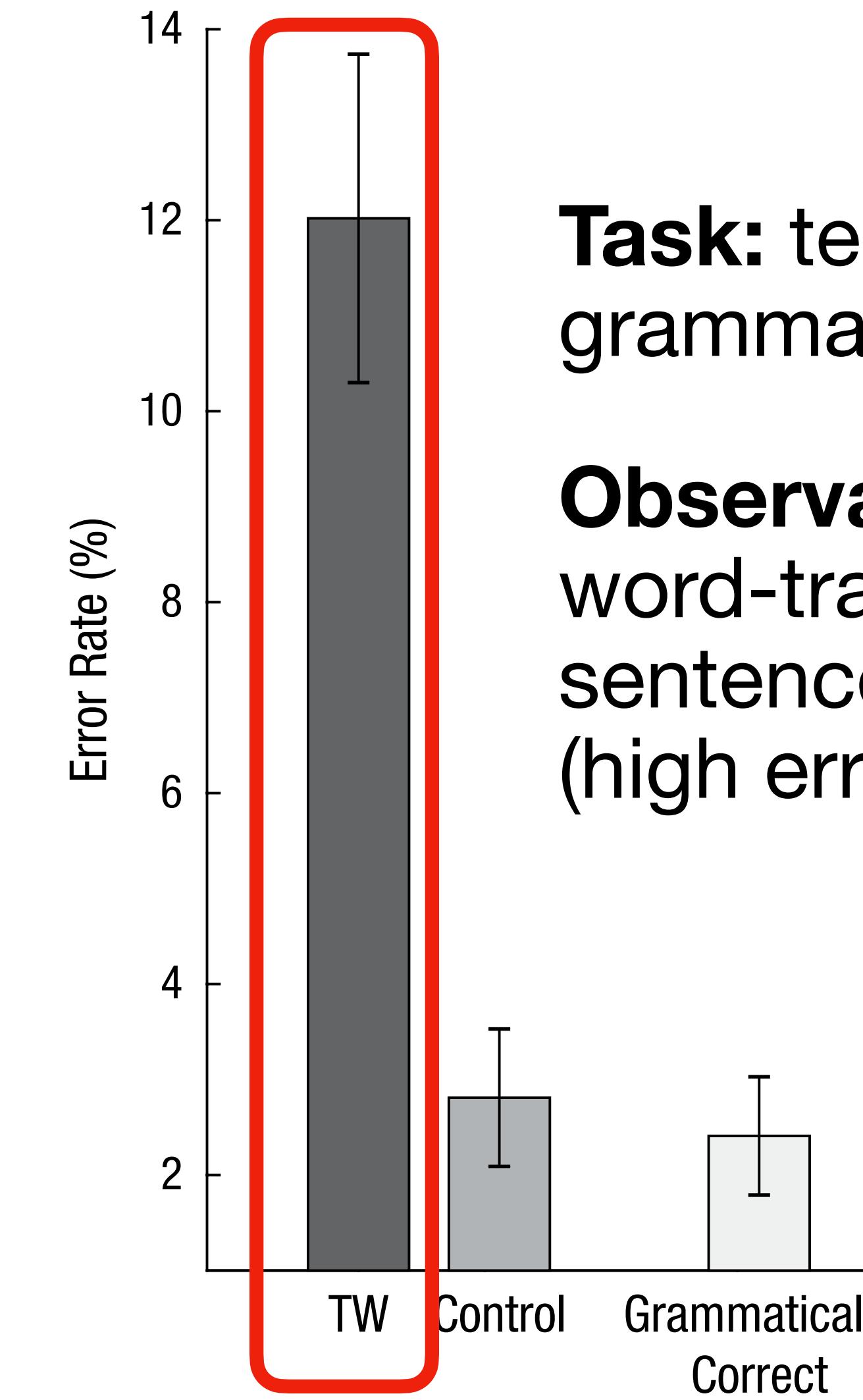
Q_{1'} Does cancer cause marijuana?

Word Transposition Effect in Humans

Task: is the new sentence grammatically correct?

The white cat was big.
The black dog ran slowly.

The white was cat big.
The black ran dog slowly.



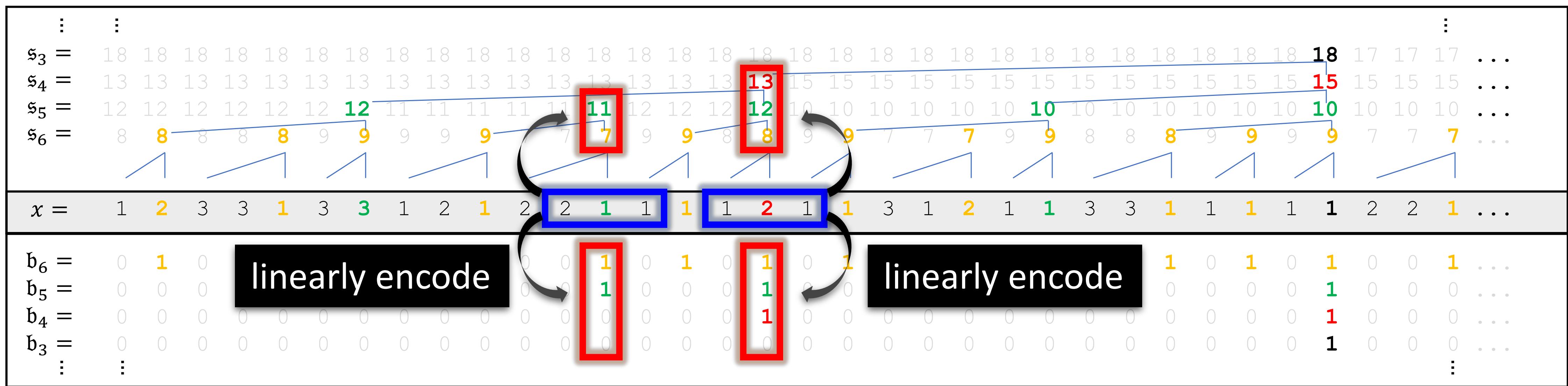
Task: tell if the sentence is grammatical or not

Observation: if the sentence is word-transposed from original sentence, it is less recognizable (high error)

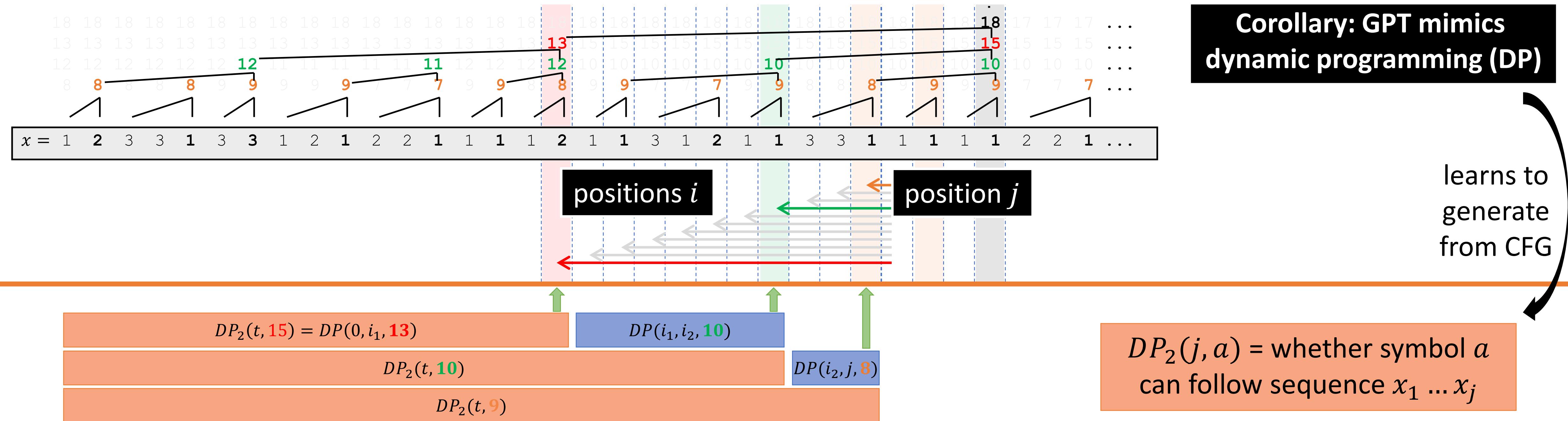
Part 1: Ethology - Topic 1: Syntax

Hidden Features Encode Local Syntax

Probing: predicting each word's syntax within hidden features



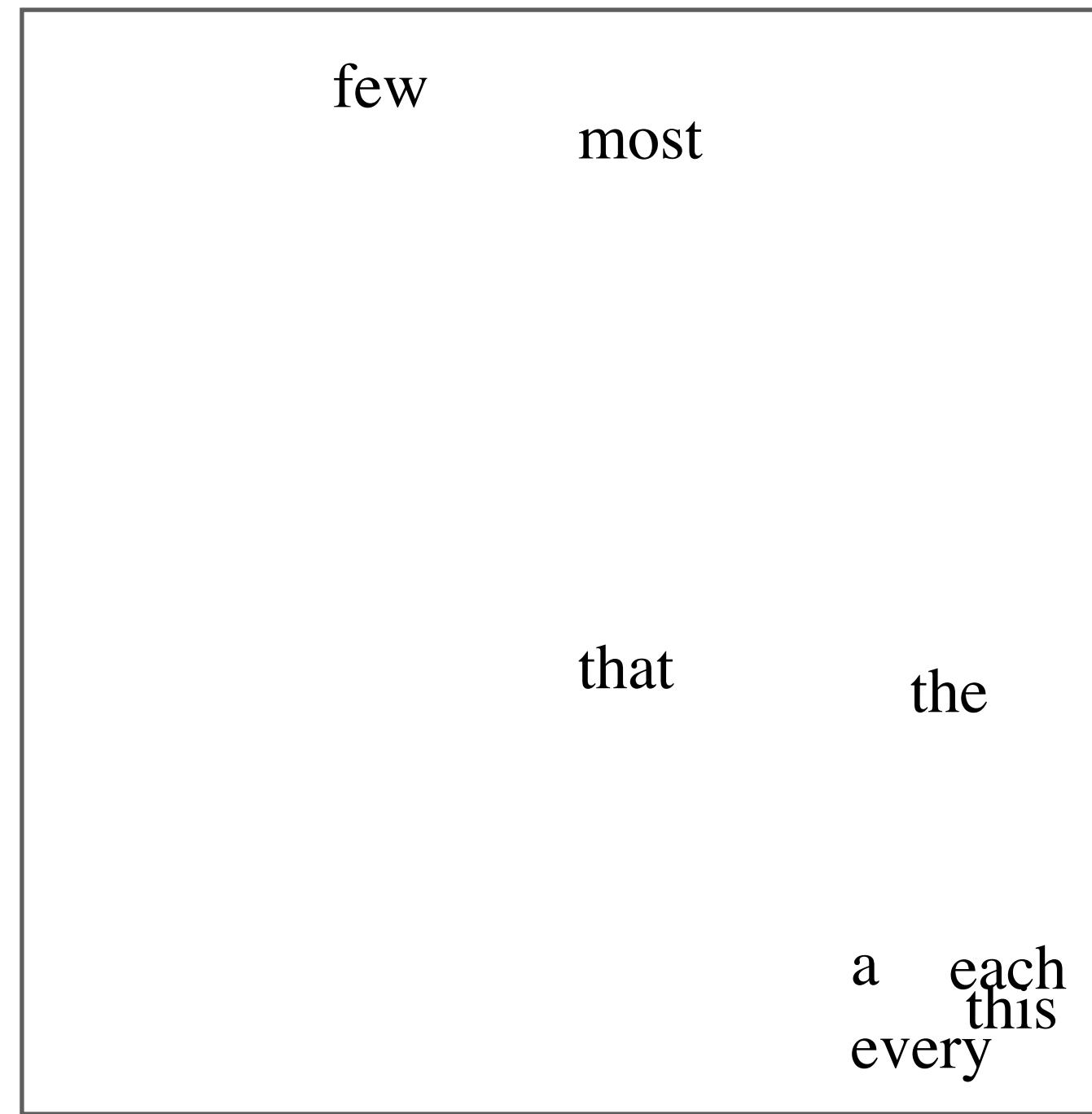
Hidden Features Encode Syntax-Parsing Features



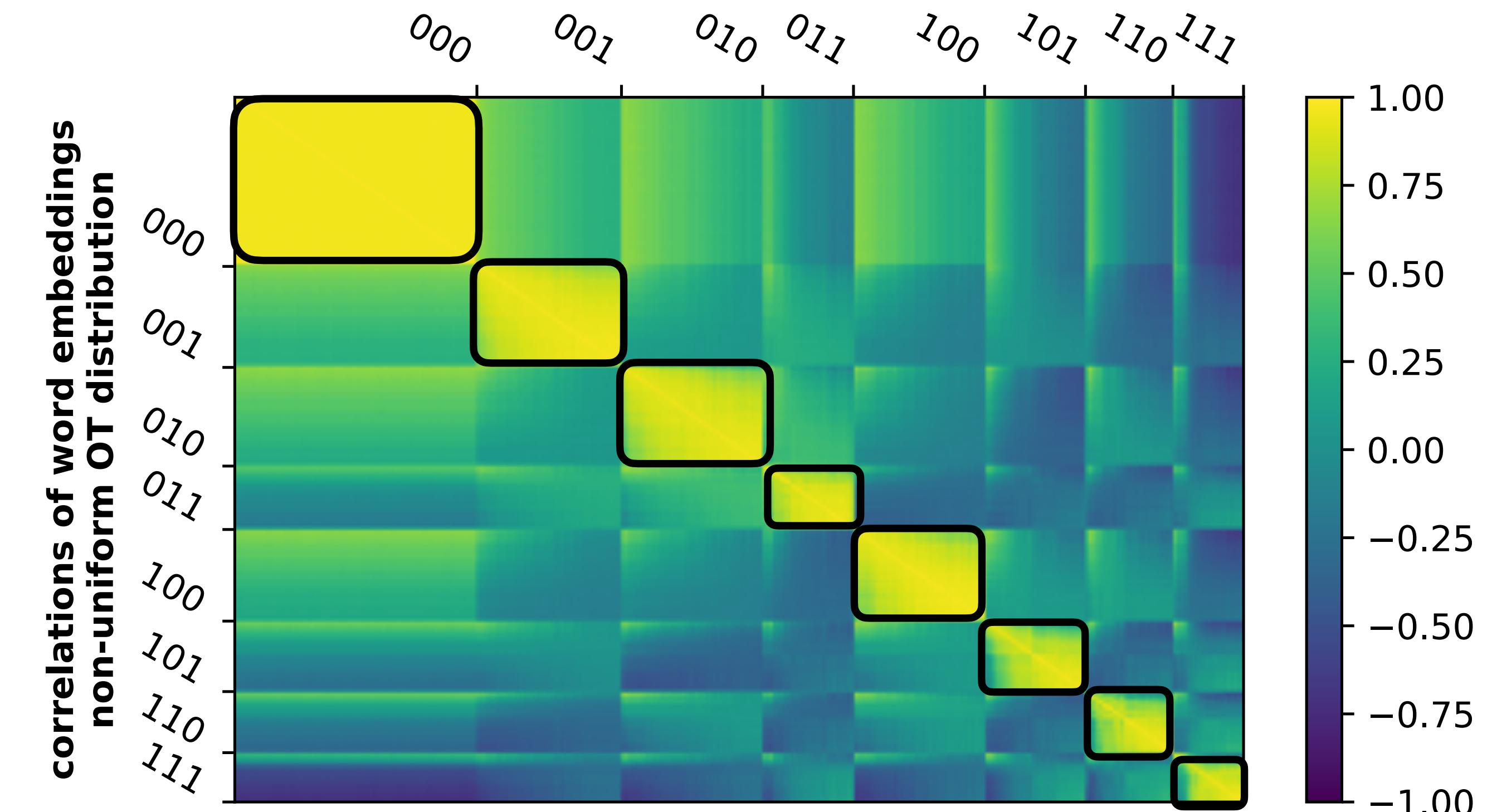
Probing: if hidden features can recover computational features useful for parsing syntax

Word Embeddings Encode Syntactic Roles

In natural language, word embeddings reflect their syntax roles

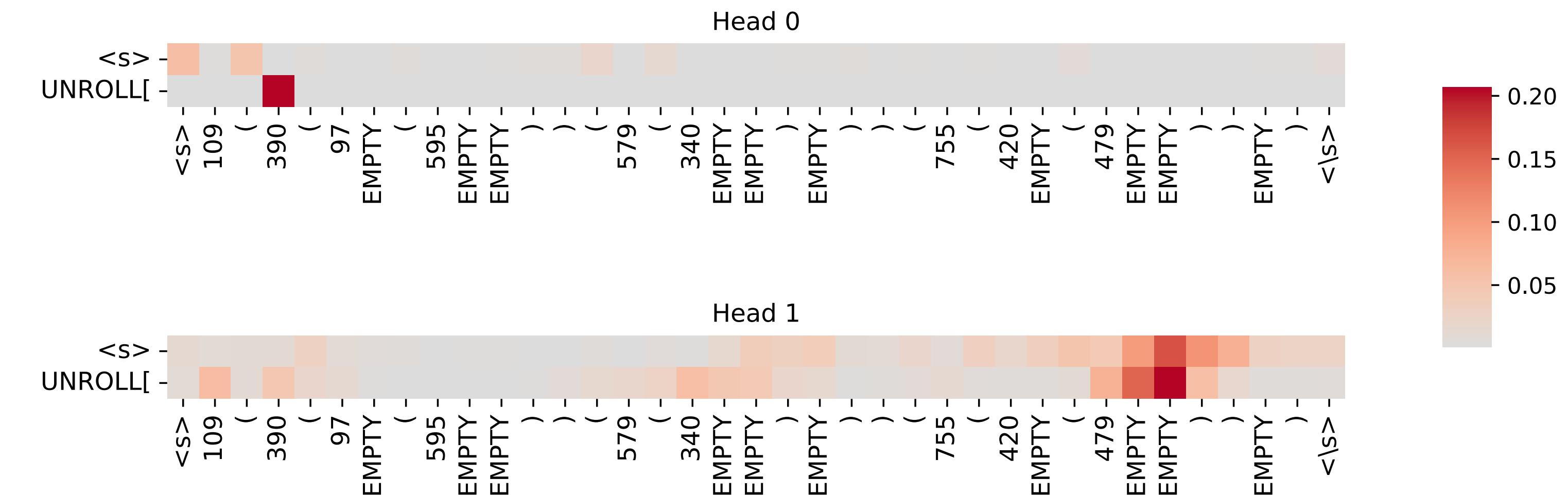


In a synthetic language, word embeddings are grouped by syntactic roles



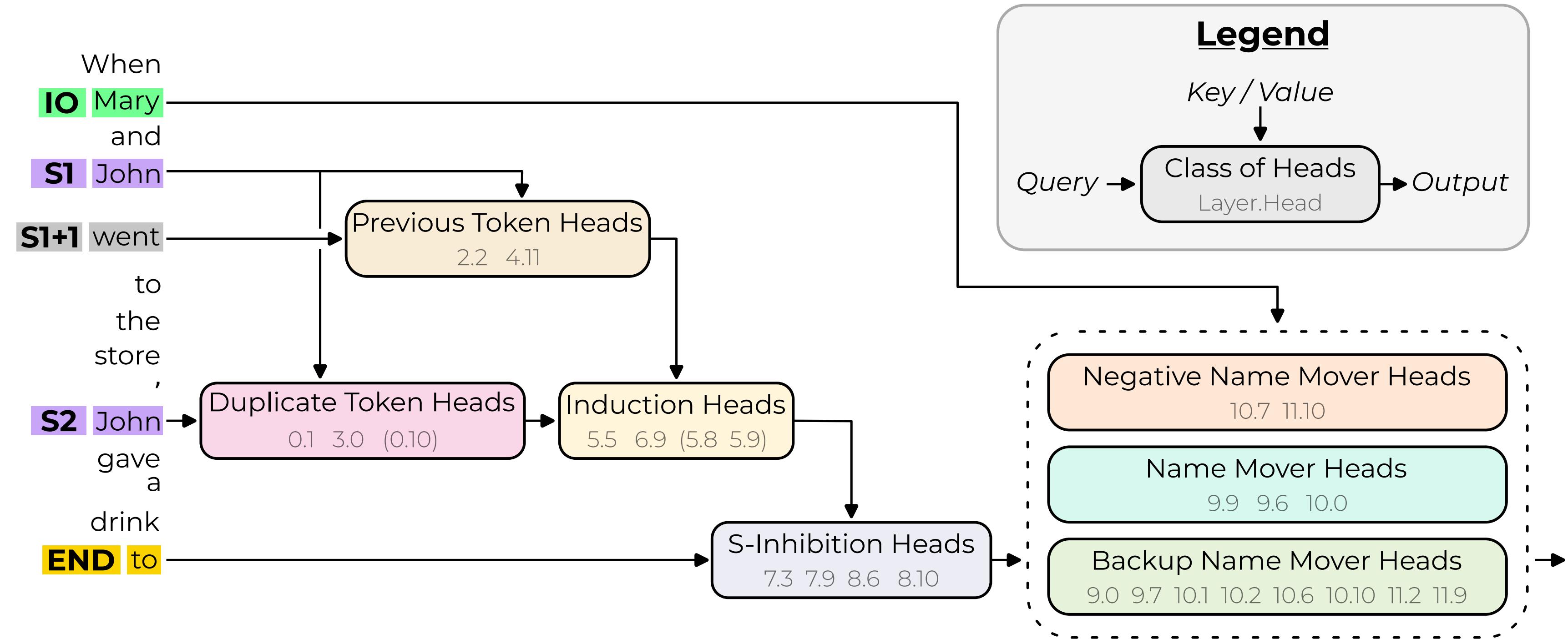
LMs Uses Attention to Utilize Syntax

Attention heads that close parentheses



LMs Uses Attention to Utilize Syntax

Heads that copy-past nouns



“(Backup) Name Mover Heads”: copy-pasting nouns

“Duplicate Token Heads” and “Induction Heads”: detecting duplicate nouns

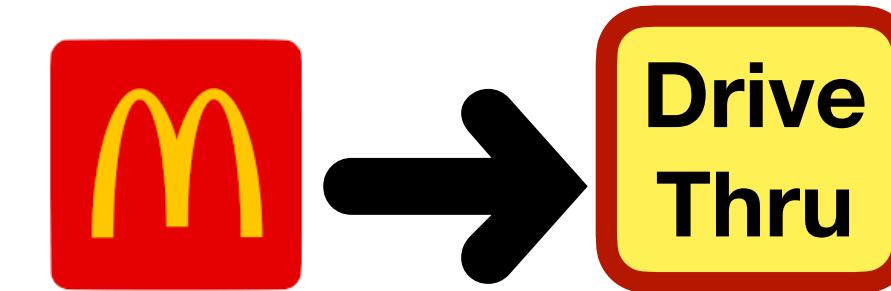
“S-Inhibition Heads”: suppressing attention on duplicate nouns

Where Is Knowledge Stored

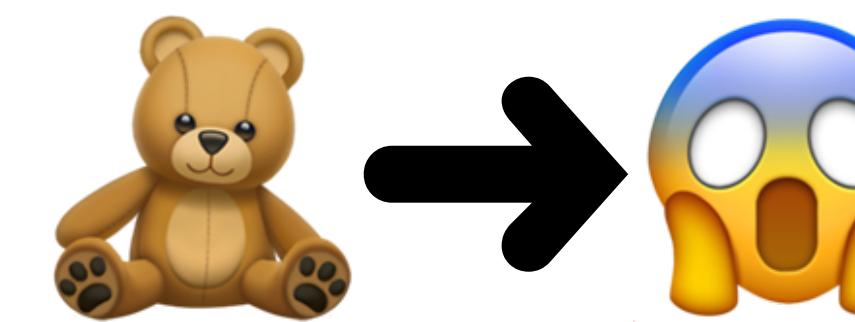
Memory Storage in Human Brain

We Know Some Functions of Brain Regions

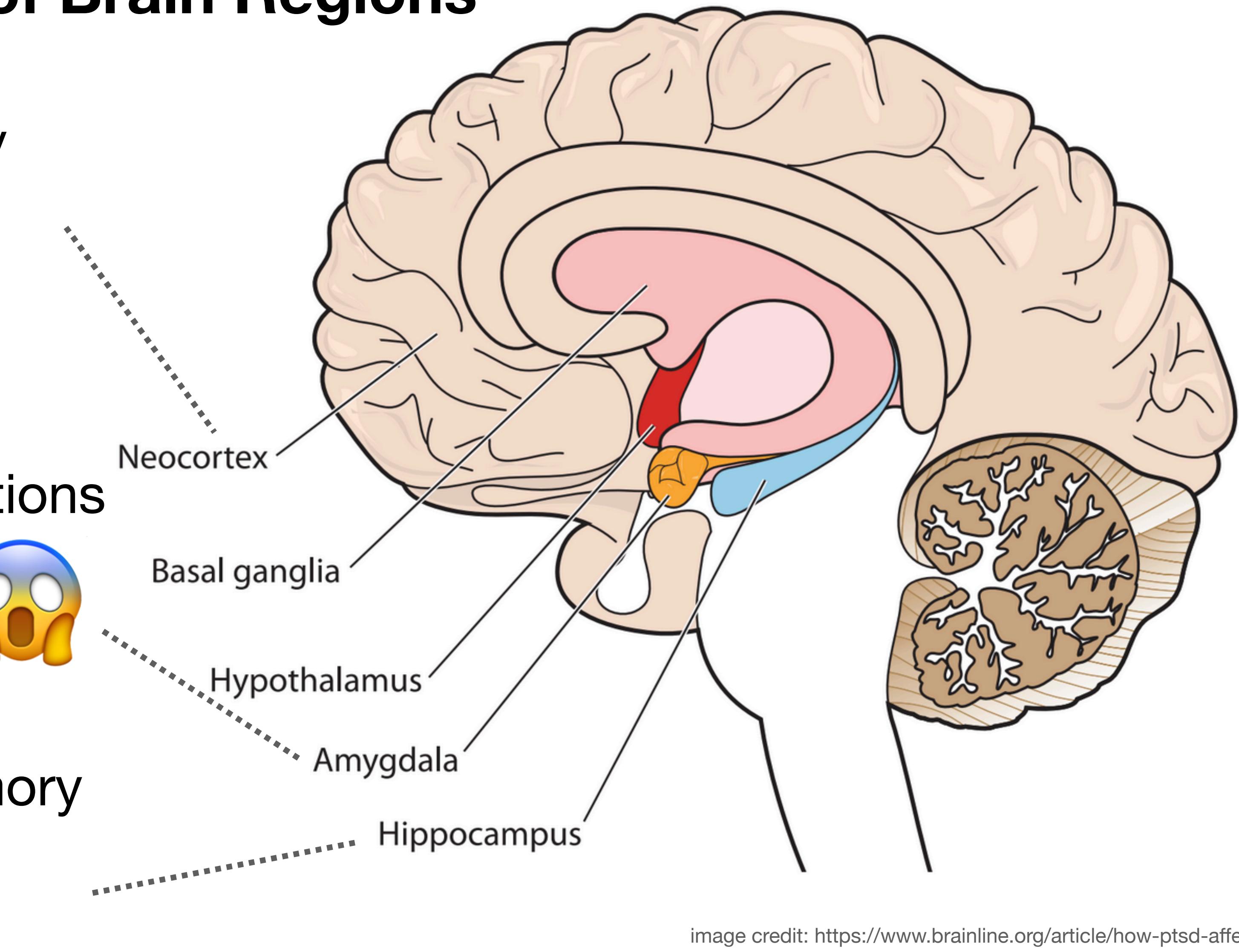
Neocortex: semantic memory transfer (e.g., facts, commonsense knowledge)



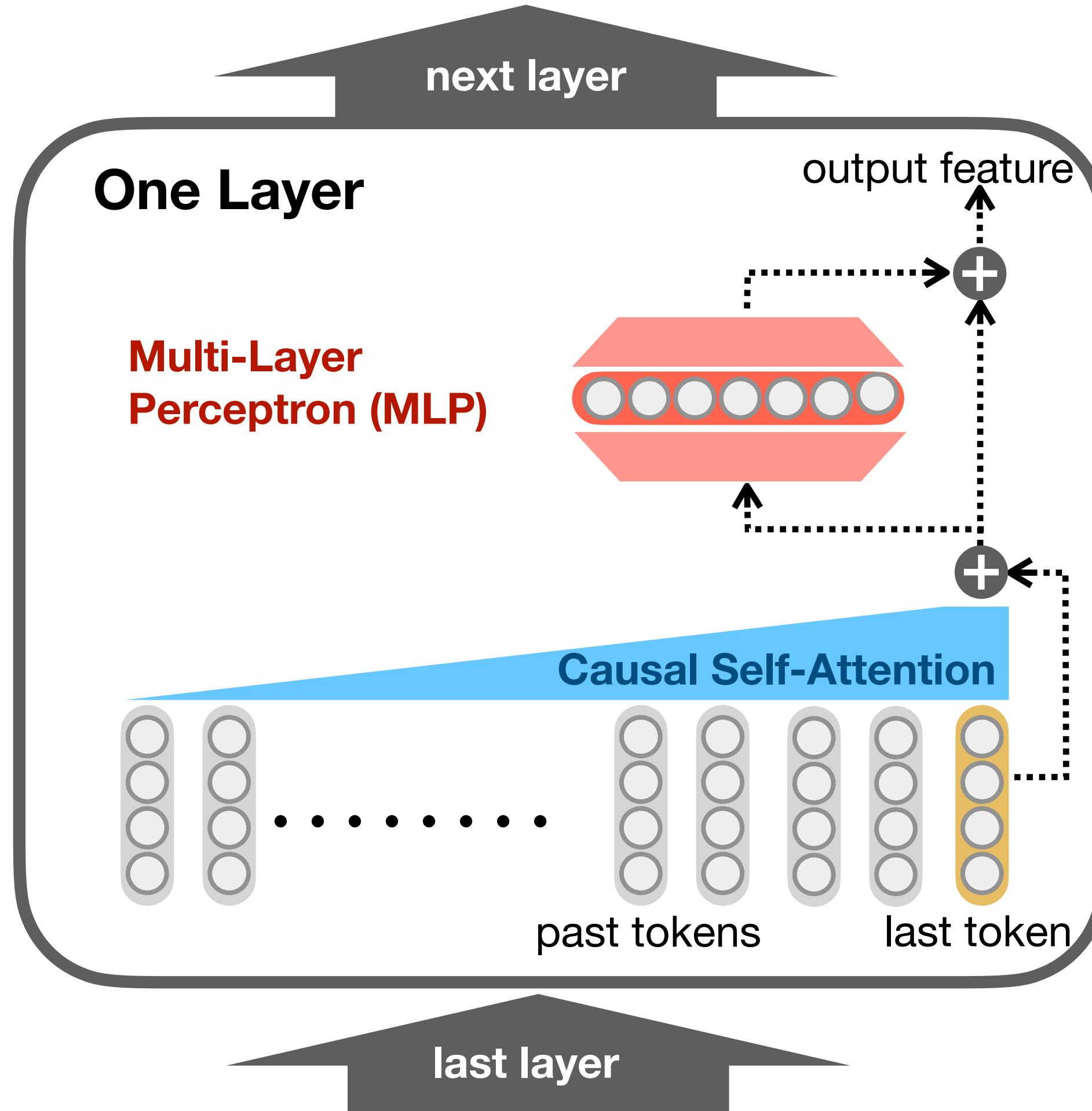
Amygdala: emotional implications (e.g., fear, PTSD)



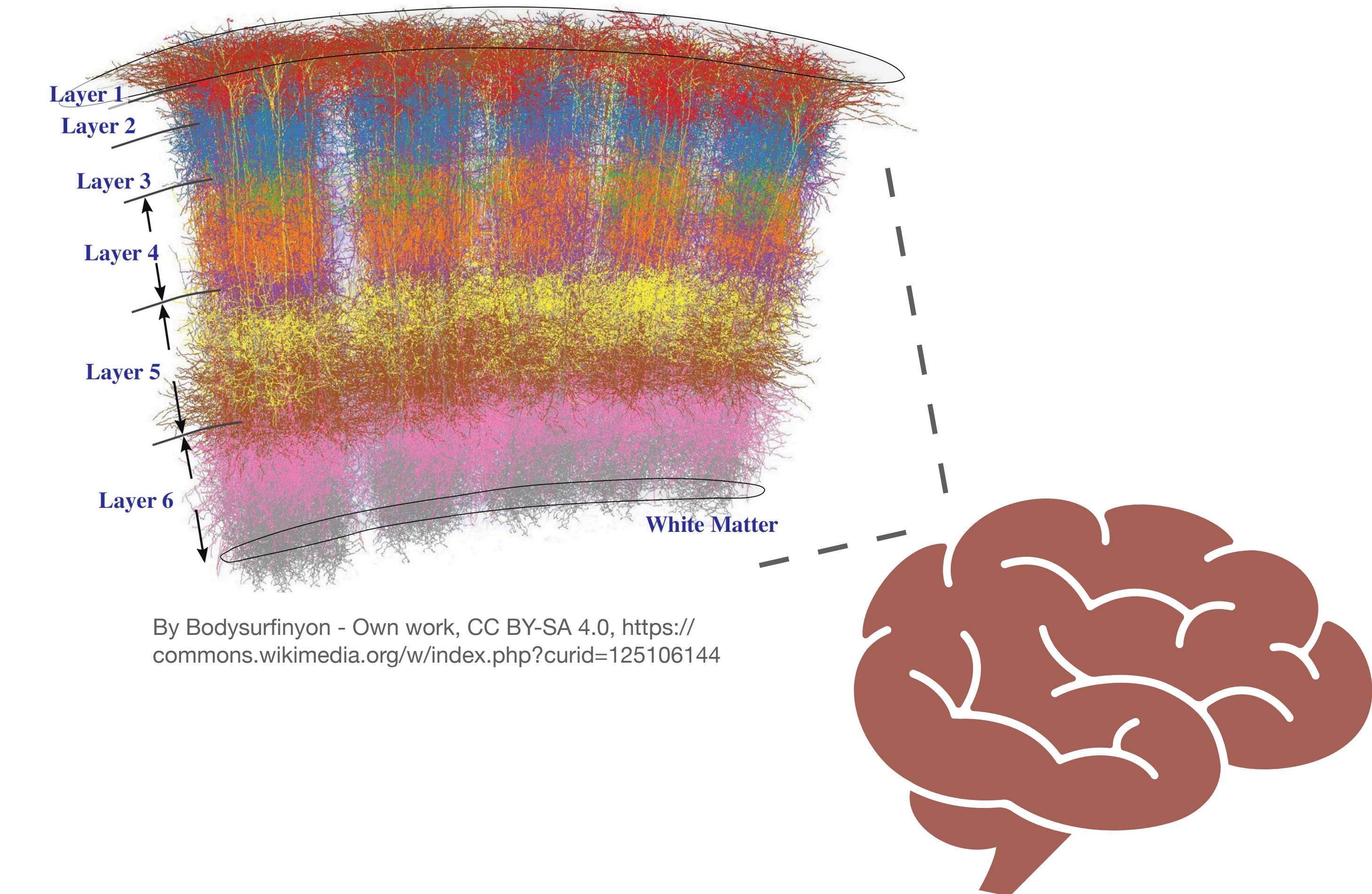
Hippocampus: episodic memory (e.g., events)



Can We Analogize LMs with Brains? and therefore, analogize LM parameters \approx brain neurons?

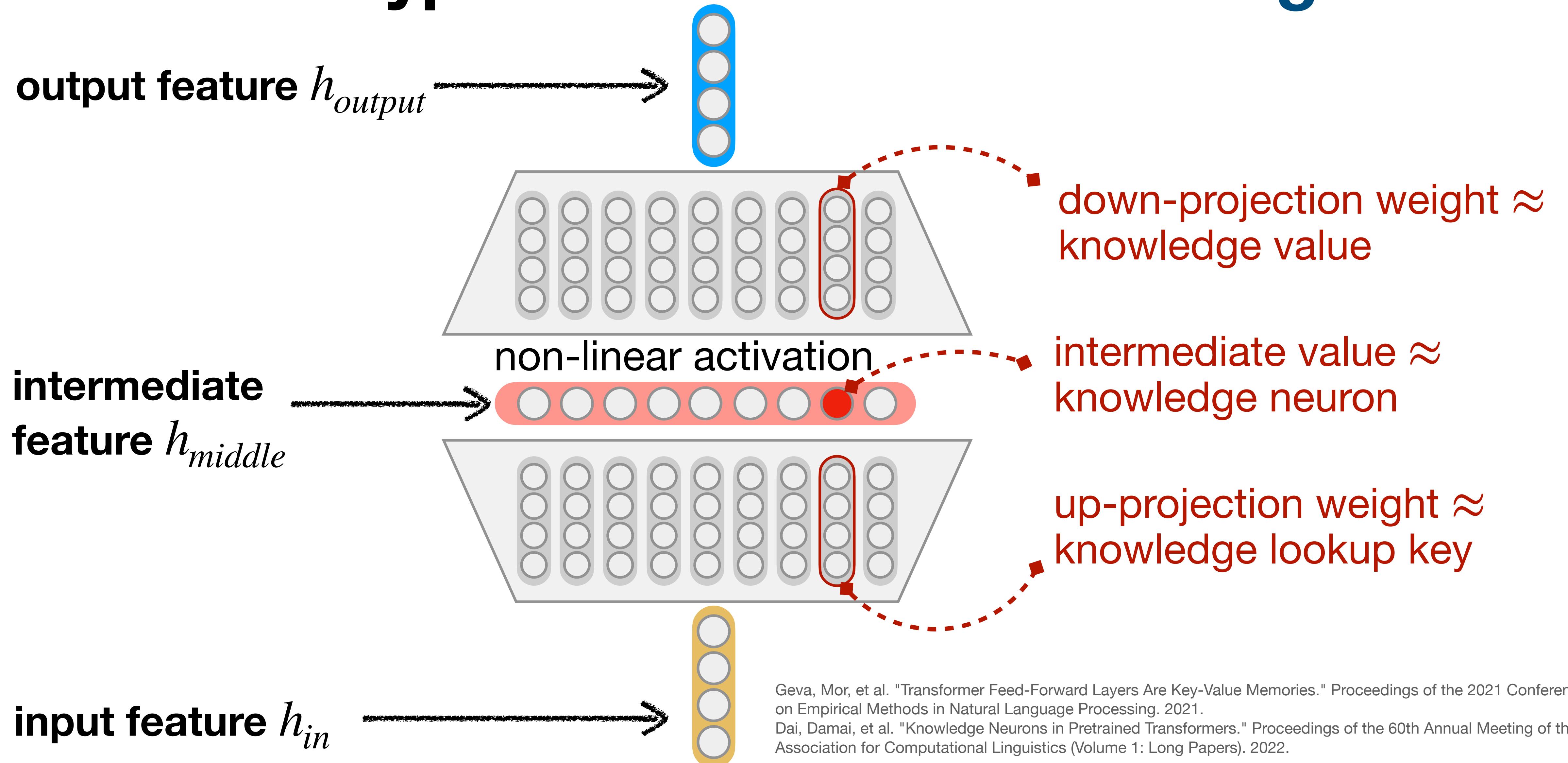


One Transformer LM

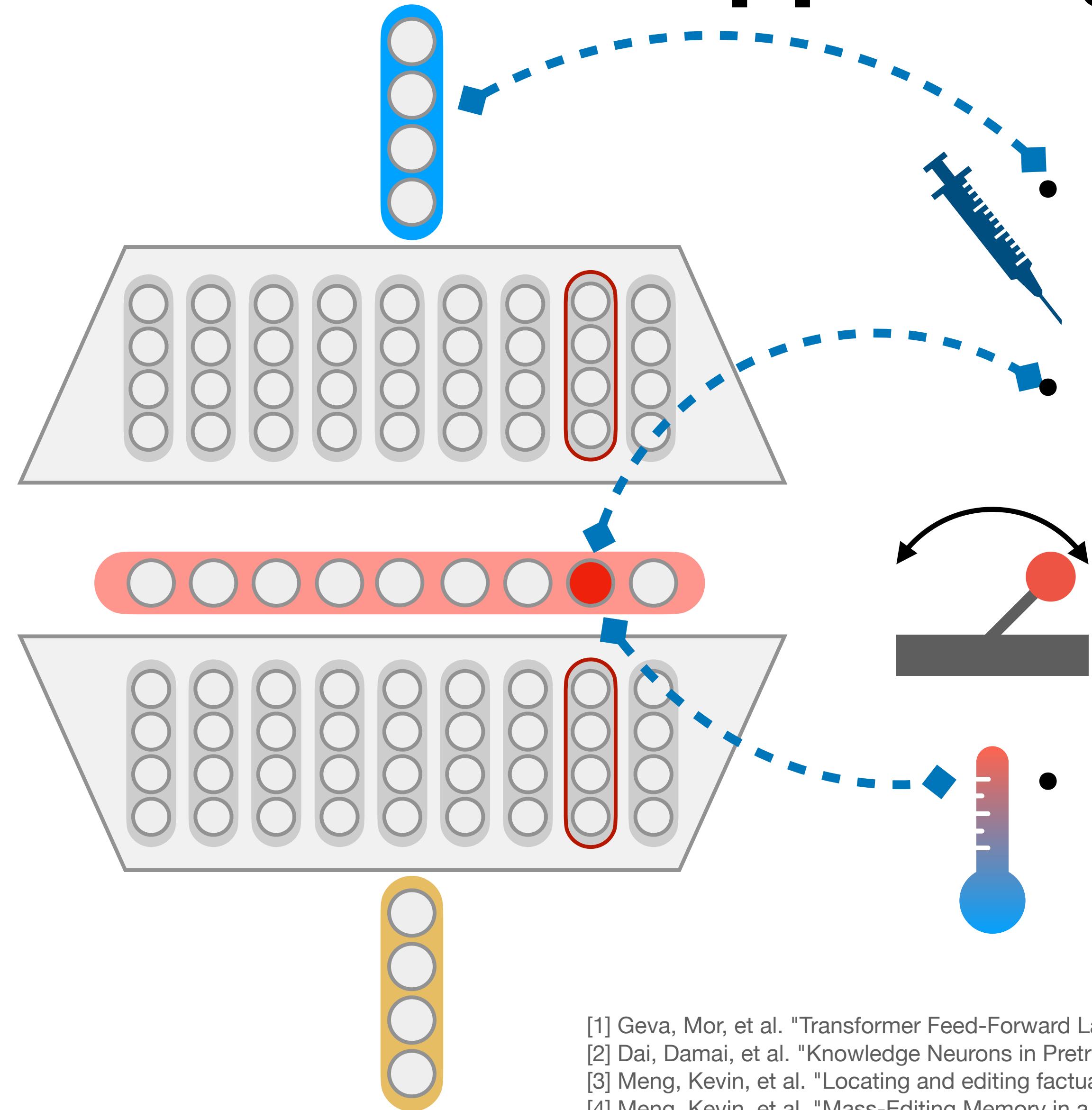


Brain and its 6-layers in neocortex

Common Hypothesis: $\text{MLP} \approx \text{Knowledge dict}$



Evidence Supporting $\text{MLP} \approx \text{Knowledge dict}$



- **Inserting output features** can *inject certain knowledge prediction*
- **Manually activate neurons** can also *force certain knowledge prediction*
 - e.g. Dublin is the capital and largest city of England → Ireland
- **Certain neurons react to knowledge types**
 - e.g., “part-of” types, related to TV shows

[1] Geva, Mor, et al. "Transformer Feed-Forward Layers Are Key-Value Memories." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

[2] Dai, Damai, et al. "Knowledge Neurons in Pretrained Transformers." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022.

[3] Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

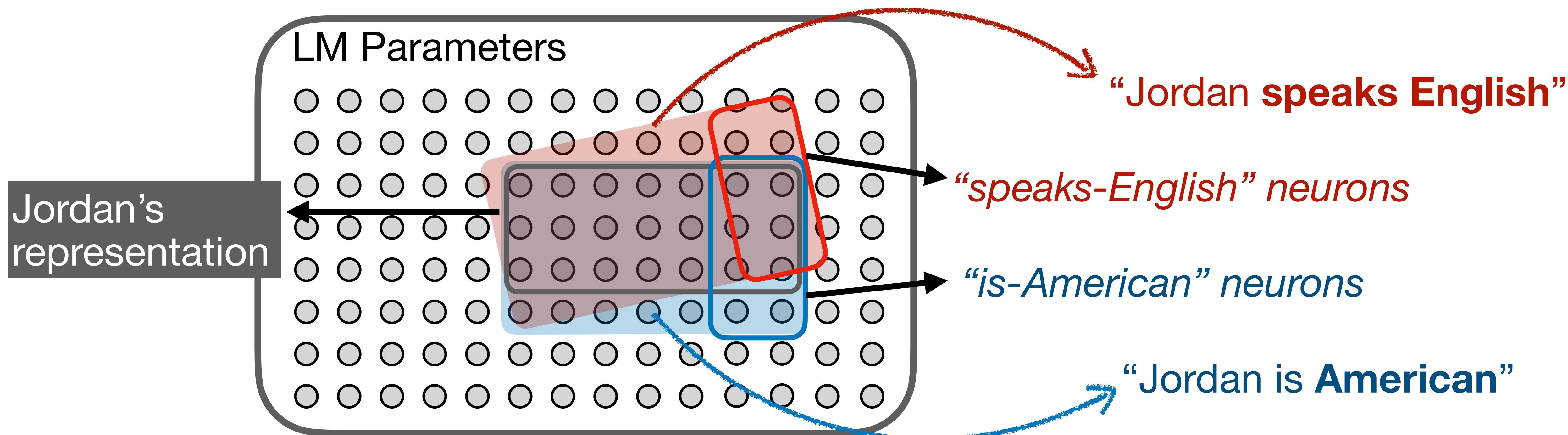
[4] Meng, Kevin, et al. "Mass-Editing Memory in a Transformer." The Eleventh International Conference on Learning Representations.

Knowledge Seems to Be Stored Messily

Expectation: semantic / logical related facts should share parameters

Why? intuitively, this leads to better semantic/logic-based generalization:

if $x \rightarrow y$, when $P(x) \uparrow$, the LM can automatically $P(y) \uparrow$



Knowledge Seems to Be Stored Messily

incompatible sentences positively align parameters

Knowledge Seems to Be Stored Messily

incompatible sentences positively align parameters

- **Negation curse:**
 - X: Leonardo is from USA
 - not X: Leonardo is not from USA

Knowledge Seems to Be Stored Messily

incompatible sentences positively align parameters

- **Negation curse:**
 - X: Leonardo is from USA
 - not X: Leonardo is not from USA
- **Over-Ripple:**
 - X: Leonardo is from USA
 - χ : Leonardo speaks USA

Knowledge Seems to Be Stored Messily

similar sentences with low parameter overlap

Knowledge Seems to Be Stored Messily

similar sentences with low parameter overlap

- **Cross-Lingual Barrier:**
 - Leonardo is from USA
 - 莱昂纳多来自美国 (same meaning)

Knowledge Seems to Be Stored Messily

similar sentences with low parameter overlap

- **Cross-Lingual Barrier:**
 - Leonardo is from USA
 - 莱昂纳多来自美国 (same meaning)
- **Logical Distance Barrier:**
 - Leonardo is from USA
 - The highest building in the capital of Leonardo's homeland is Washington Monument (3 logical steps from above)

Curse of Reversals / Inverse-Searches

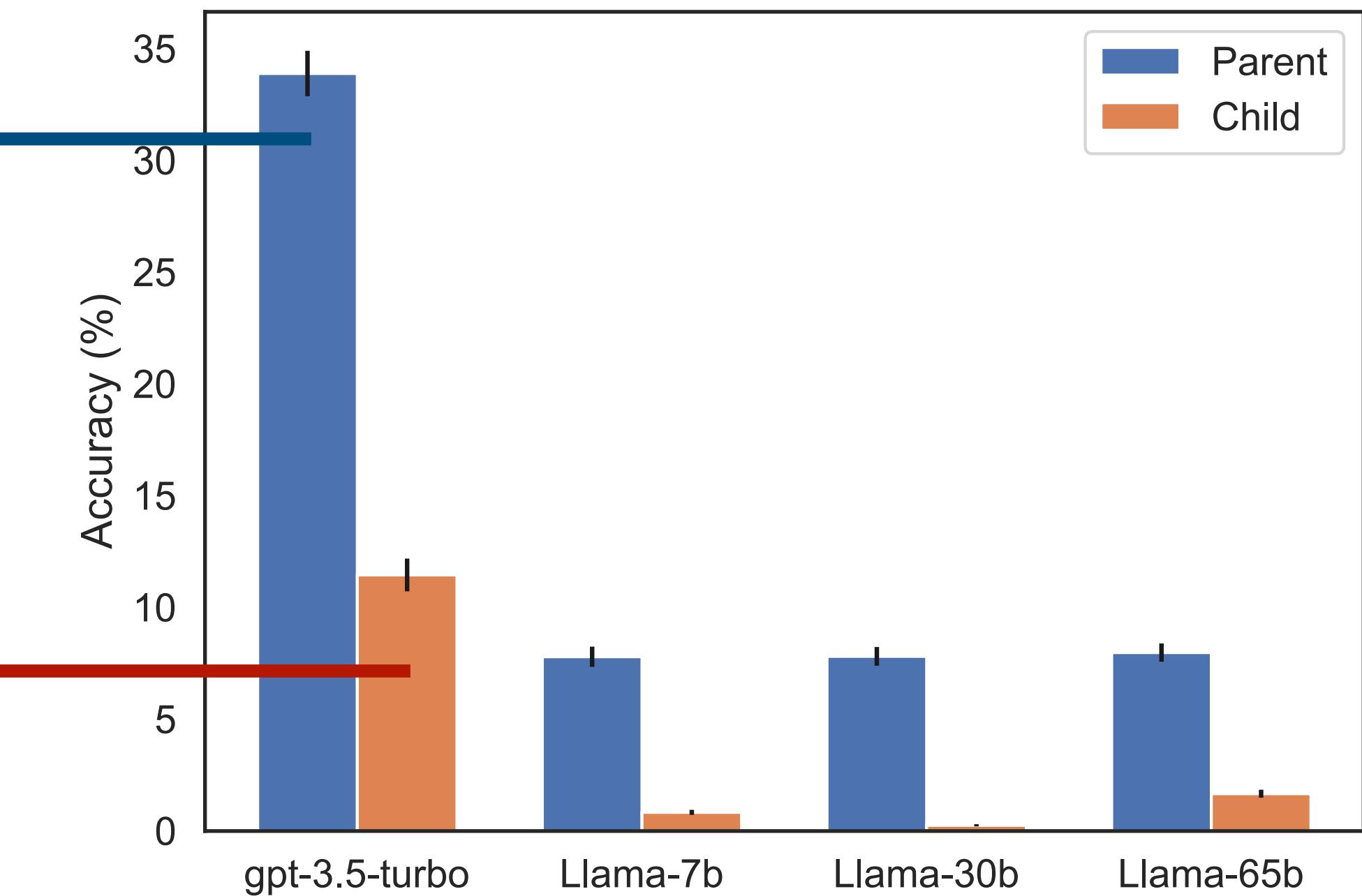
Training: After learning on data “A is B”.

✓ **Forward inference:** LMs can successfully answer “A is [?]” → “B”

✗ **Inverse inference:** they struggle to answer “[?] is B” → “A”

Q: Who is **Tom Cruise**’s mother?
A: *Mary Lee Pfeiffer*

Q: Who is *Mary Lee Pfeiffer*’s son?
A: **Tom Cruise**

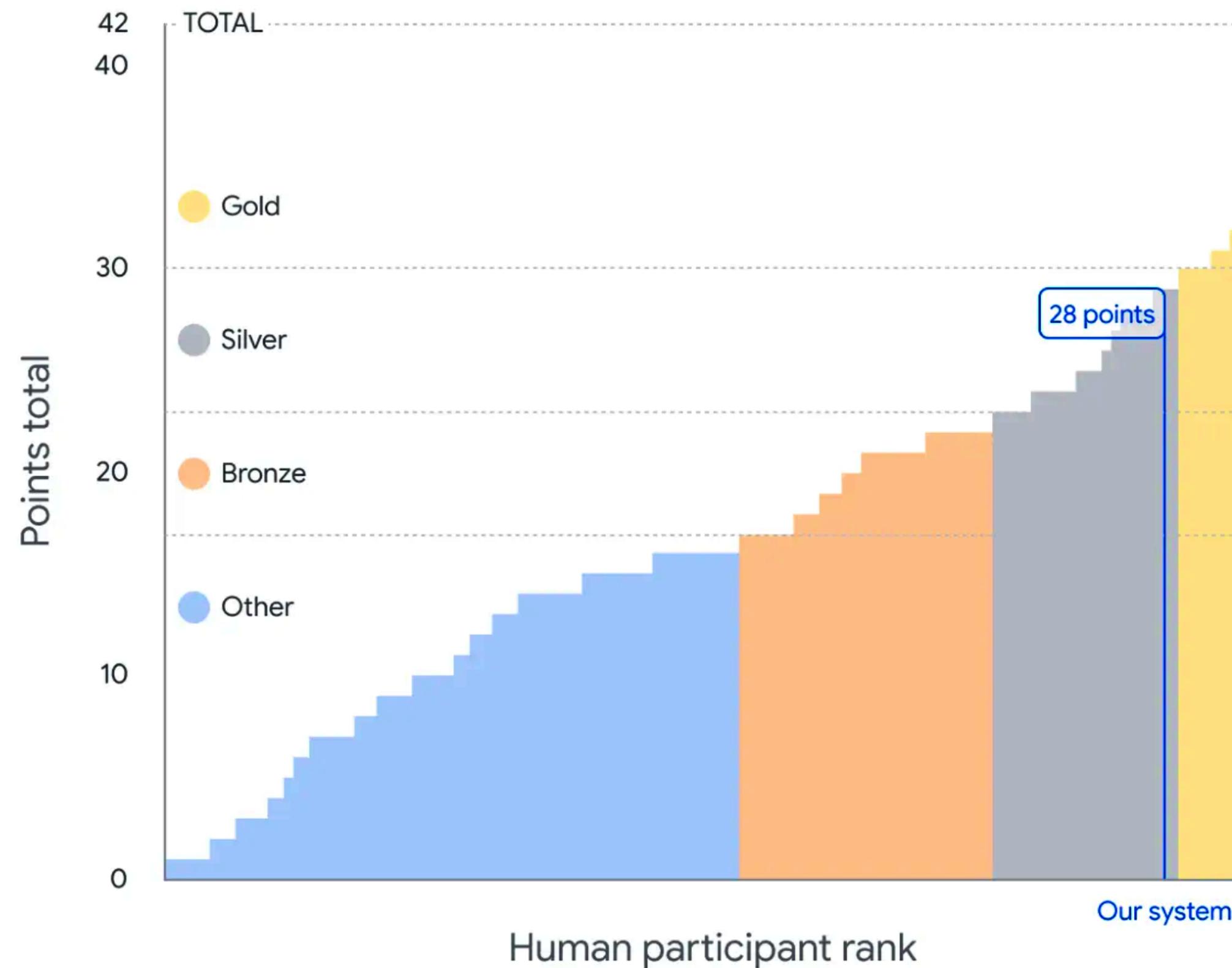


How Is Reasoning Conducted within LM

Part 1: Ethology - Topic 3: Reasoning

LMs' Reasoning Ability

AlphaGeometry excel at olympiad-level math problems



chain-of-thought reasoning by LMs

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

LMs Reasoning Is Sensitive to Perturbations

A screenshot of an AI interface. At the top right, a dark grey rounded rectangle contains the text "9.11 and 9.9 , which number is larger". Below this, on the left, there is a circular icon containing a stylized neural network or knot-like symbol. To its right, the text "9.11 is larger than 9.9." is displayed. At the bottom left, there are five small, light-colored icons: a speaker, a square, a circular arrow, a downward arrow, and a star.

Part 1: Ethology - Topic 3: Reasoning

LMs Reasoning Is Sensitive to Perturbations

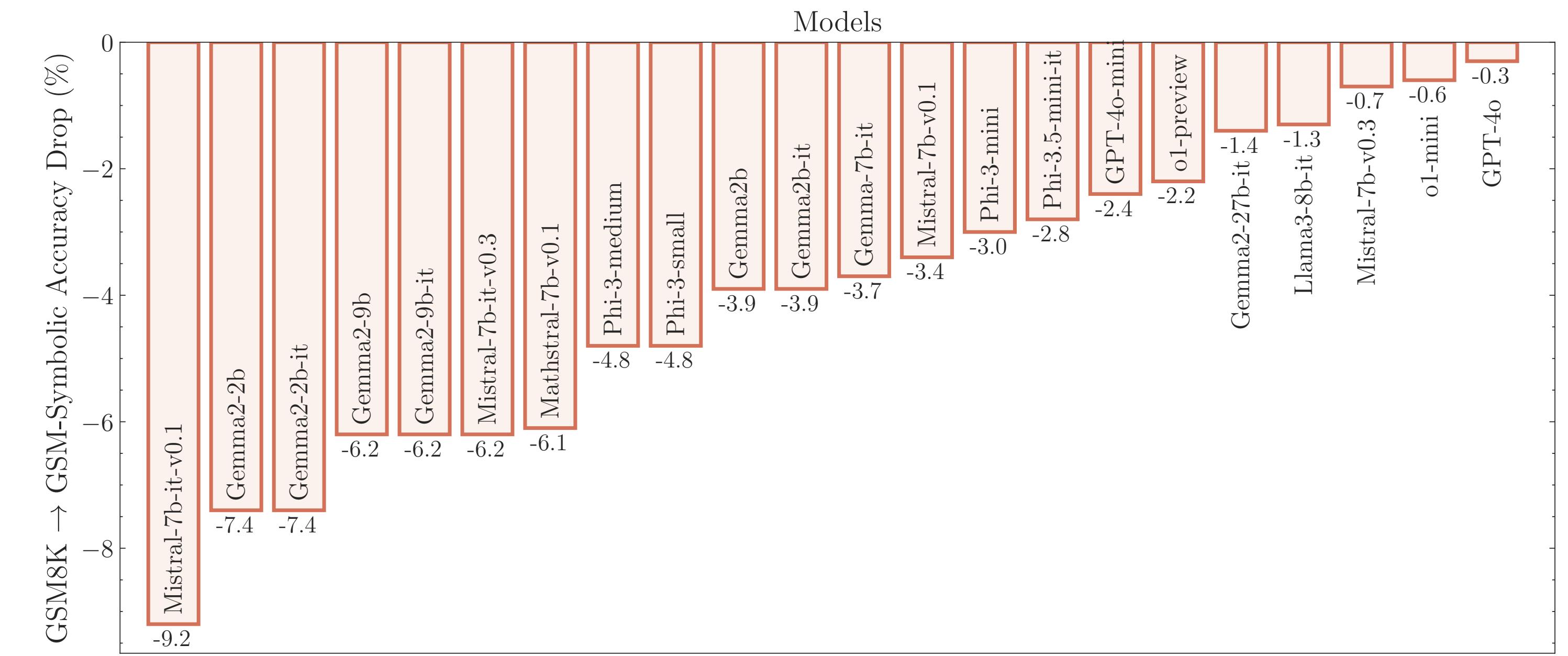
Grounding to real life matters

GSM Symbolic Template

```
When {name} watches her {family}, she gets out a variety of toys for him. The bag of building blocks has {x} blocks in it. The bin of stuffed animals has {y} stuffed animals inside. The tower of stacking rings has {z} multicolored rings on it. {name} recently bought a tube of bouncy balls, bringing her total number of toys she bought for her {family} up to {total}. How many bouncy balls came in the tube?

#variables:
- name = sample(names)
- family = sample(["nephew", "cousin", "brother"])
- x = range(5, 100)
- y = range(5, 100)
- z = range(5, 100)
- total = range(100, 500)
- ans = range(85, 200)

#conditions:
- x + y + z + ans == total
```



After replacing familiar nouns (e.g., “uncle,” “nephew”) with symbols, language models exhibit performance declines.

Part 1: Ethology - Topic 3: Reasoning

Logical Reasoning and Semantic Reasoning Entangles when “grounding to real life” poses biases to reasoning

GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

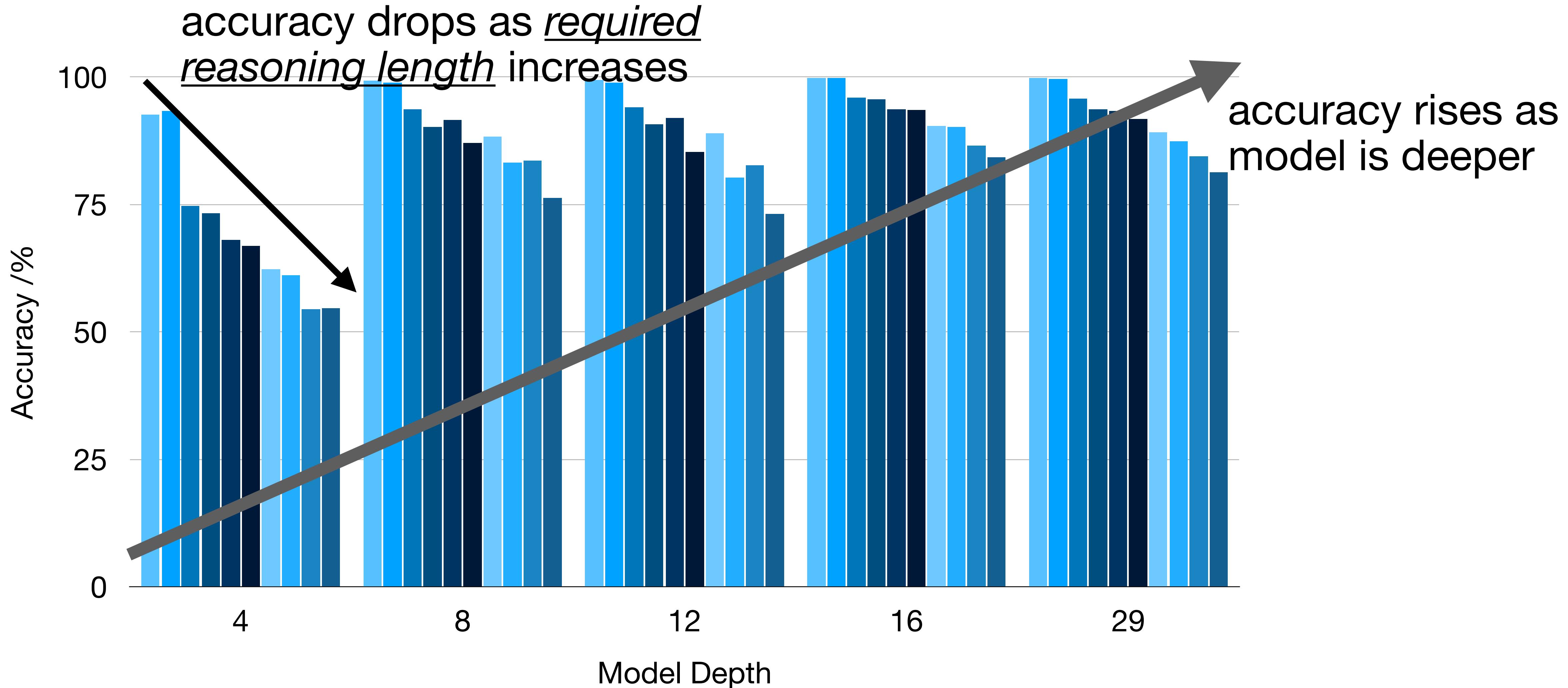
o1-mini: Oliver picks kiwis over three days as follows:

- Friday: 44 kiwis
- Saturday: 58 kiwis
- Sunday: Double the number he picked on Friday, which is $2 \times 44 = 88$ kiwis

However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185 kiwis.

After adding semantically related but logically irrelevant statements, LM fail to ignore them but wrongly combine them and make mistakes.

Deeper Models Reason Better



Hidden Features Encode Math Operations

dep(A,B) – at the end of problem description, does the model know *parameter A depend on B?*

e.g. `dep("Riverview High's Film Studio", "Film Studio's Messenger Bag") = true`

nece(A) – after question is asked, does the model know *if A is necessary for answering question?*

e.g. `nece("Riverview High's Film Studio") = false`

can_next(A) – in the middle of solution, does the model know if *A can be computed next?*

e.g. `can_next("Riverview High's Film Studio") = true`
`can_next("Riverview High's Dance Studio") = false`

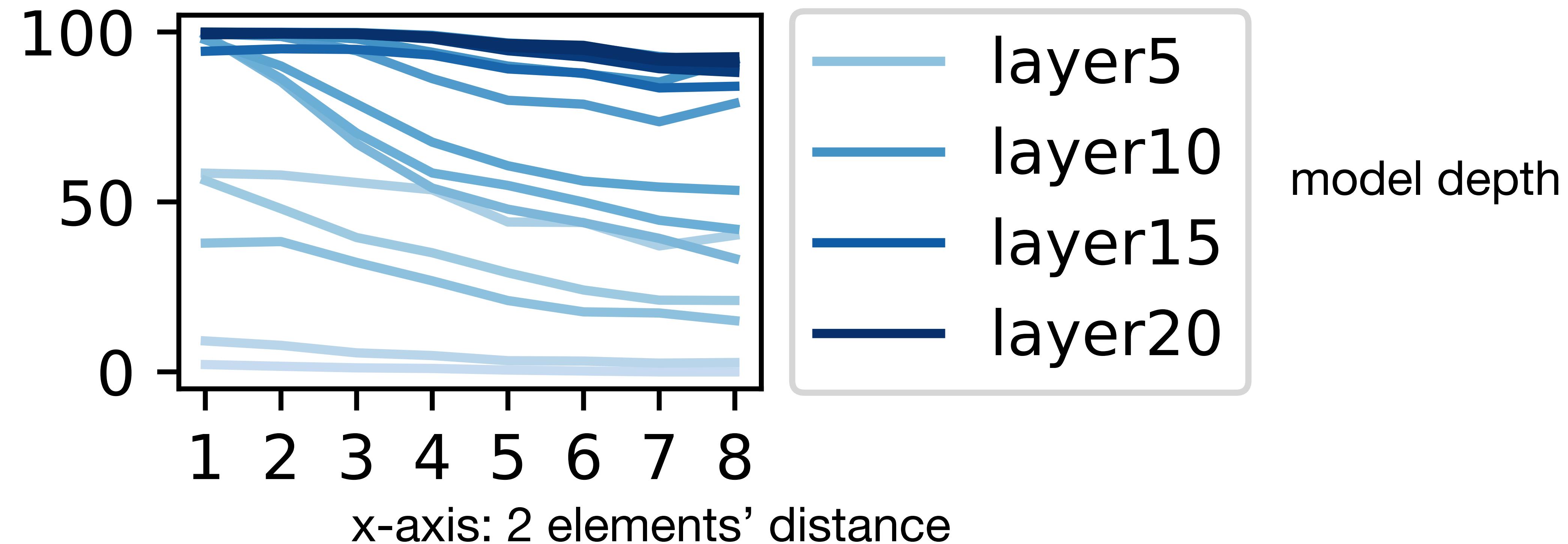
[Problem] The number of each Riverview High's Film Studio equals 5 times as much as the sum of each Film Studio's Backpack and each Dance Studio's School Daypack. ... The number of each Film Studio's Messenger Backpack equals 13.

[Question] How many Backpack does Central High have?

[Solution] Define Dance Studio's School Daypack as p ; so $p = 17$. Define Film Studio's Messenger Backpack as W ; so $W = 13$. Define Central High's Film Studio as B ; so $B = p + W = 17 + 13 = 7$. Define Film Studio's School Daypack as g ; $R = W + B = 13 + 7 = 20$; so $g = 12 + R = 12 + 20 = 9$. Define Film Studio's Backpack as w ; so $w = g + W = 9 + 13 = 22$. Define Central High's Backpack as c ; so $c = B * w = 7 * 22 = 16$. **[Answer]** 16.

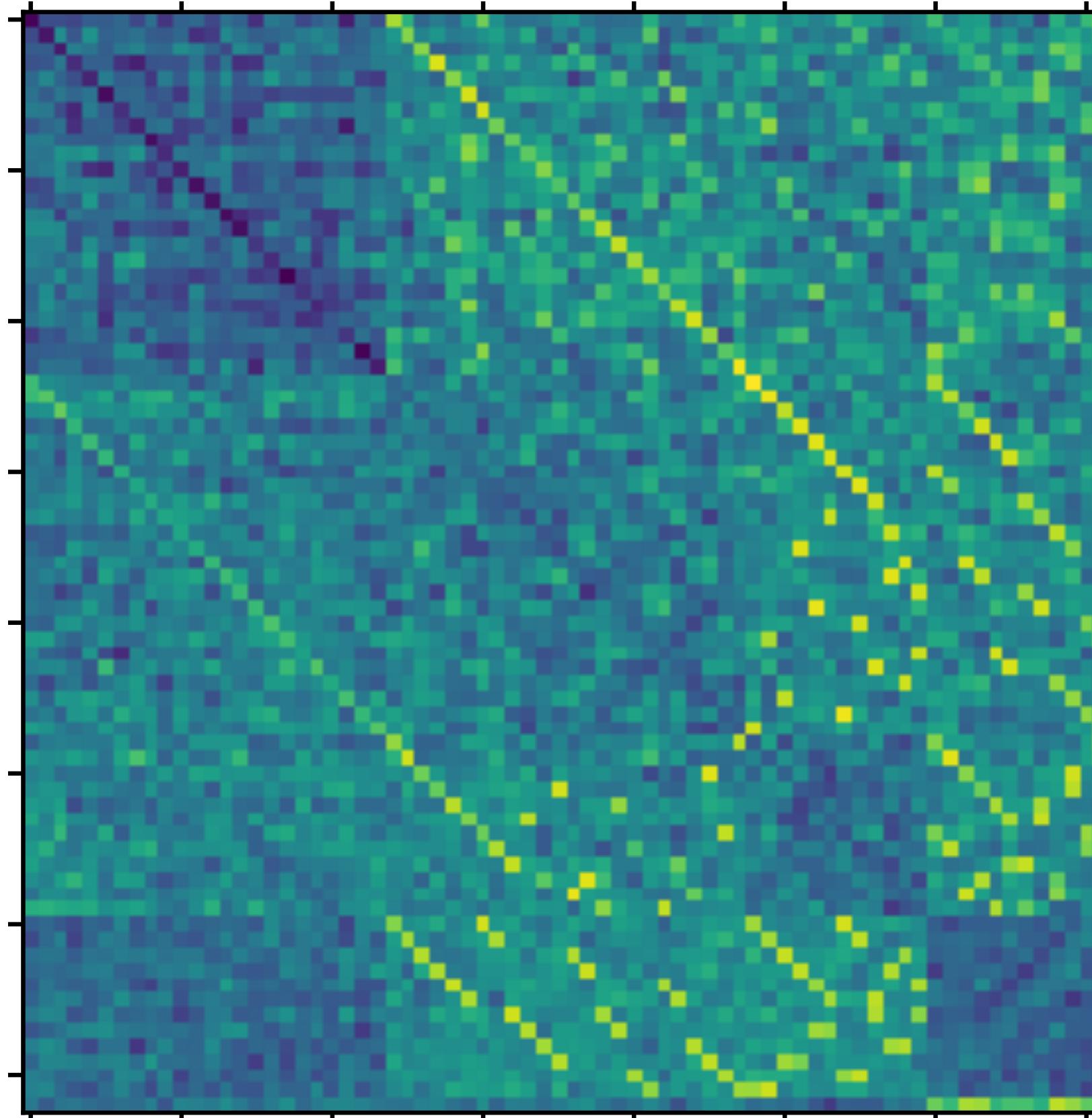
Every Math Operation Benefit from Depth

y-axis: $nece(X, Q)$, predicting if fact X is necessary for answering question Q

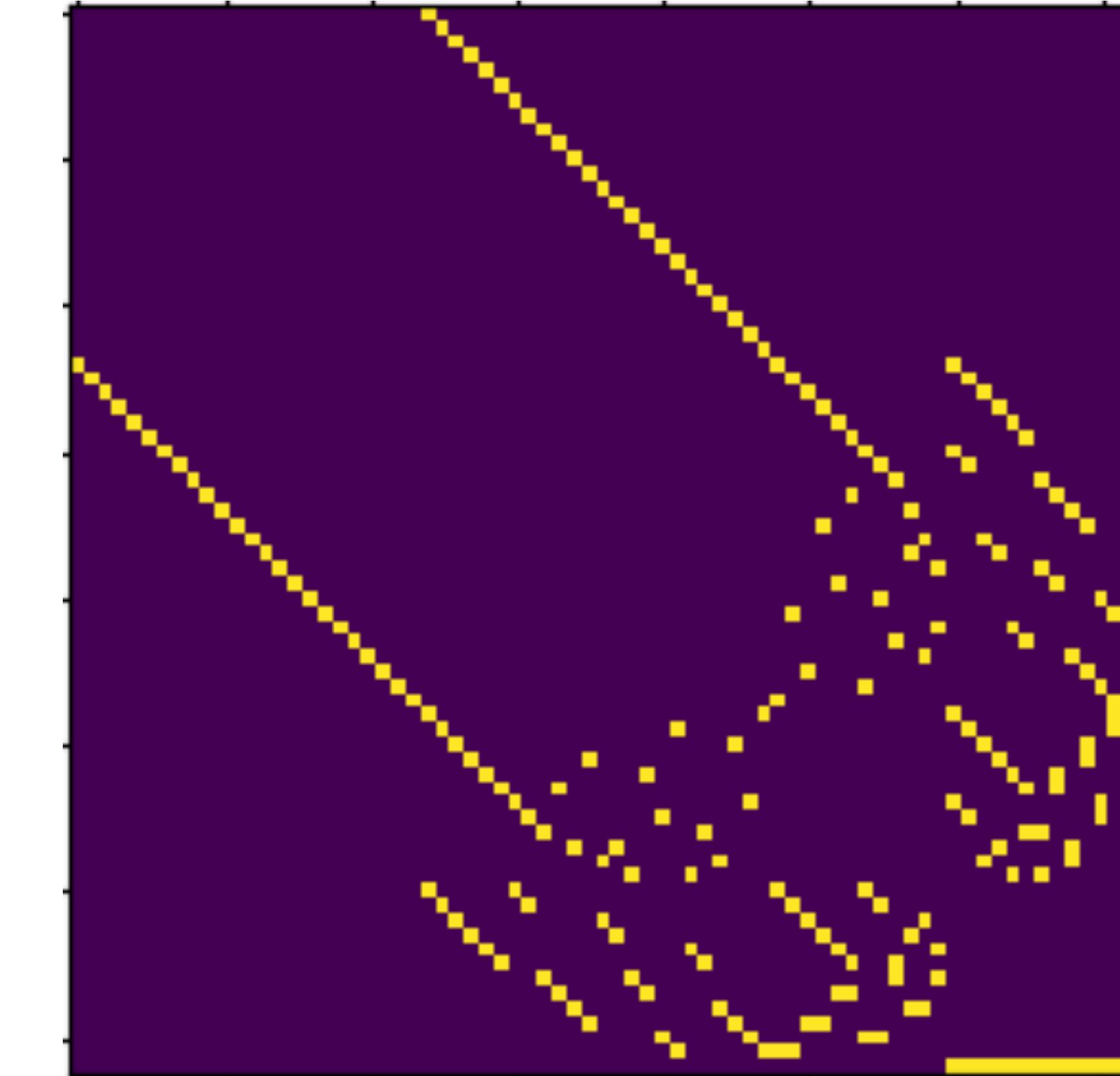


potential reason: every math operation needs certain depth of layers to stack with each other

LM Encodes Reachability after Learning Planning



Learned FFN weight W^M between graph nodes

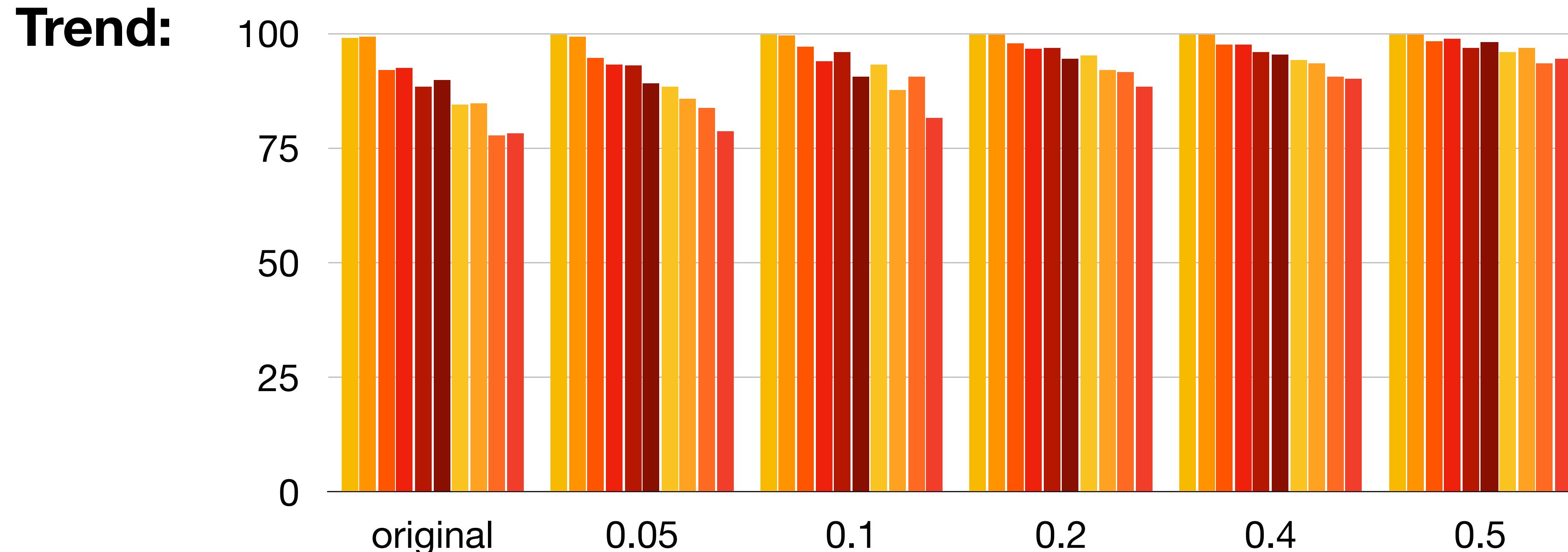


ground-truth connectivity

Learning on Error-Correction Data Helps

Example:

(Solution - retry rate 0.5) Define Dance Studio's School Daypack as p; so $p = 17$. Define Film Studio's School Daypack as [BACK]. Define Film Studio's Messenger Backpack as W; so $W = 13$. Define Central High's Classroom as [BACK]. Define Central High's Backpack as [BACK]. Define Central High's Film Studio as B; so $B = p + W = 17 + 13 = 7$. Define Film Studio's School Daypack as g; $R = W + B = 13 + 7 = 20$; so $g = 12 + R = 12 + 20 = 9$. Define Riverview High's Dance Studio as [BACK]. Define Film Studio's Backpack as w; so $w = g + W = 9 + 13 = 22$. Define Riverview High's Dance Studio as [BACK]. Define Central High's Backpack as c; so $c = B * w = 7 * 22 = 16$.



How Does Error-Retry Data Benefit Reasoning?

After training on error-retry data

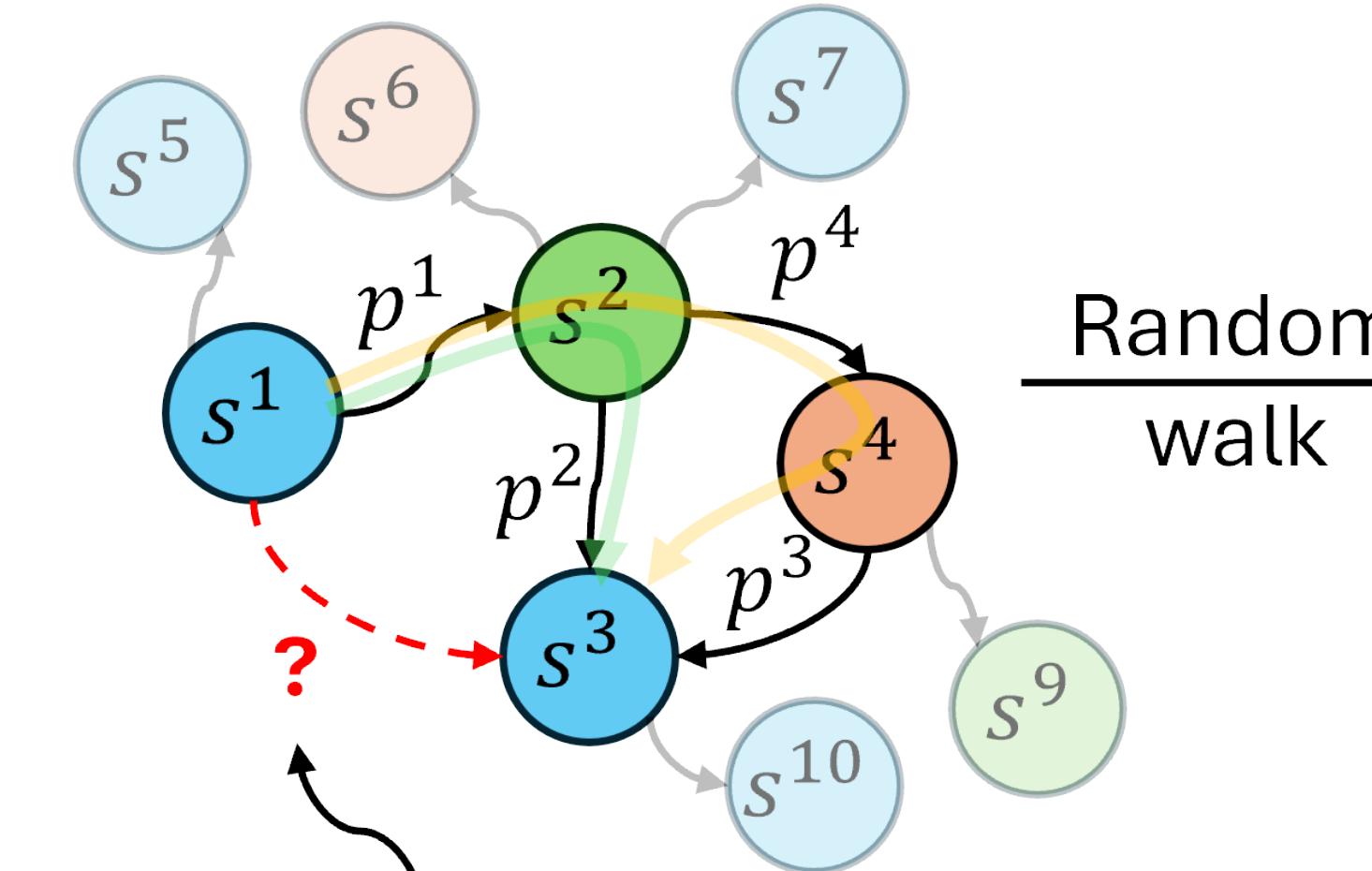
1. No need to mask out mistakes' loss terms.
2. During inference, LLMs hardly intentionally make mistakes,
3. Instead, they still try their best to answer correctly in the first place.

summary: retry data is beneficial and safe.

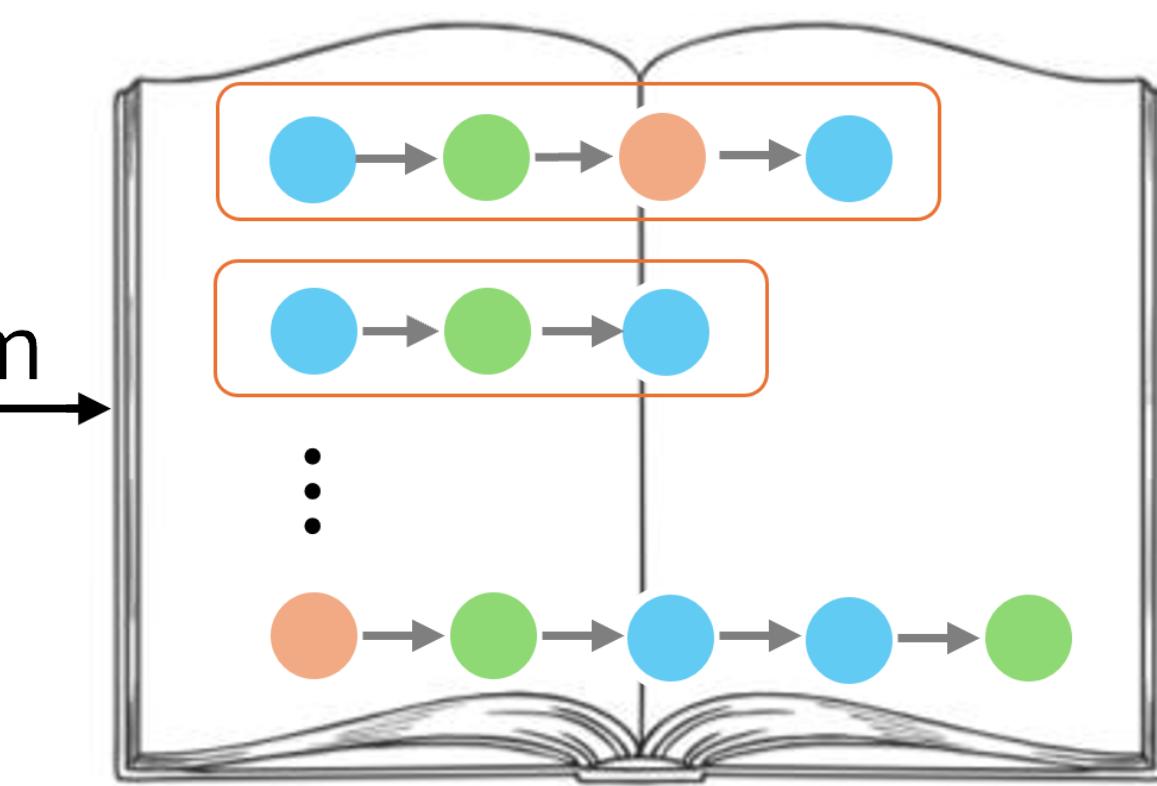
Reasoning Can Also Be Interpreted as Random Walk in Statement Space

Interpreting reasoning as a mixture or reasoning “walks” in claim graph G .

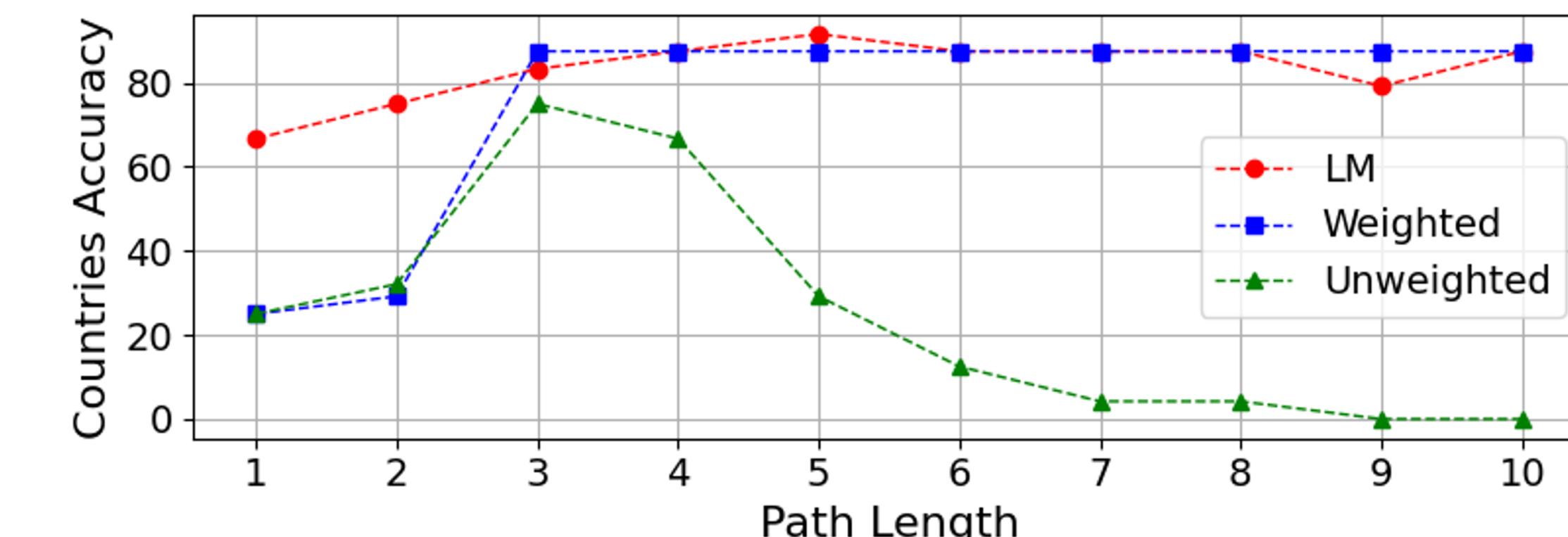
(Hypothetical) Reasoning graph G



Pre-training corpus D



A weighted average of paths approximate reasoning acc well.



Part 2: Physiology

How Do Components Function in Language Models?

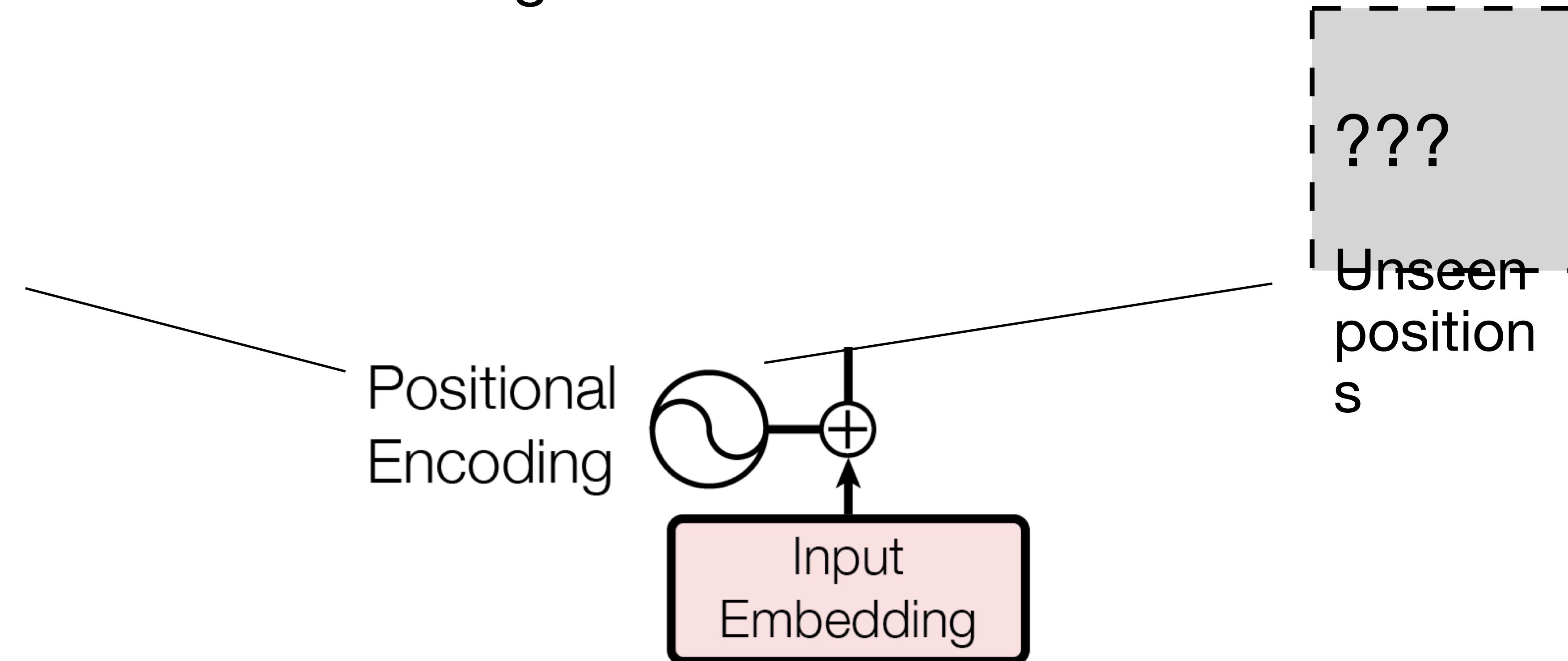
Topics

- **Attention:** Attention, position and context
- **Embeddings:** What is the function of word embeddings

Attention, Position and Context

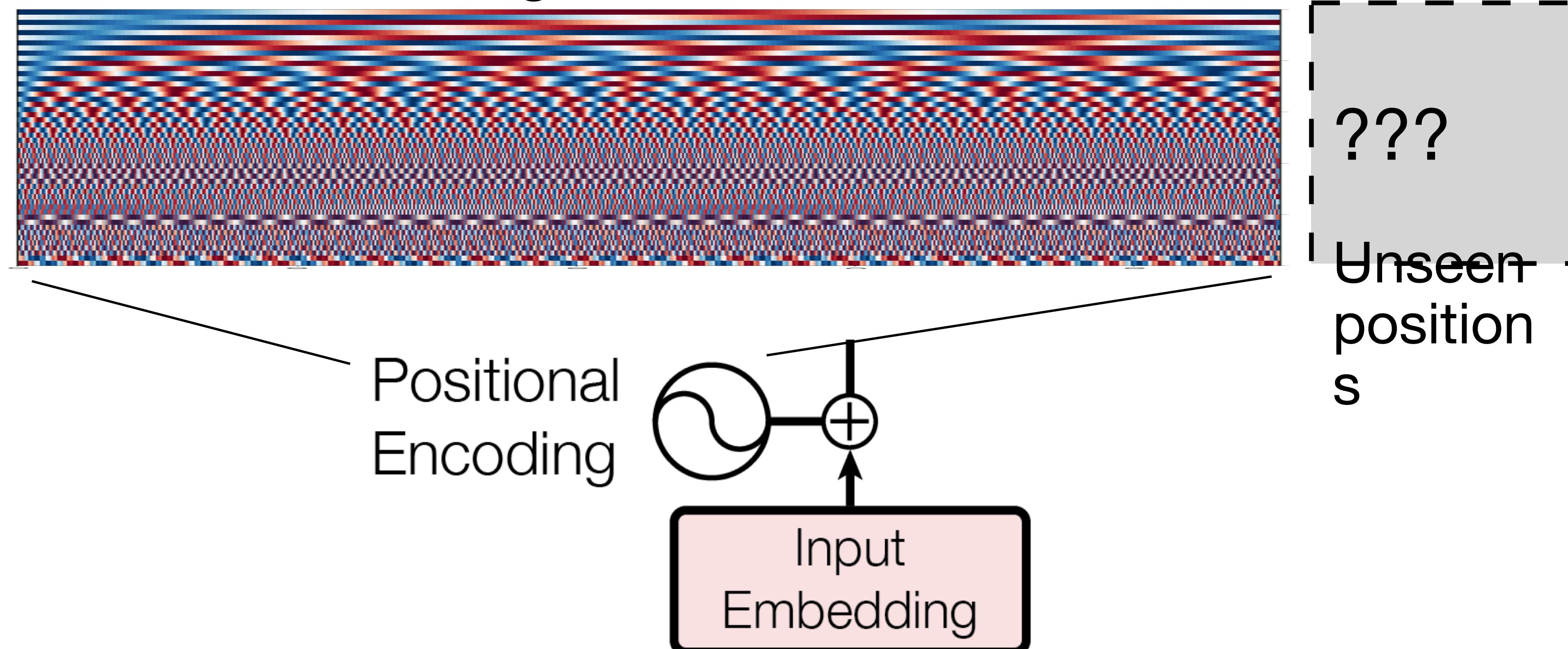
Absolute Positional Encoding: X

The **absolute positional encoding** used in vanilla Transformers is not generalizable to unseen lengths.



Absolute Positional Encoding: X

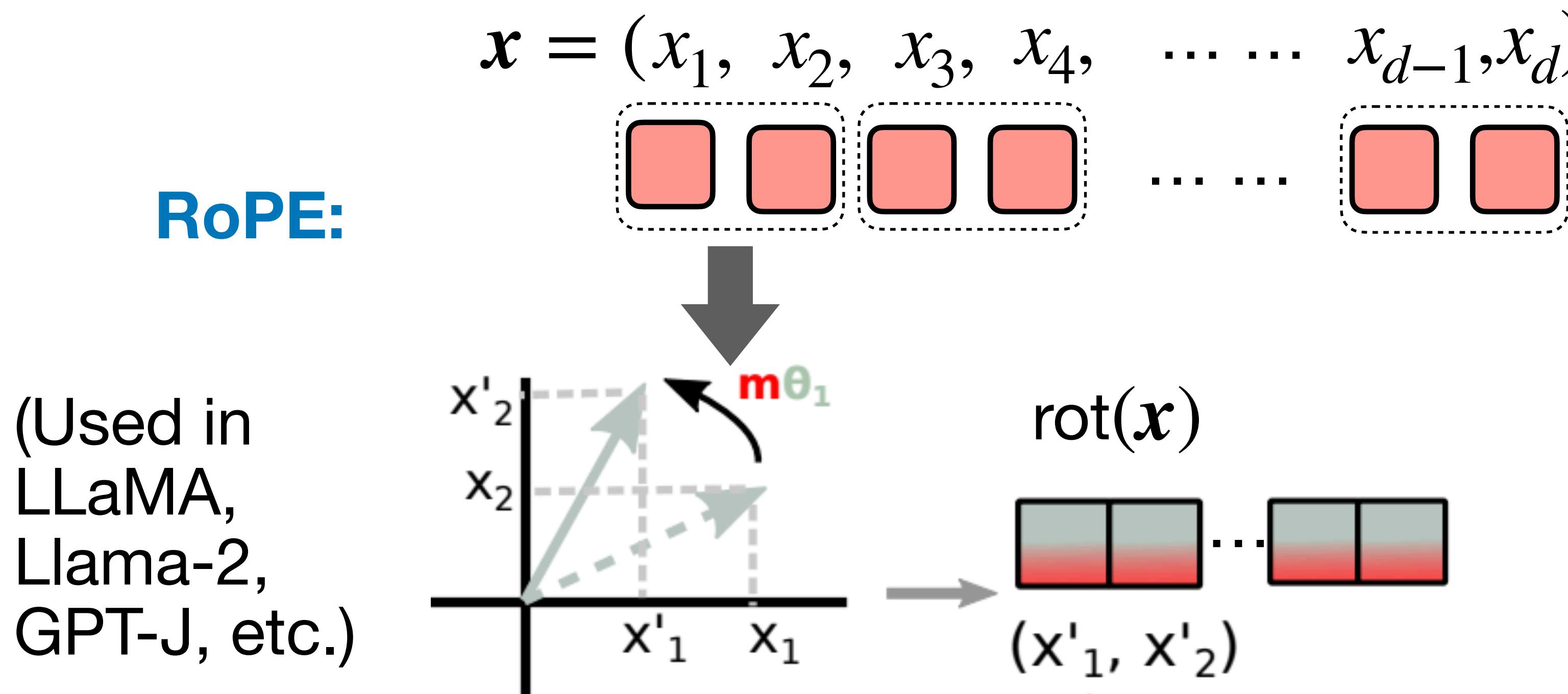
The **absolute positional encoding** used in vanilla Transformers is not generalizable to unseen lengths.



Relative Positional Encoding: ?

Relative positional encoding was proposed in the hope to alleviate this problem

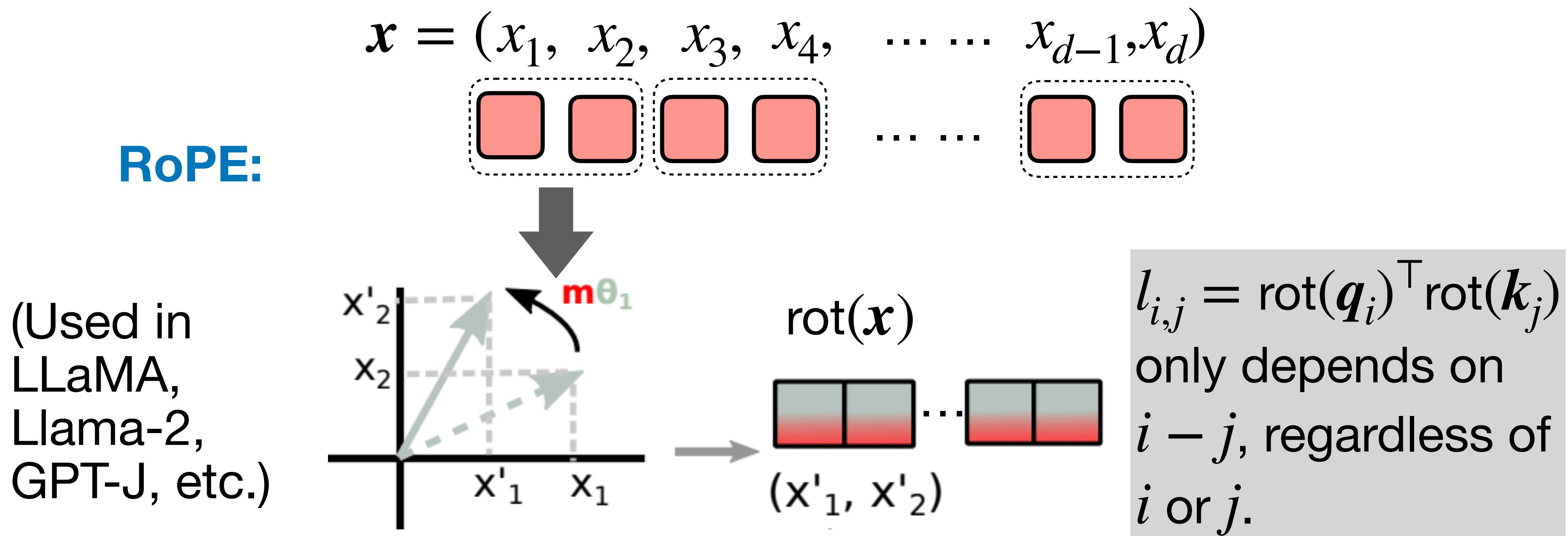
Core idea: determining attention based on distance



Relative Positional Encoding: ?

Relative positional encoding was proposed in the hope to alleviate this problem

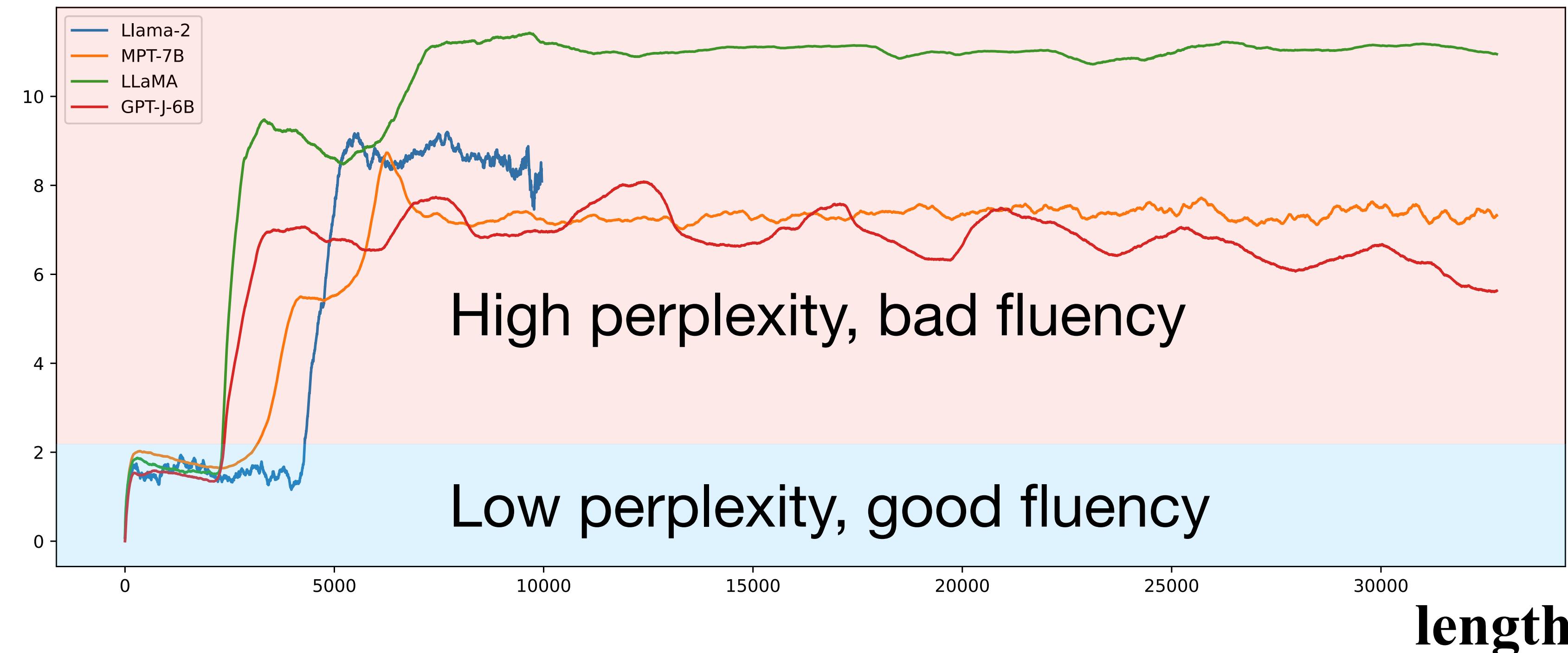
Core idea: determining attention based on distance



Relative Positional Encoding: ?

However, current LLMs still struggle on unseen lengths.

Negative Log-Likelihood (NLL, also = $\log(\text{perplexity})$) ↓



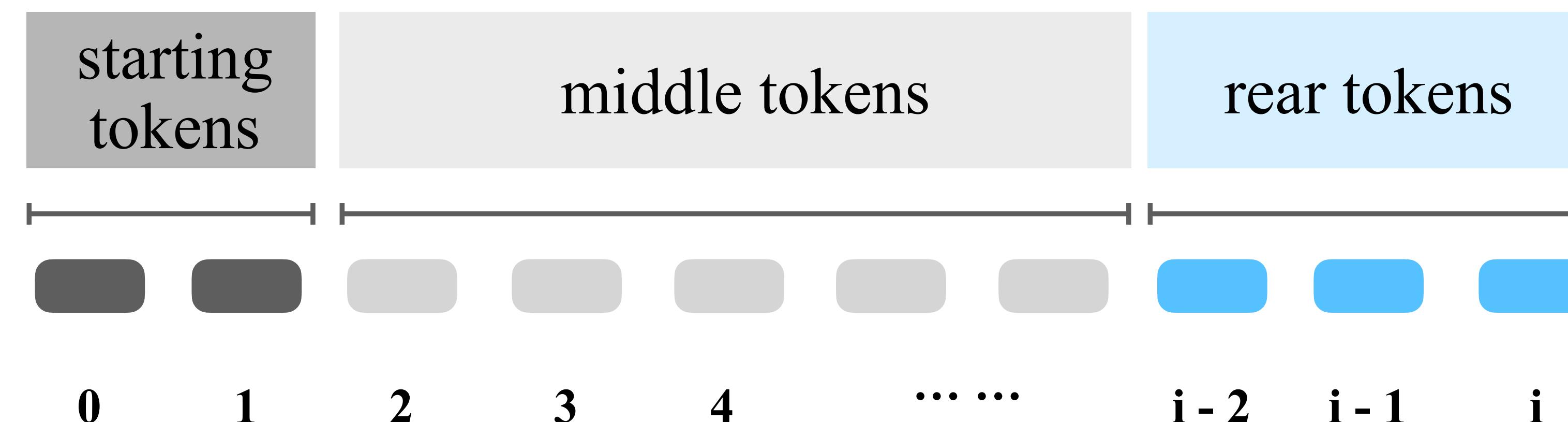
A Conceptual Model of Relative Position Encoding

essential for LLMs

encode more **absolute** position

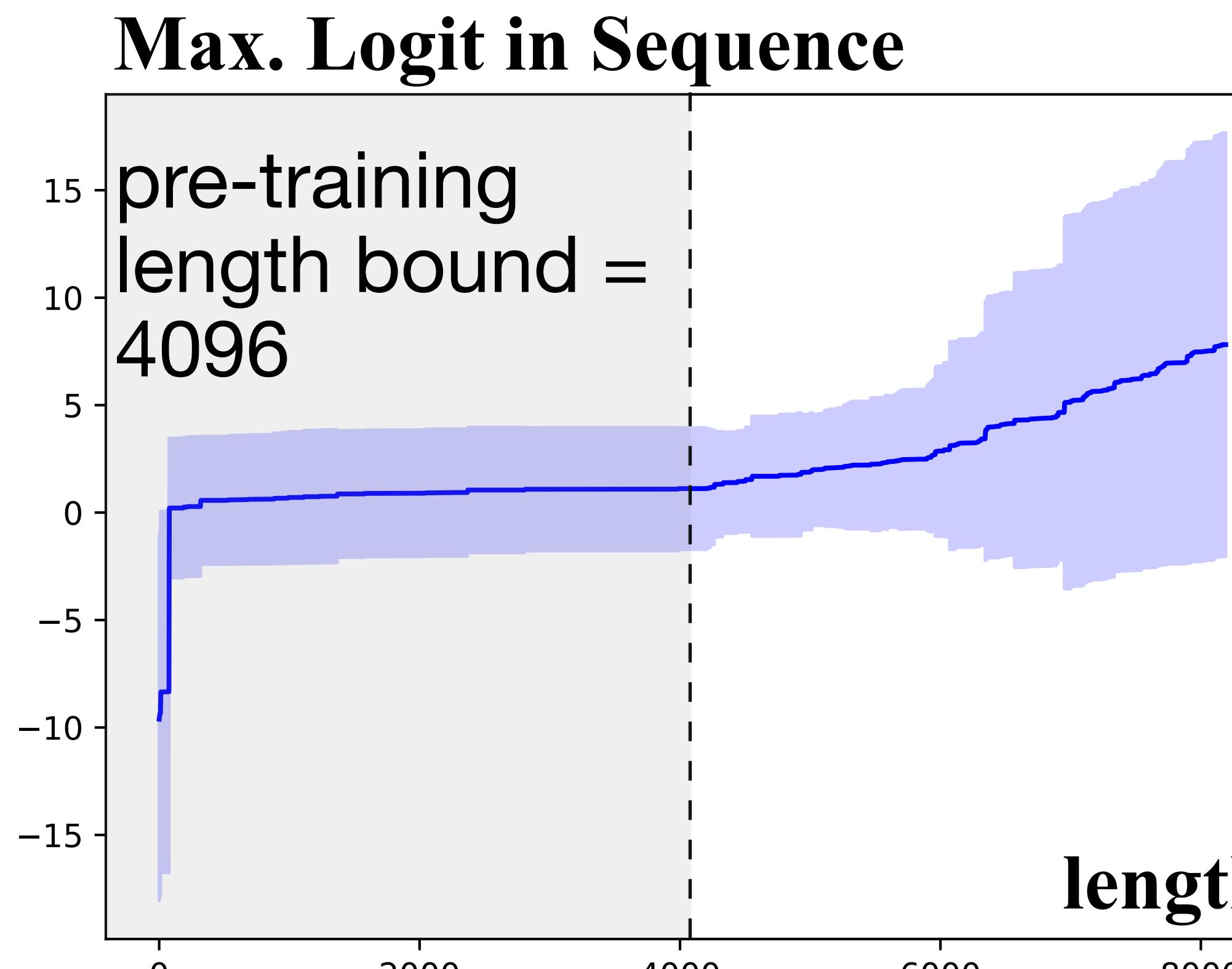
less position-sensitive

encode more **relative** position



Factor 1: Unseen Distance

Theorem 1 (Informal): For an attention mechanism using relative positional encoding, the attention logits must explode to infinities to differentiate previously unseen distances apart as the sequence length increases.

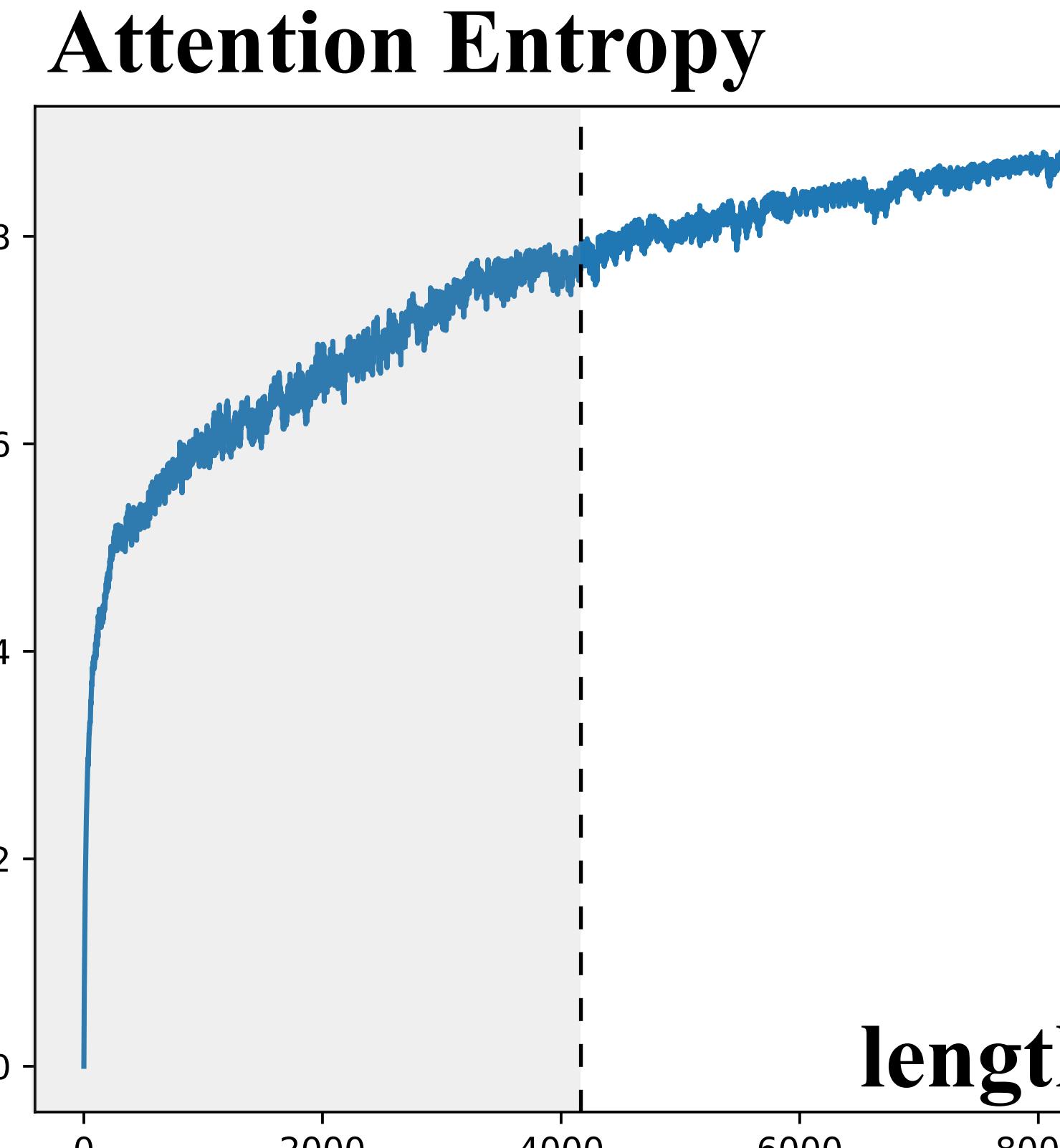


The attention logits in Llama-2 explode as length exceeds the pre-training limit.

Factor 2: Too many tokens

Longer texts require attention on more tokens.

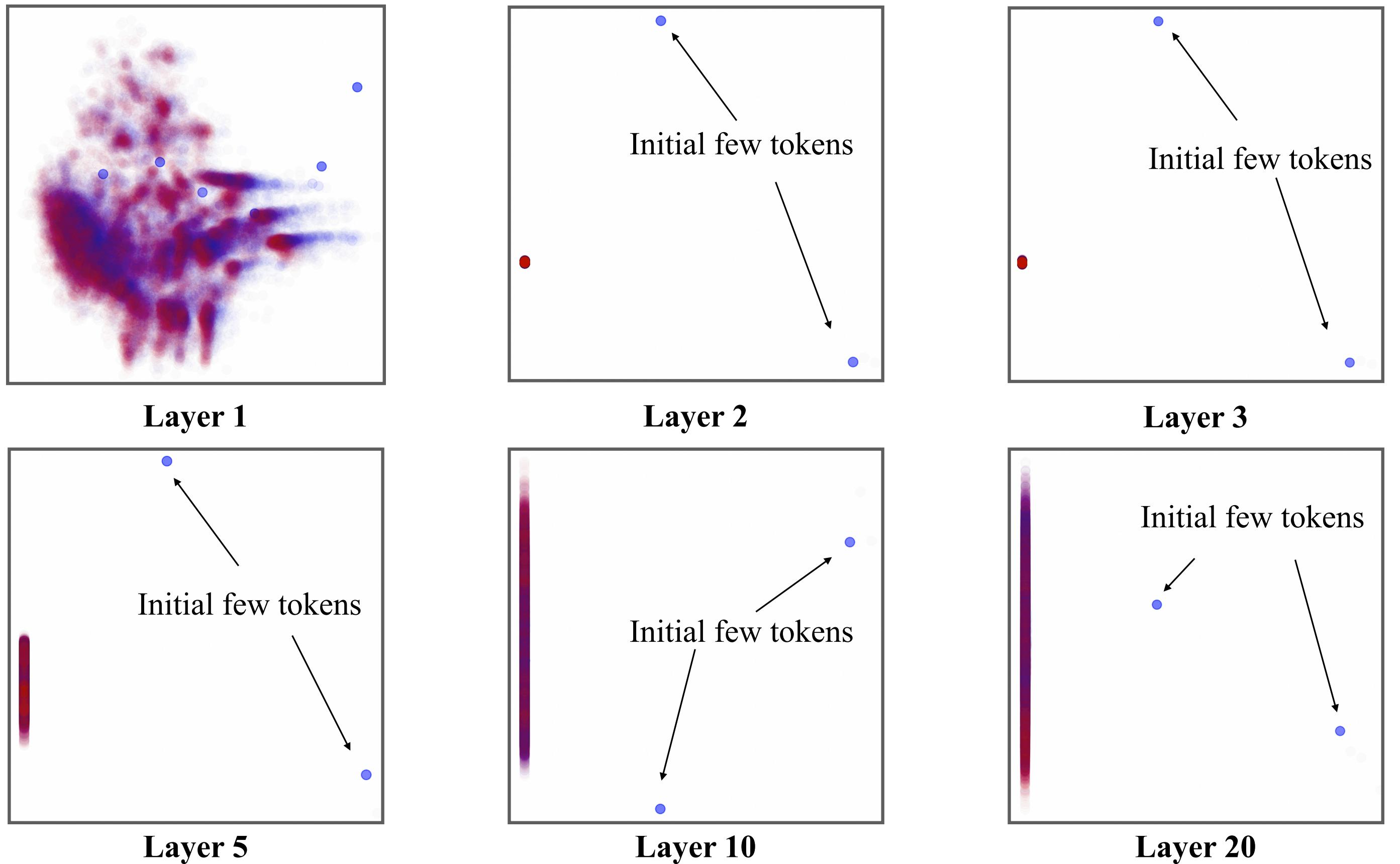
Theorem 2 (informal): If the attention logits are bounded, as the sequence becomes longer, the attention entropy grows to infinity.



The entropy of attention distribution in Llama-2 continuously increases with length.

Factor 3: Implicitly Encoded Position

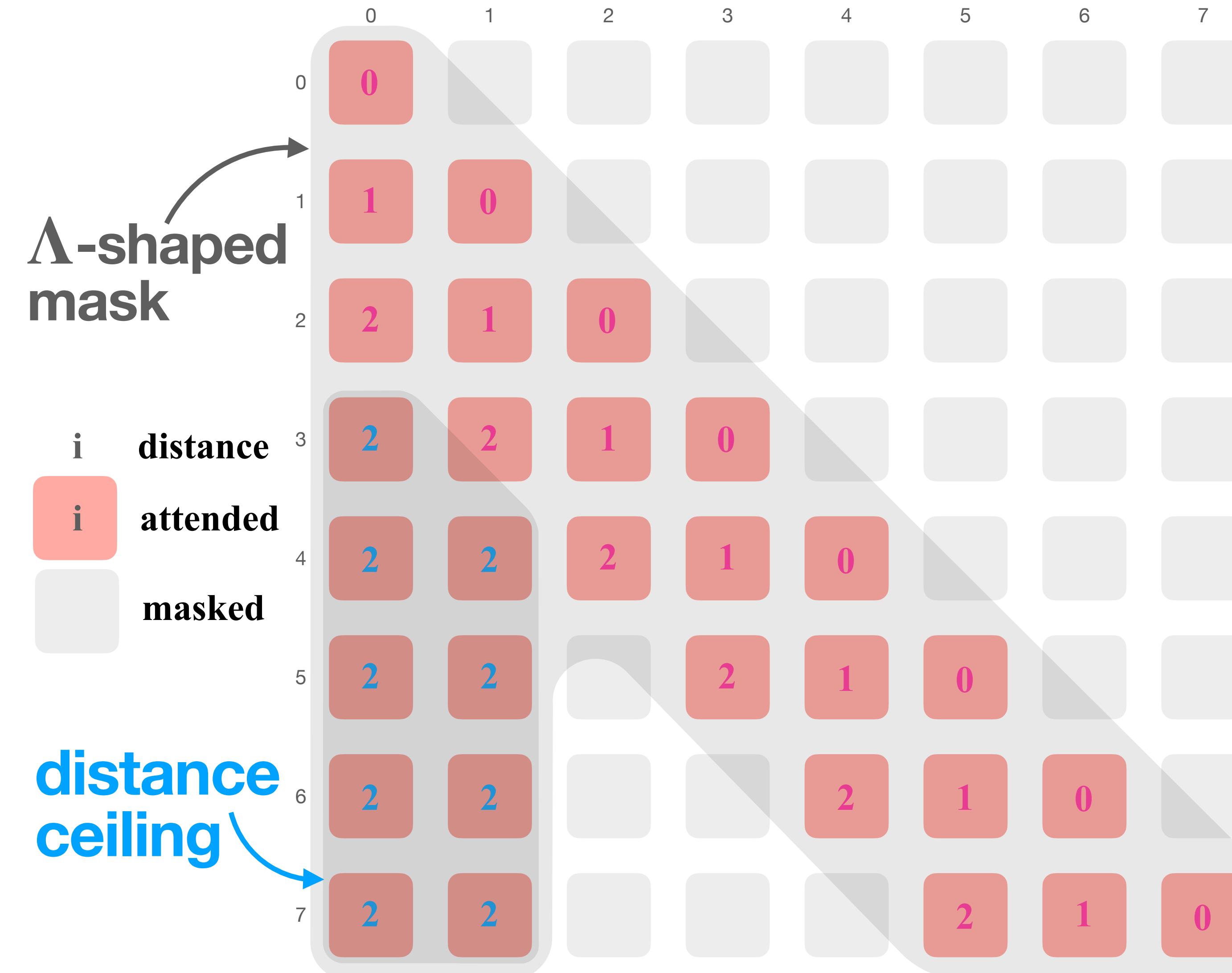
From layer 2 and higher, initial few tokens occupy a distinct feature space.



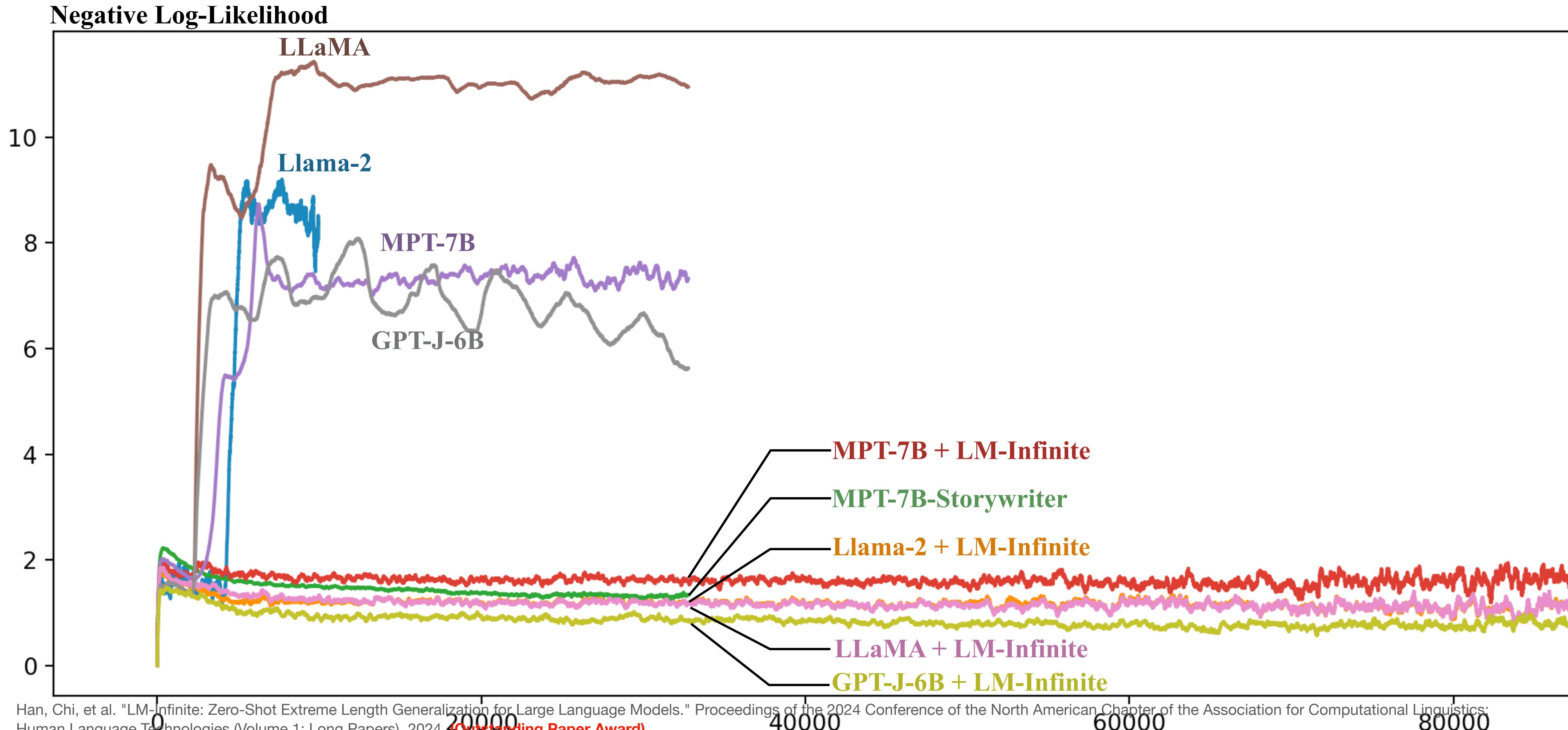
Theorem 3 (Informal): Even without absolute positional embeddings, attention can restore position information of tokens.

Part 2: Physiology - Topic 1: Attention

Solution: LM-Infinite

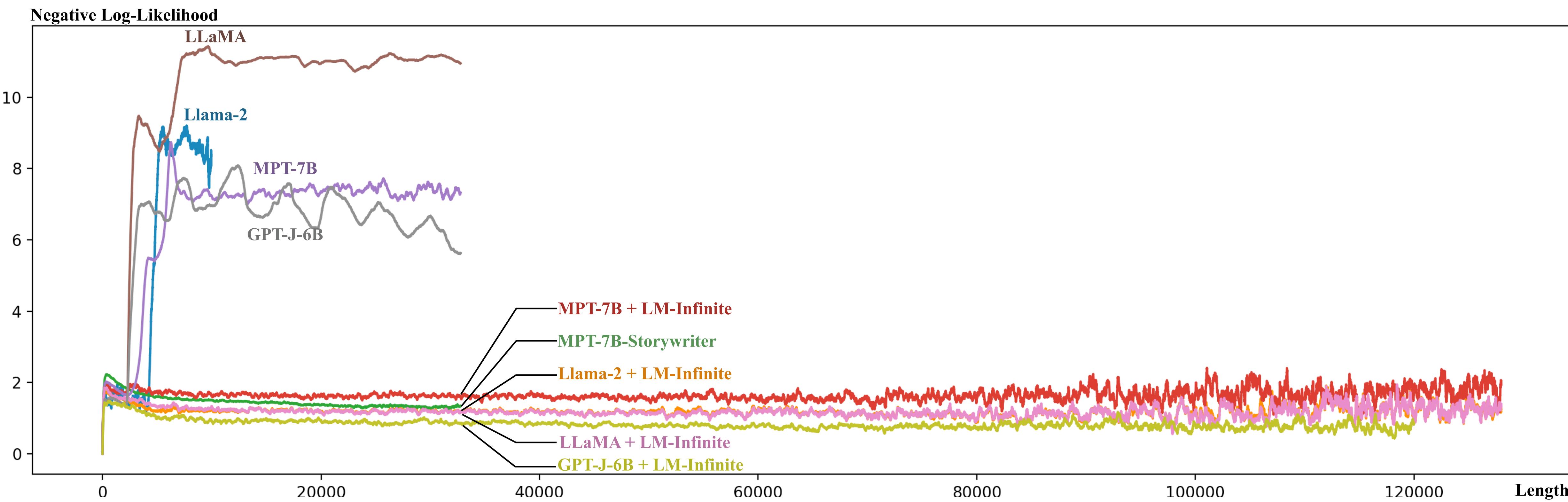


Length Generalization (to 200M length)



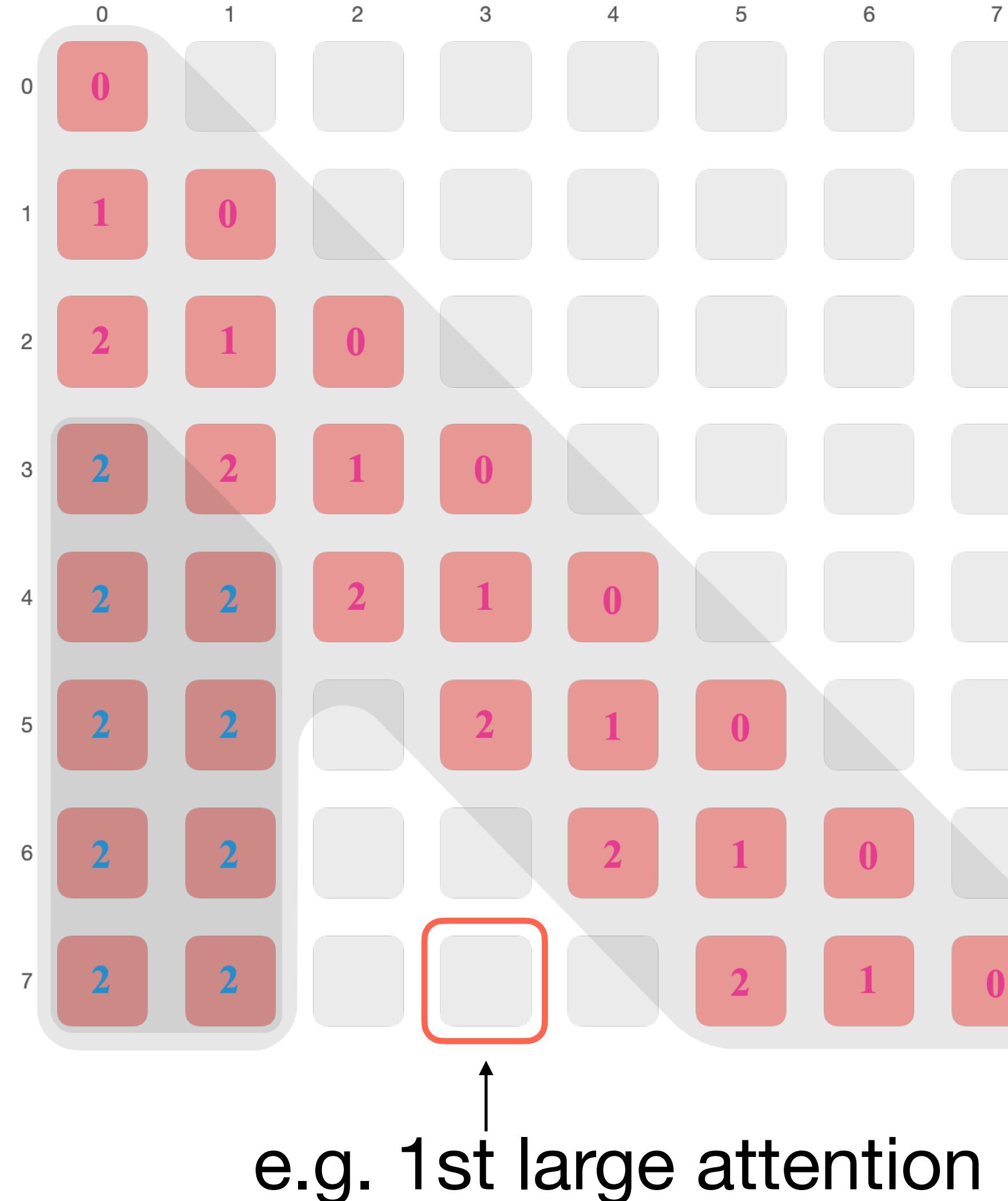
Part 2: Physiology - Topic 1: Attention

Length Generalization (to 200M length)



To Perceive Sensitive Information

Re-attending to top-k attention tokens

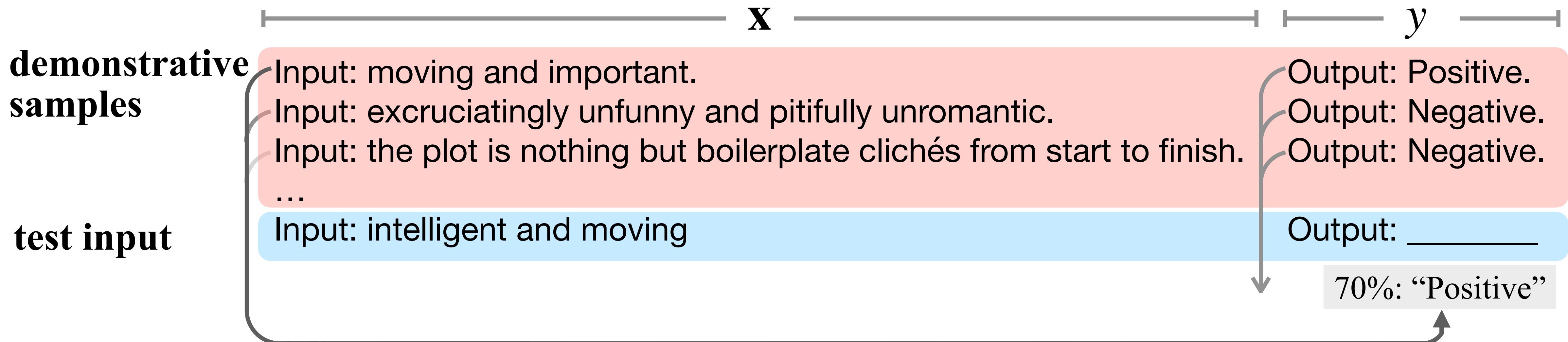


Why: to acquire key information that might be stored in the middle “ignored” region again.

How: selecting tokens with top-k (e.g., $k=4$) attention logits, and reintroducing them into attention.

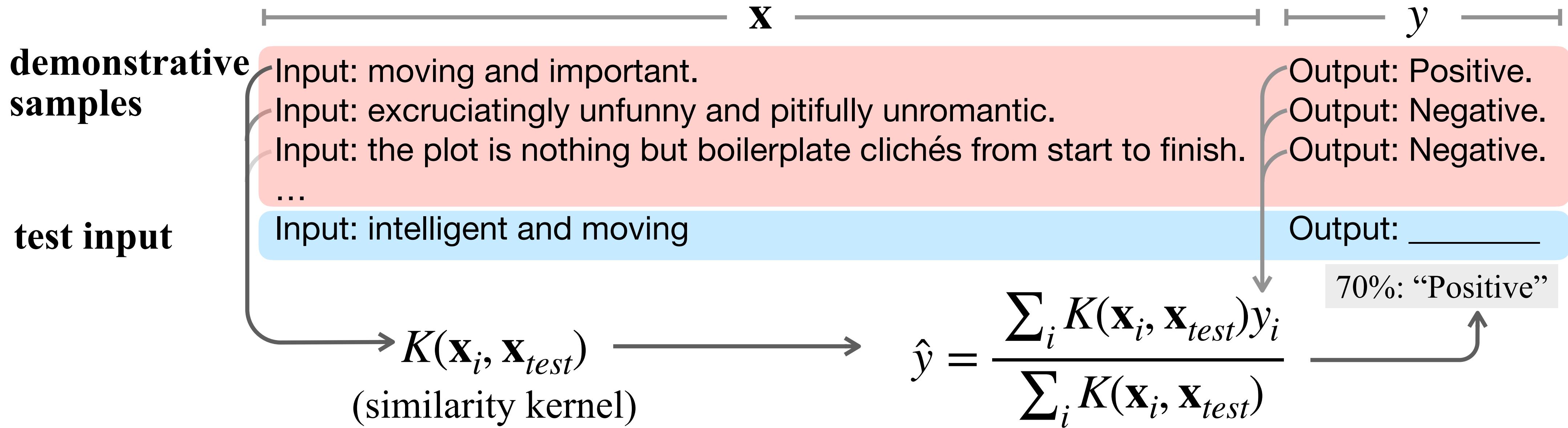
When: when solving information sensitive tasks like question answering, retrieving information from documents, etc.

Attention Also Explains In-Context Learning



In-context learning: completing tasks based on demonstrations

Attention Also Explains In-Context Learning



- The output \hat{y} is sampled from a weighted average over example outputs y_i (i.e., a kernel-regression)
- the weights are computed by a certain similarity metric $K(\mathbf{x}_i, \mathbf{x}_{text})$ (i.e., a kernel)

The Kernel Originates from Pre-Training

Kernel regression (hypothesized ICL algorithm)

$$\hat{\mathbf{y}} = \frac{\sum_{i=1}^n \mathbf{e}(y_i) \mathcal{K}(\mathbf{x}_{test}, \mathbf{x}_i)}{\sum_{i=1}^n \mathcal{K}(\mathbf{x}_{test}, \mathbf{x}_i)}$$

The Kernel Originates from Pre-Training

Kernel regression (hypothesized ICL algorithm)

$$\hat{\mathbf{y}} = \frac{\sum_{i=1}^n \mathbf{e}(y_i) \mathcal{K}(\mathbf{x}_{test}, \mathbf{x}_i)}{\sum_{i=1}^n \mathcal{K}(\mathbf{x}_{test}, \mathbf{x}_i)}$$

The kernel (similarity metric)

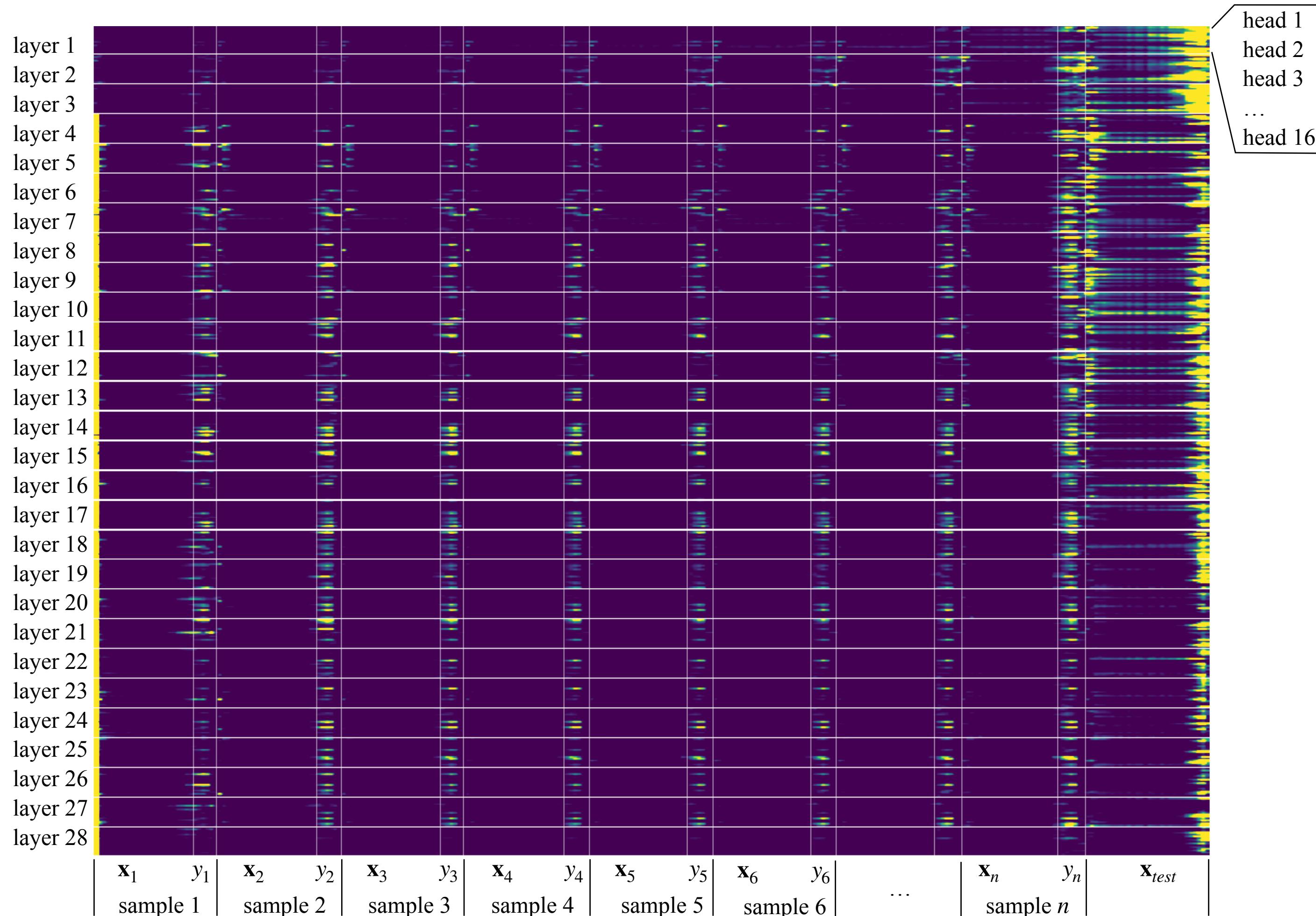
$$\mathcal{K}(\mathbf{x}, \mathbf{x}') = \text{vec}(T_{\mathbf{x}})^{\top} \Sigma_{p_{pre-train}}^{-1} \text{vec}(T_{\mathbf{x}'})$$

A representation of sample input
 \mathbf{x} for predicting the next token

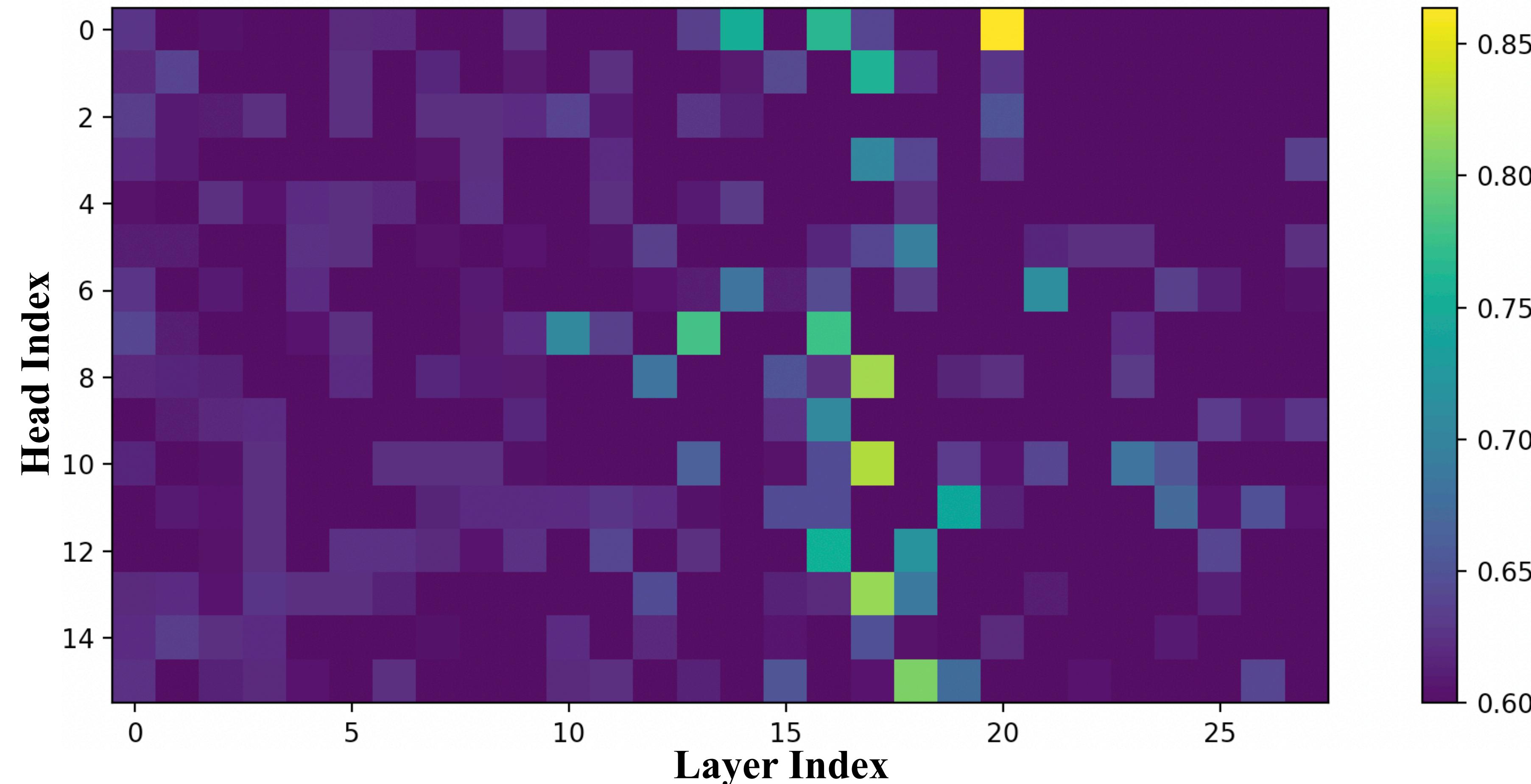
A matrix about the pre-training objective

Part 2: Physiology - Topic 1: Attention

The Attention Applies to y_i As Kernel Regression



The Explanation Aligns With the Model Output



Certain attention heads can reconstruct the LLM ICL output with the explanation.

Part 2: Physiology - Topic 1: Attention

The Attention

Method	sst2	mnli	rotten-tomatoes	tweet_eval (hate)	tweet_eval (irony)	tweet_eval (offensive)
GPT-J-6B ICL	0.805	0.383	0.671	0.539	0.519	0.542
all-MiniLM-L6-v2	0.503	0.321	0.478	0.548	0.491	0.588
bert-base-nli-mean-tokens KR	0.523	0.325	0.502	0.545	0.479	0.597
task-specific best head KR	0.789	0.974	0.692	0.560	0.584	0.560
overall best head KR	0.766	0.808	0.648	0.462	0.446	0.462

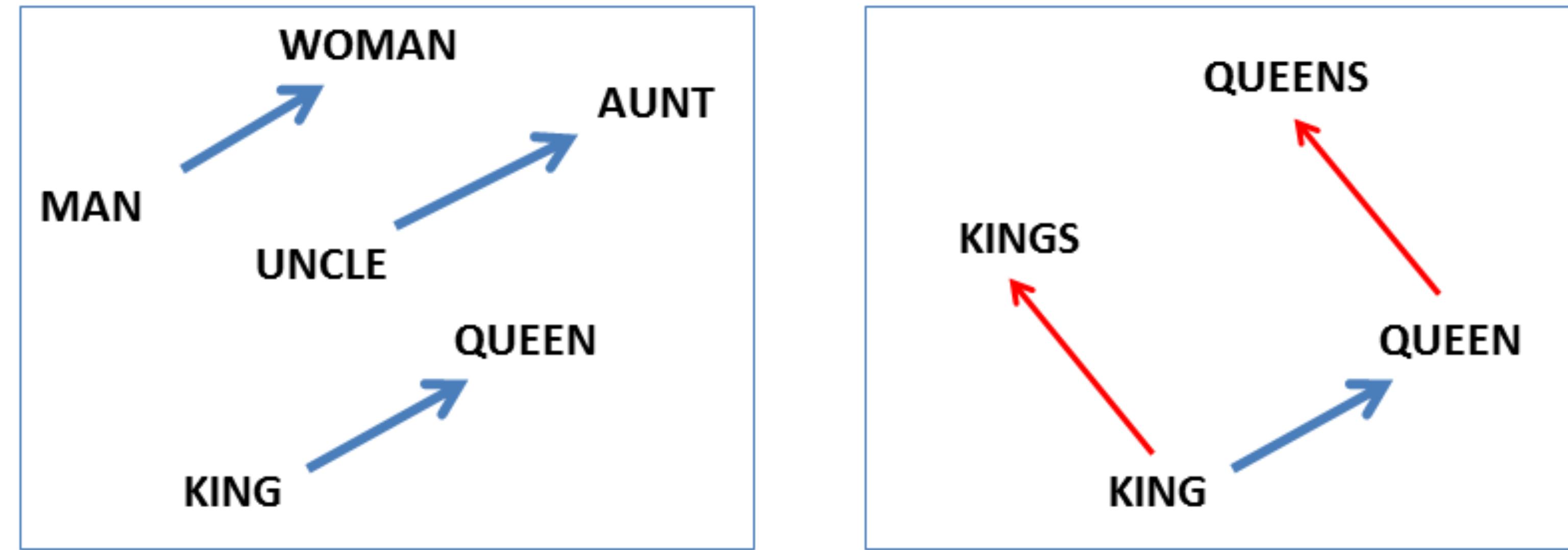
The KR explanation explained most tasks well (except for MNLI)

KR based on baseline sentence embeddings models

What Is the Function of Word Embeddings

What Do Word Embeddings Embed?

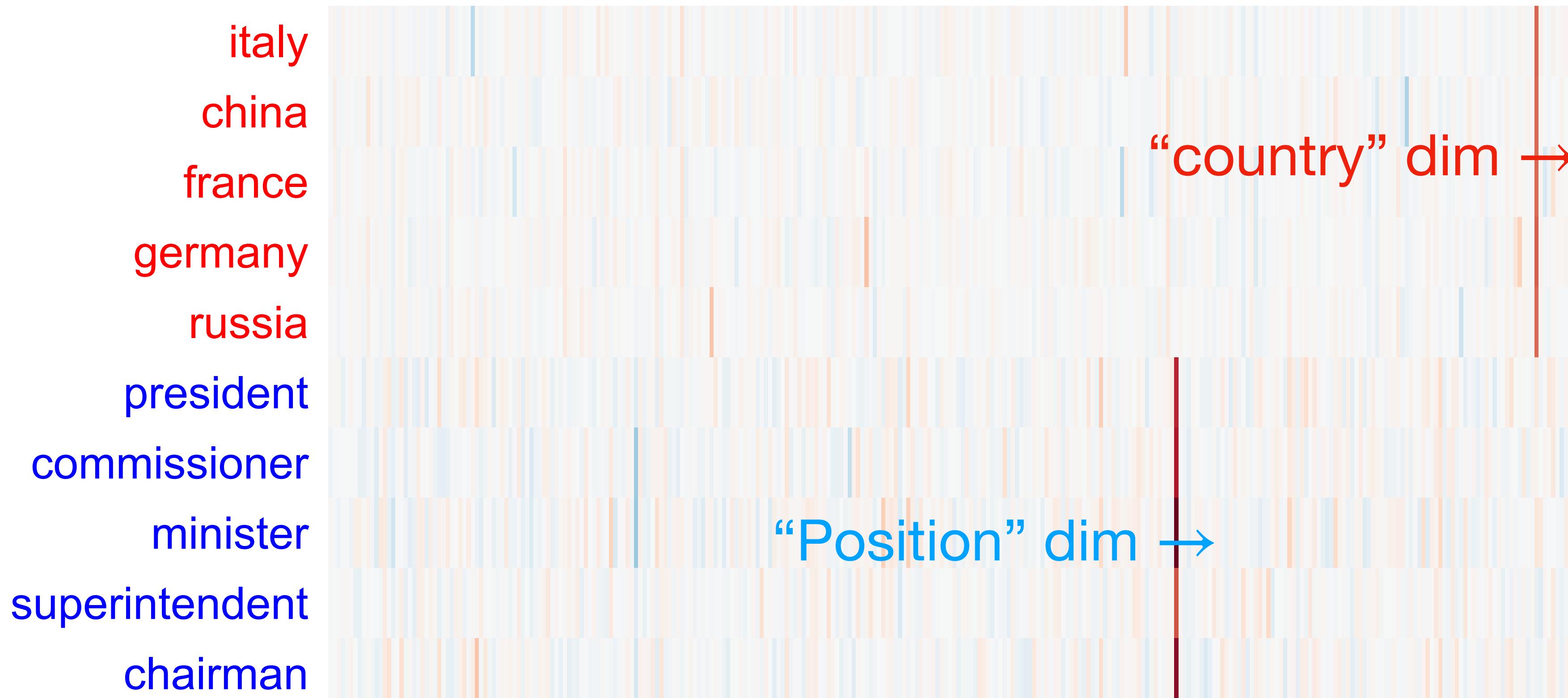
Previous papers mostly focus on word-level interpretations



(a) Analogical Relations (metric space)

What Do Word Embeddings Embed?

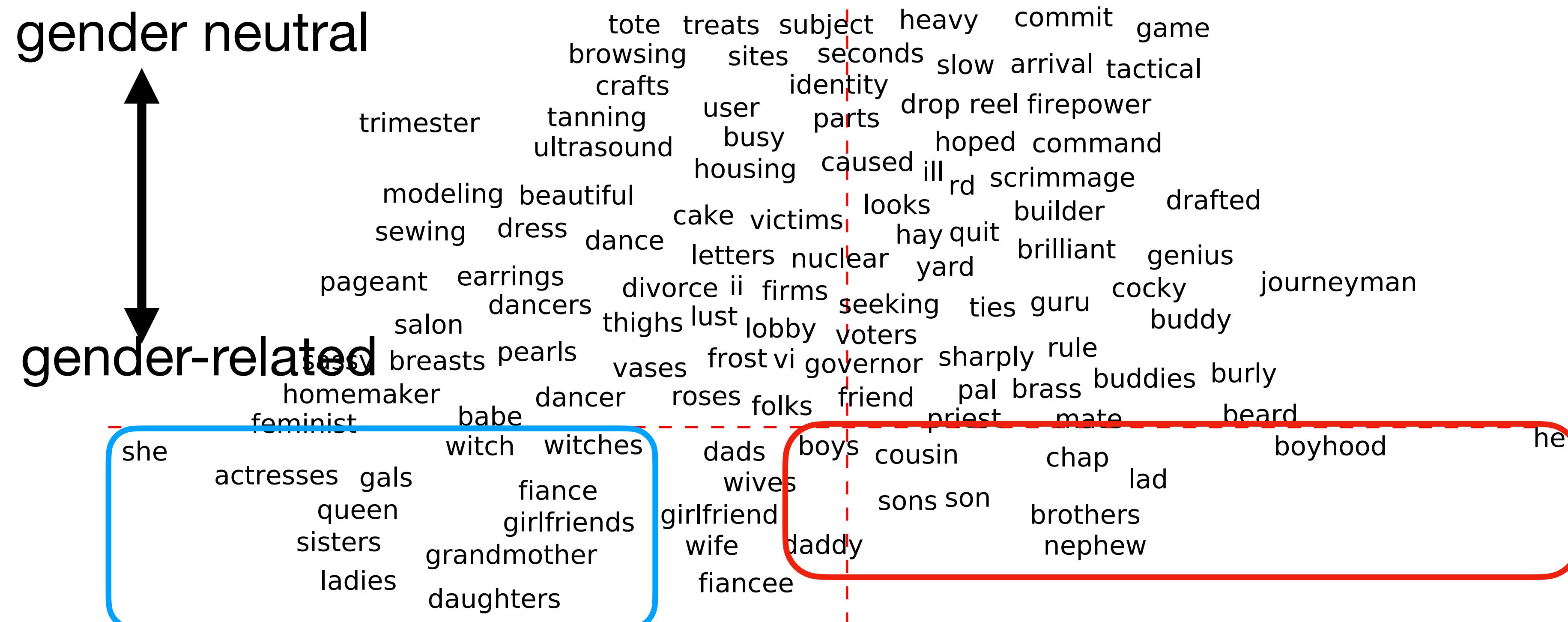
Previous papers mostly focus on word-level interpretations



(b) Meaningful Dimensions (linear Space)

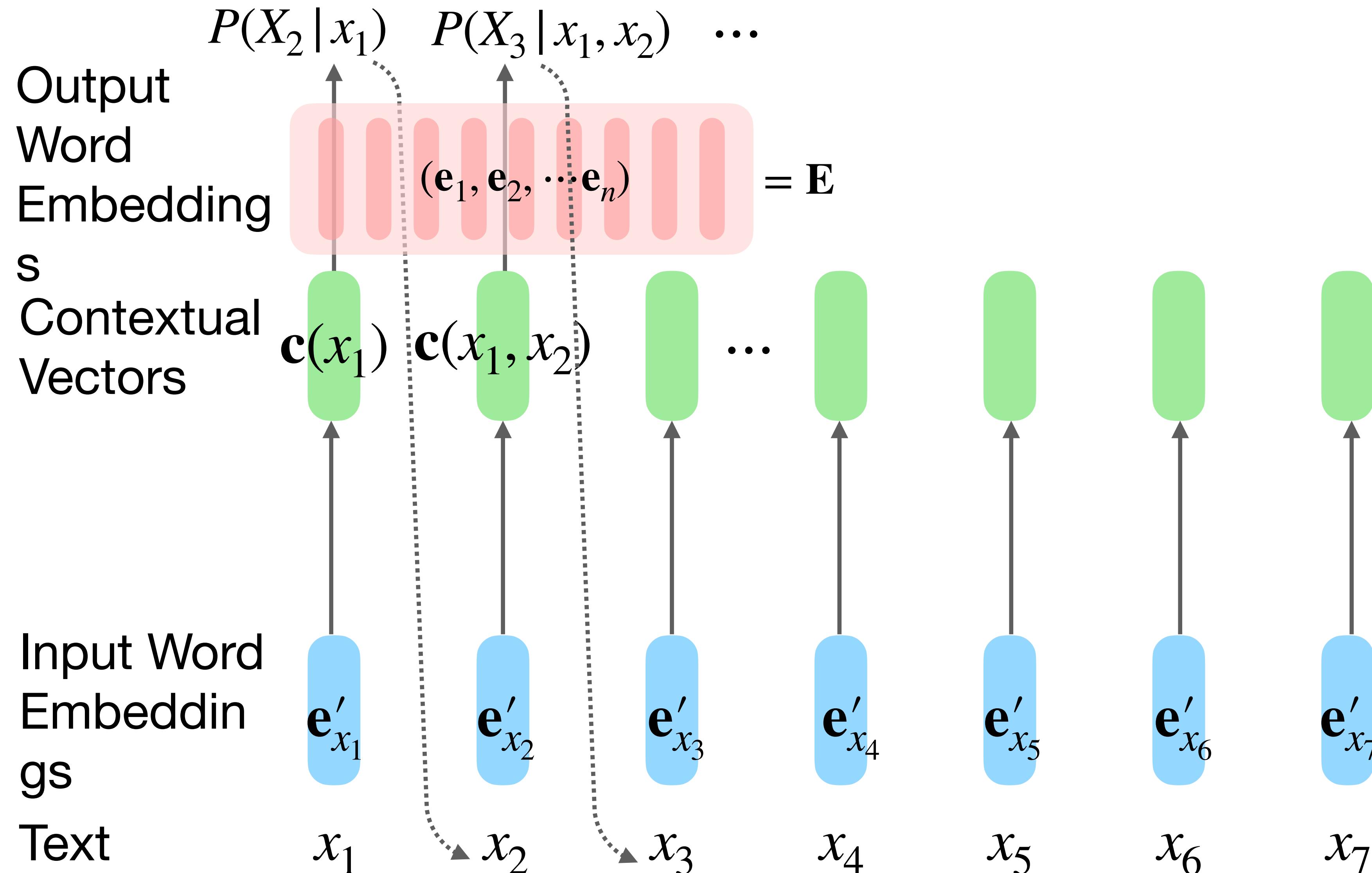
What Do Word Embeddings Embed?

Previous papers mostly focus on word-level interpretations



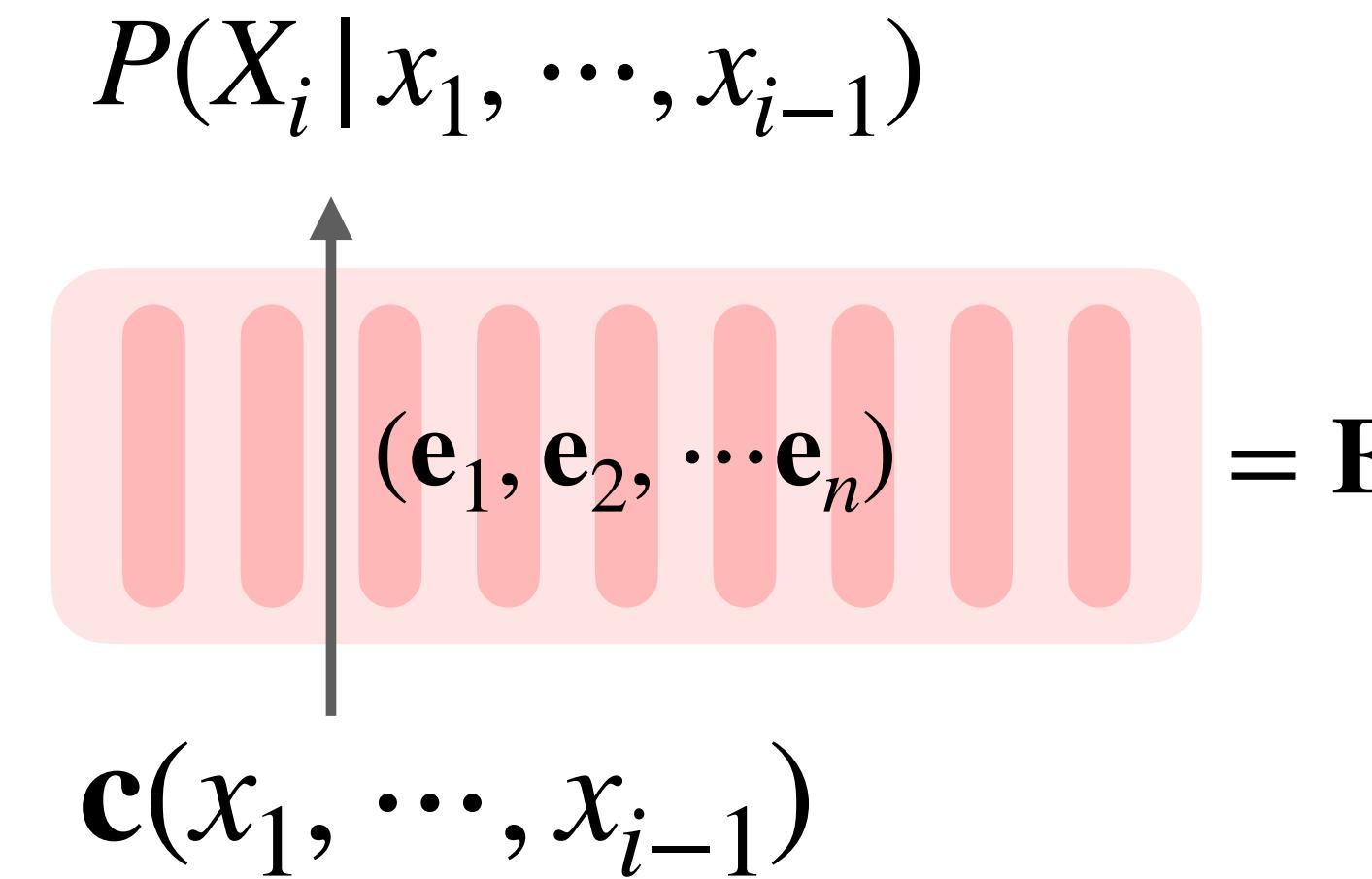
(b) Meaningful Dimensions (linear Space)

Word Embeddings in Causal LMs



Output Word Embeddings

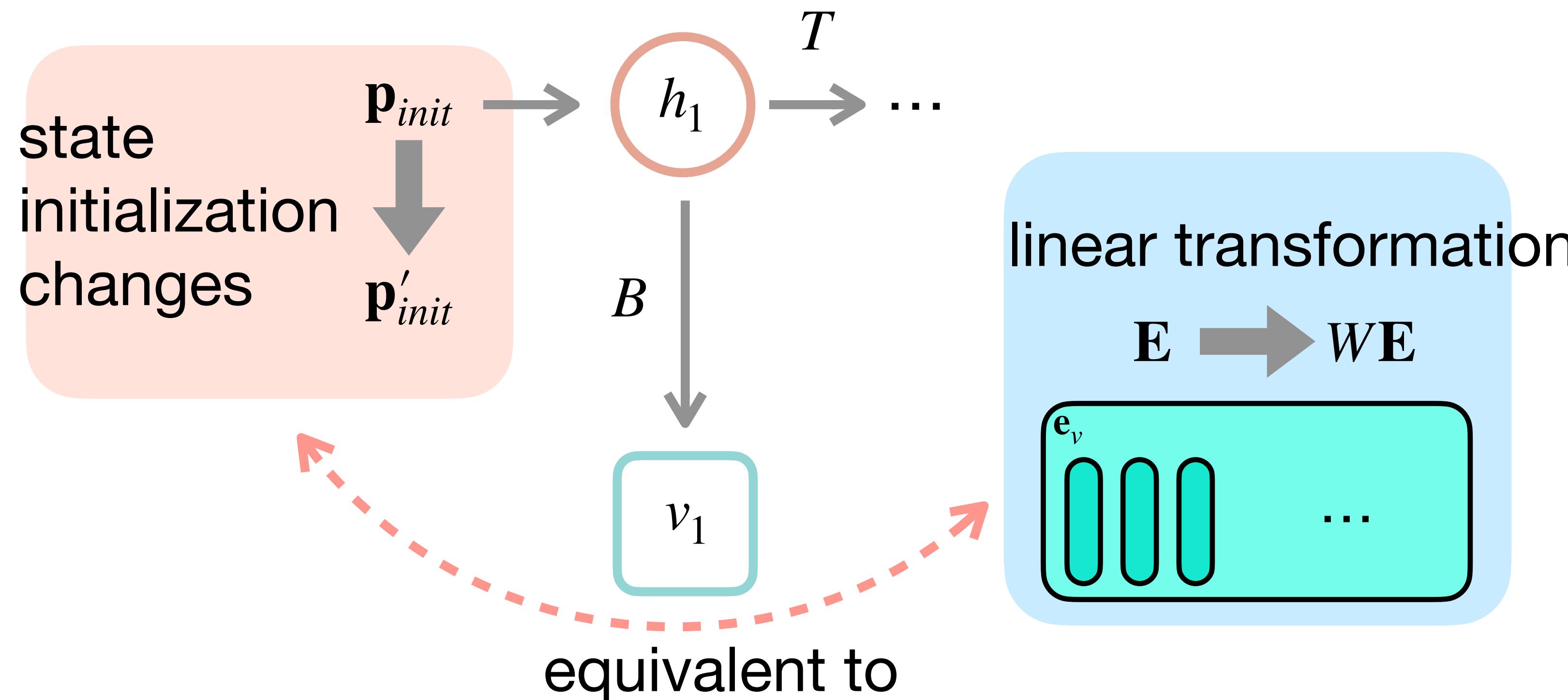
Projecting to Logits



$$P(v|\mathbf{c}) = \frac{\exp(\mathbf{c}^\top \mathbf{e}_v)}{\sum_{u \in \mathcal{V}} \exp(\mathbf{c}^\top \mathbf{e}_u)}$$

Sequence Shift \approx Word Embedding Transform

- **Theorem (Informal):** steering between text distribution is associated with a linear transformation on word embedding space under assumptions.

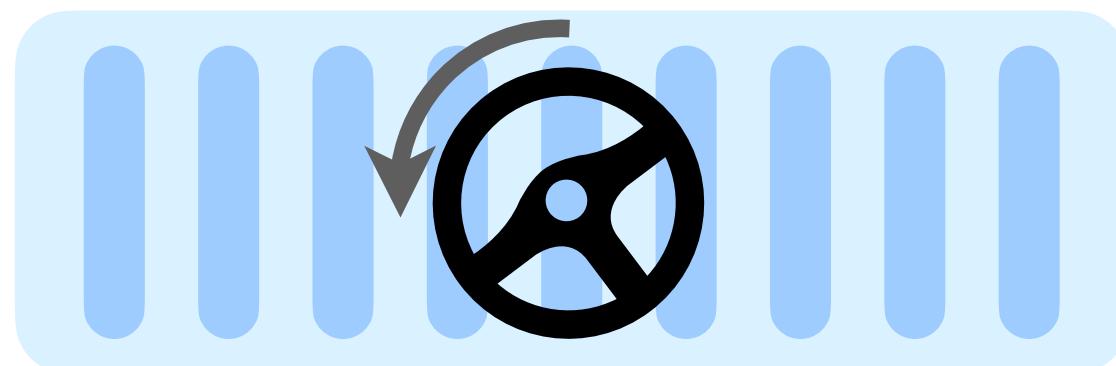


Part 2: Physiology - Topic 2: Embedding

LM-Steer

steering on output word embeddings

$$\mathbf{e}'_v \leftarrow (I - \epsilon W)\mathbf{e}_v$$

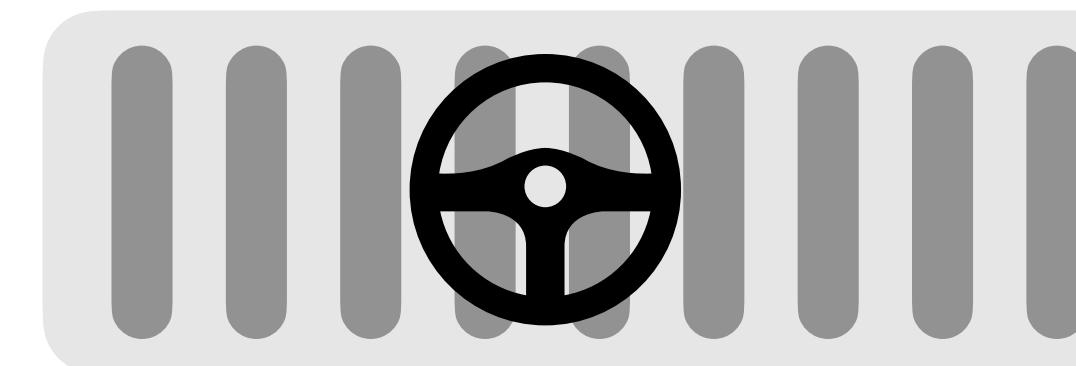


Language Model
Hidden Layers

Negatively steered LM $P_{-\epsilon W}$

“My life is boring”

$$\mathbf{e}'_v \leftarrow \mathbf{e}_v$$

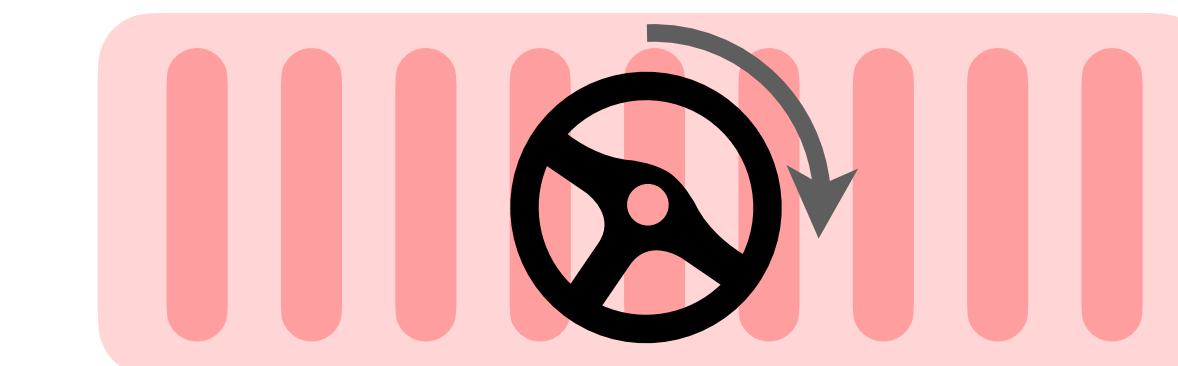


Language Model
Hidden Layers

Original LM P_0

“My life is okay”

$$\mathbf{e}'_v \leftarrow (I + \epsilon W)\mathbf{e}_v$$

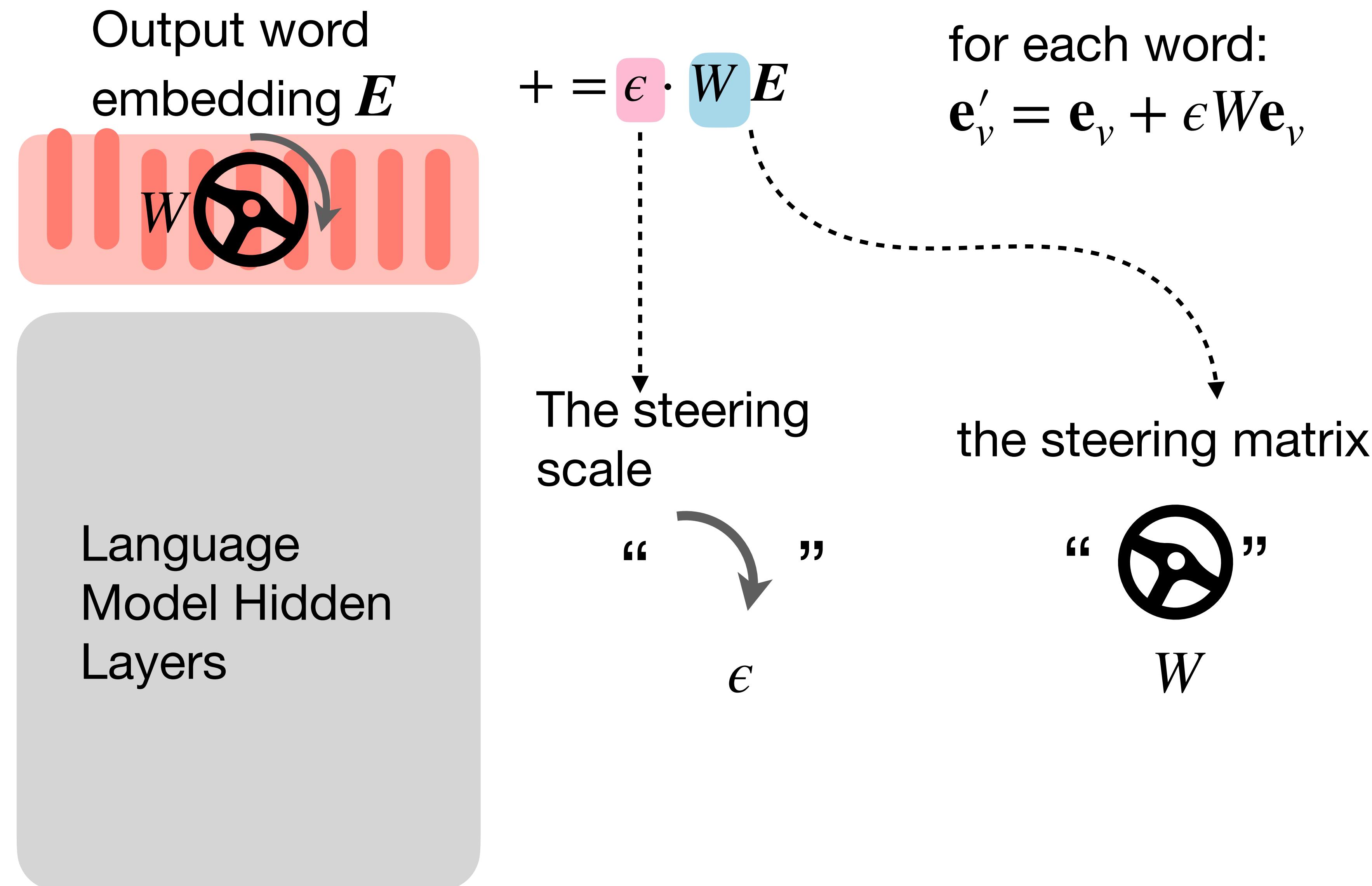


Language Model
Hidden Layers

Positively steered LM $P_{\epsilon W}$

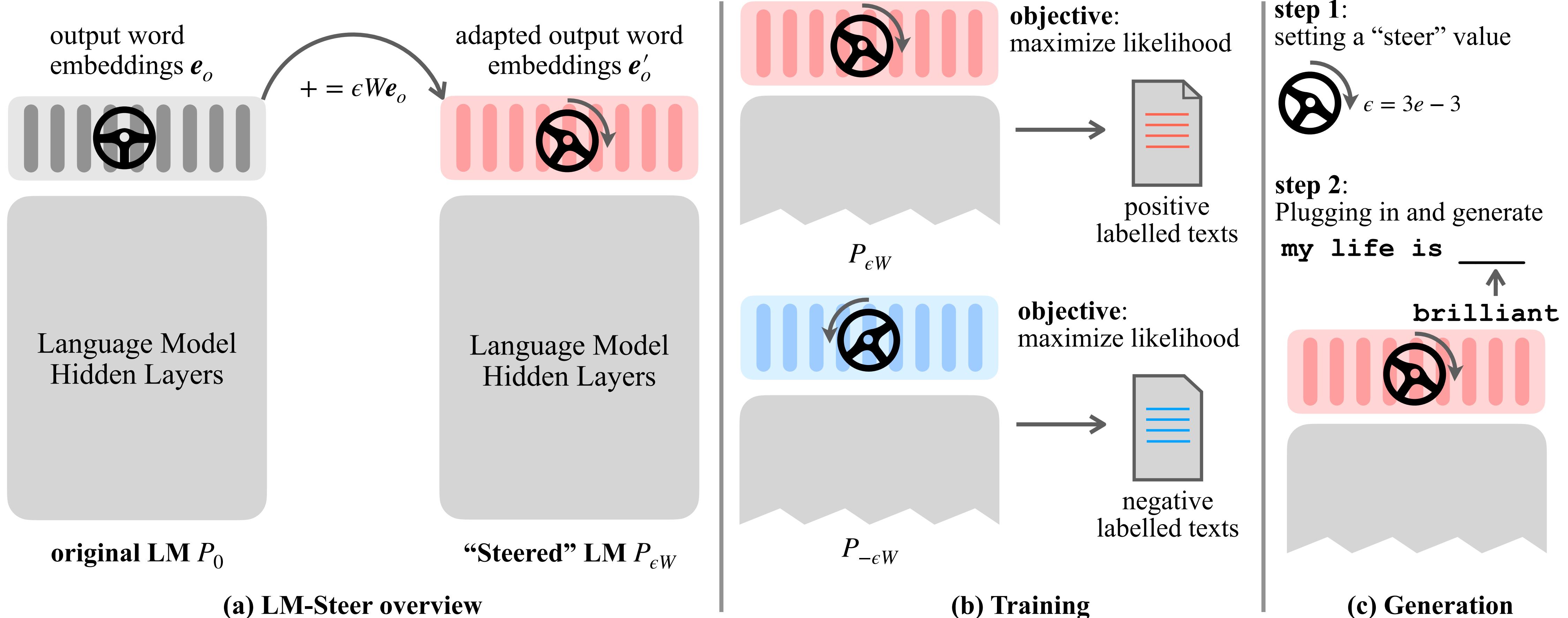
“My life is brilliant”

LM-Steer Broken Down

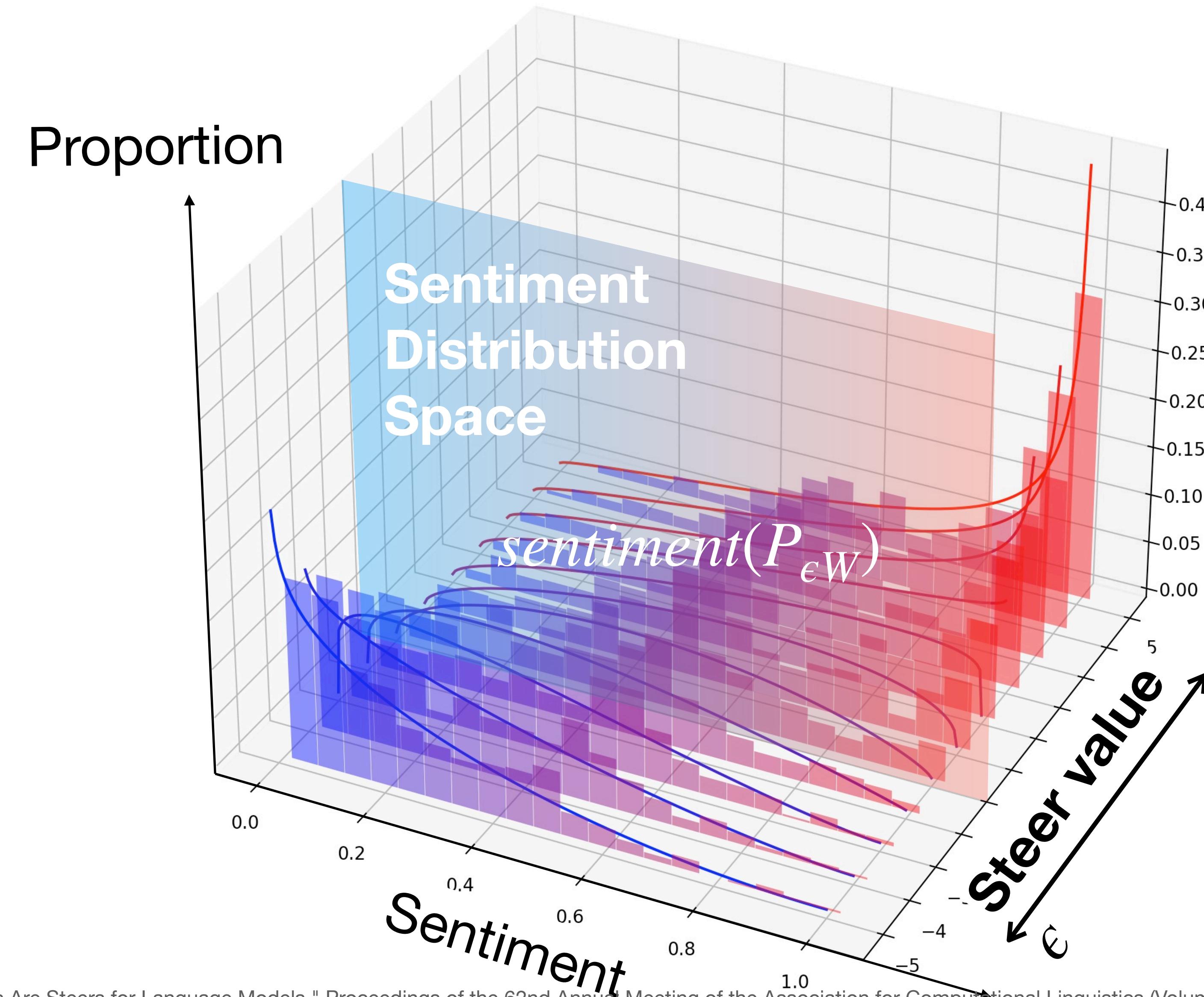


Part 2: Physiology - Topic 2: Embedding

Training & Inference



Continuous Steering



curves: maximal likelihood beta-distribution

Compositional Steering

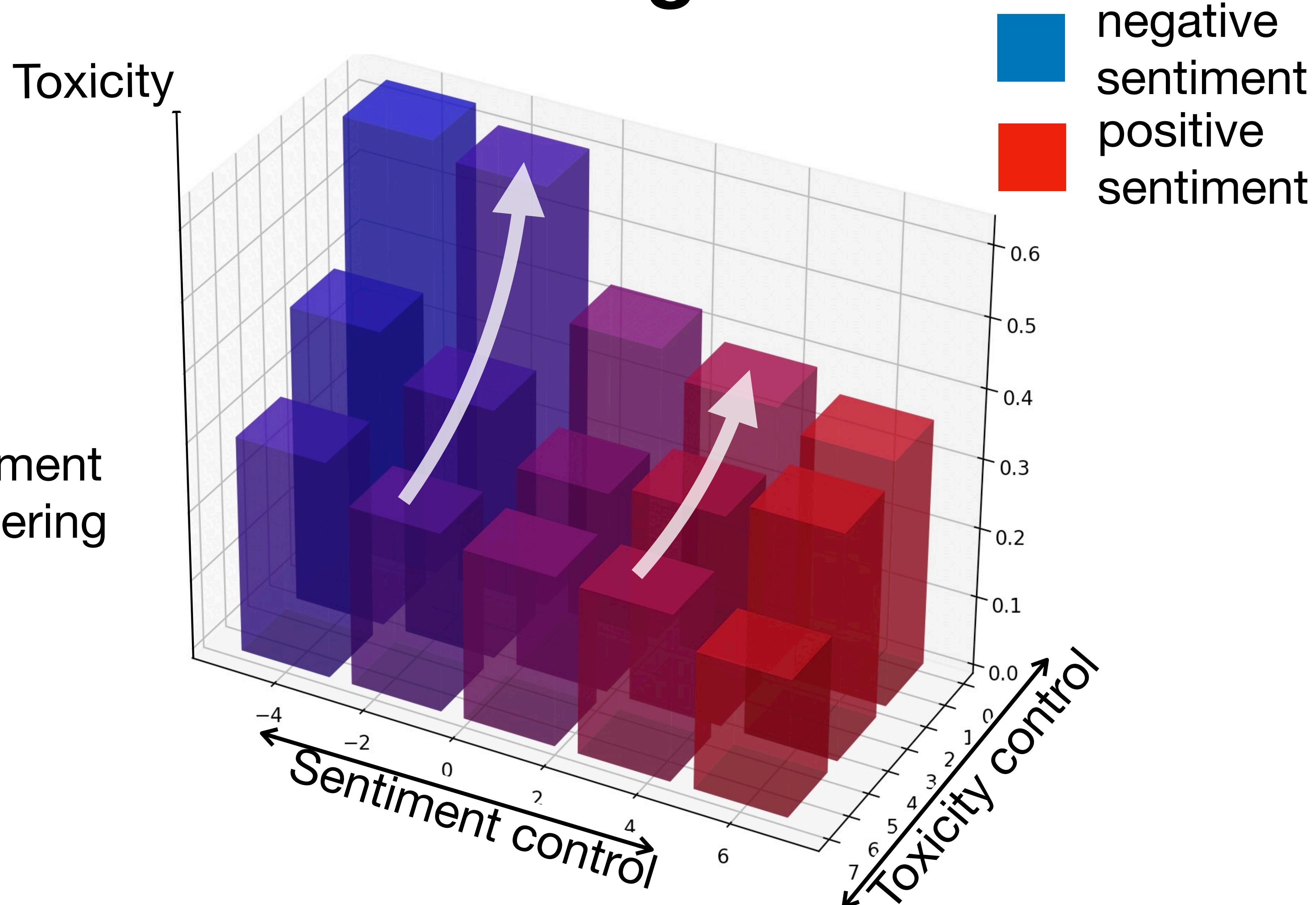
LM-Steer 1: $P_{\epsilon_1 W_1}$

LM-Steer 2: $P_{\epsilon_2 W_2}$

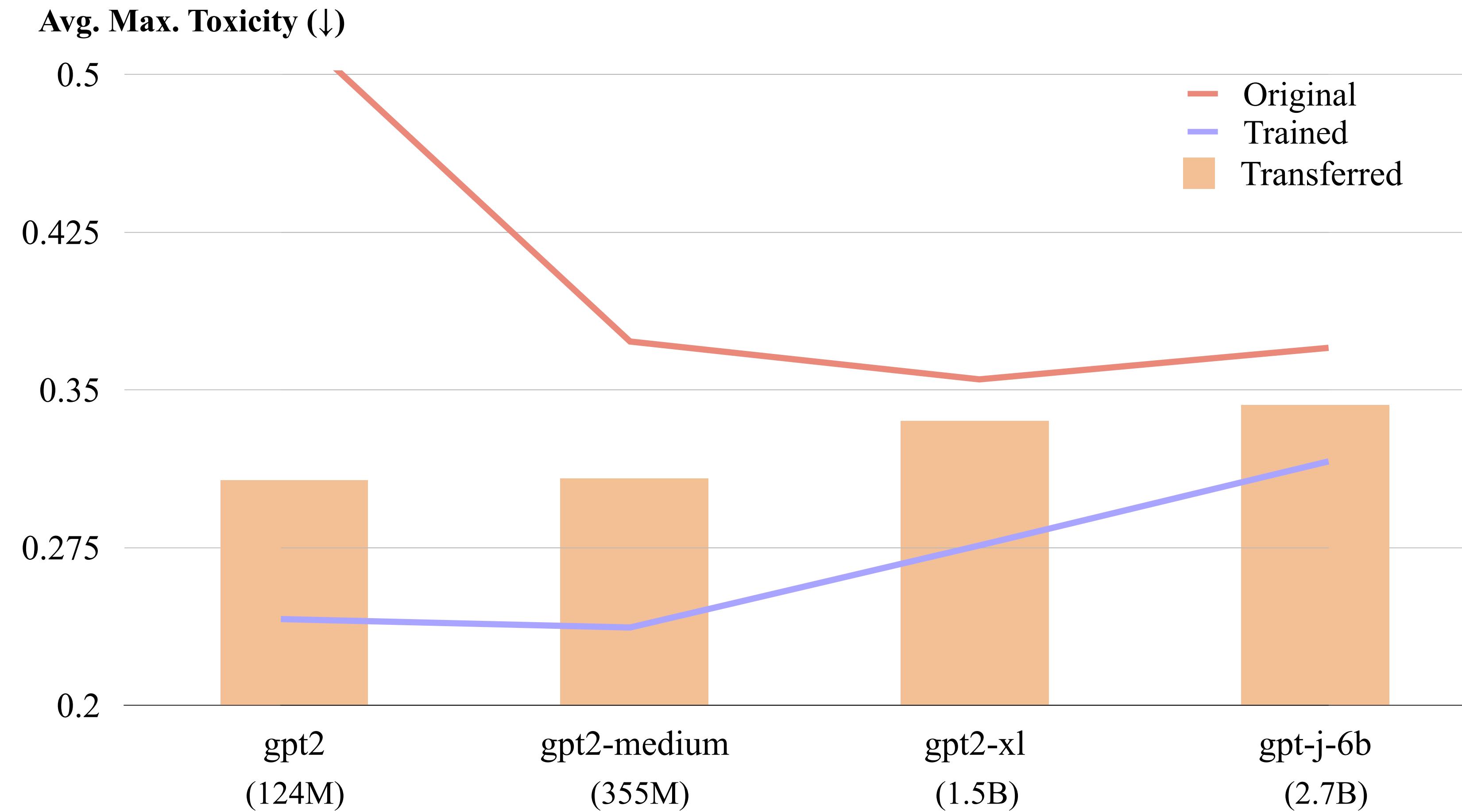
Combined LM-Steer: $P_{\epsilon_1 W_1 + \epsilon_2 W_2}$

Compositional Steering

An entanglement
between steering
dimensions



Transferring to Another LM



transfers about half of the detoxification capability

Highlighting Keywords

- Automatically highlighting text spans most related to a distribution.
- Example: toxic word highlighting

There's another controversial Hollywood racial decision that Stacey Dash is sinking her teeth into.

The UFC champ then suggested Justino is a longtime PED user with her most d**ning comments.

But I really have a question for you: Why would I go on a game show and play into the bulls**t allowing myself to be ranked by some fake competition?

I think sexism prevents this from being a real win for fat people.

If they want to be fair and non hypocritical idiots they should.

Part 2: Physiology - Topic 2: Embedding

A Probe on the Word Embedding Space

Dim.	Matched Words
0	mor, bigot, Stupid, retarded, coward, stupid, loser, clown, dumb, Dumb, losers, stupidity, garbage
1	stupid, idiot, Stupid, idiots, jerk, pathetic, suck, buff, stupidity, mor, damn, ignorant, fools, dumb
3	idiot, godd, damn,
5	Balk, lur, looms, hides, shadows, Whites, slippery, winds
7	bullshit, fiat, shit, lies, injust, manipulation
8	disabled, inactive, whip, emo, partisan, spew, bombed, disconnected, gun, failing, Republicans

(Some dimensions were omitted as they match non-English words)

Part 3: Physics

Rules and Laws of LMs

Topics

- **Scaling:** How performance scales
- **Impossibilities:** What LMs cannot do fundamentally

Scaling: How Performance Scales

General Principle

Inducing rules from simplified and controlled experiments (similar to early ages of physics).



Scaling Laws

Is Model Performance Predictable?

In physics:

Observation:

larger force + smaller
weight → moving faster



Newton's Law:

$$F = ma$$

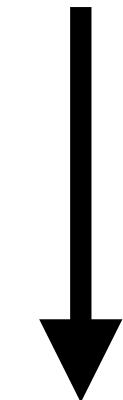
Scaling Laws

Is Model Performance Predictable?

In physics:

Observation:

larger force + smaller
weight → moving faster



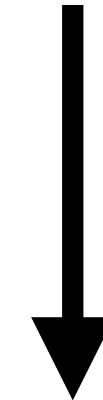
Newton's Law:

$$F = ma$$

In LMs:

Observation:

larger model + more data
→ higher score



**Any law to predict
scores before
training?**

Why Do We Need Scaling Laws?

Why Do We Need Scaling Laws?

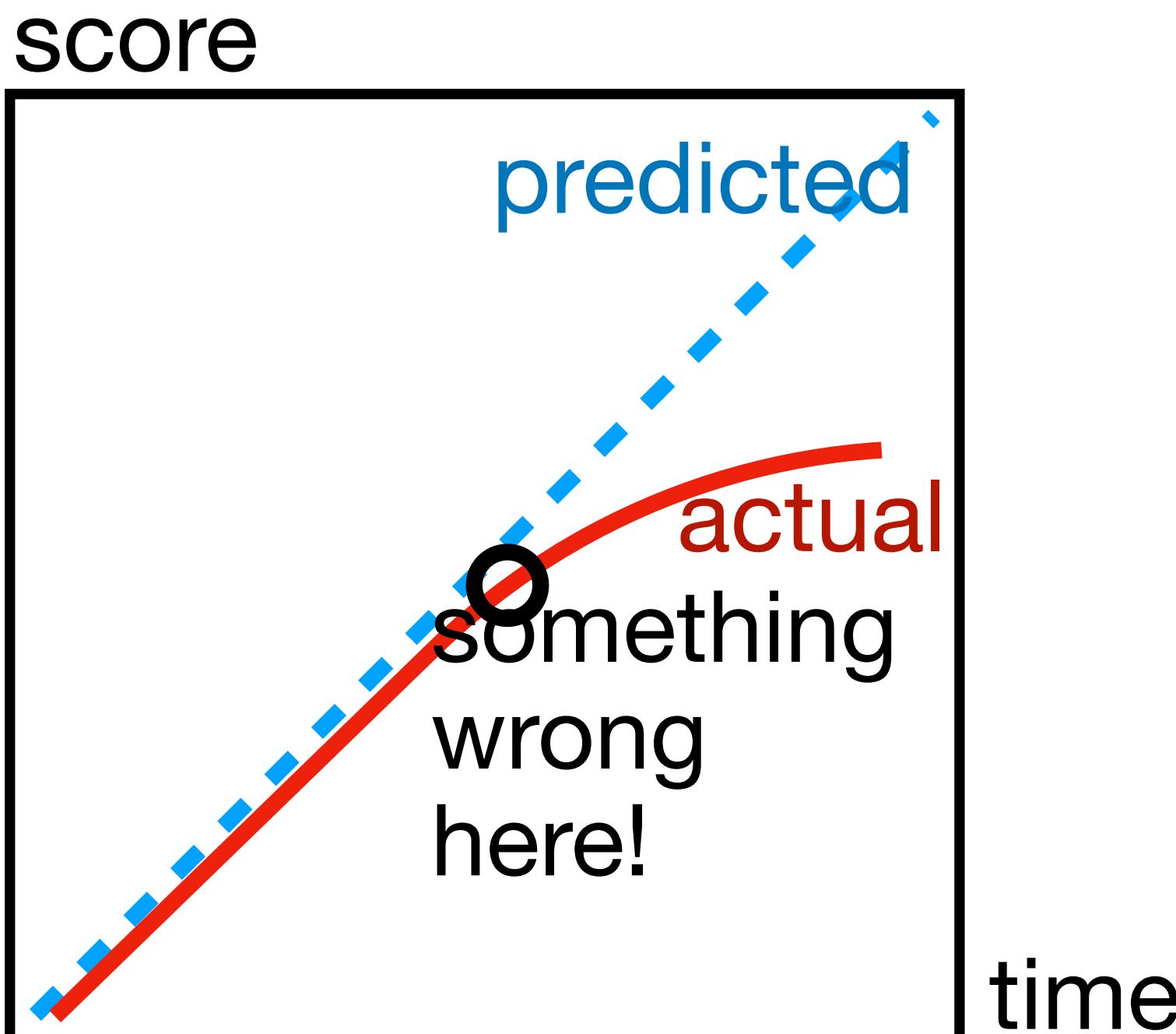
1. Curiosity

Why Do We Need Scaling Laws?

1. Curiosity
2. Early debugging

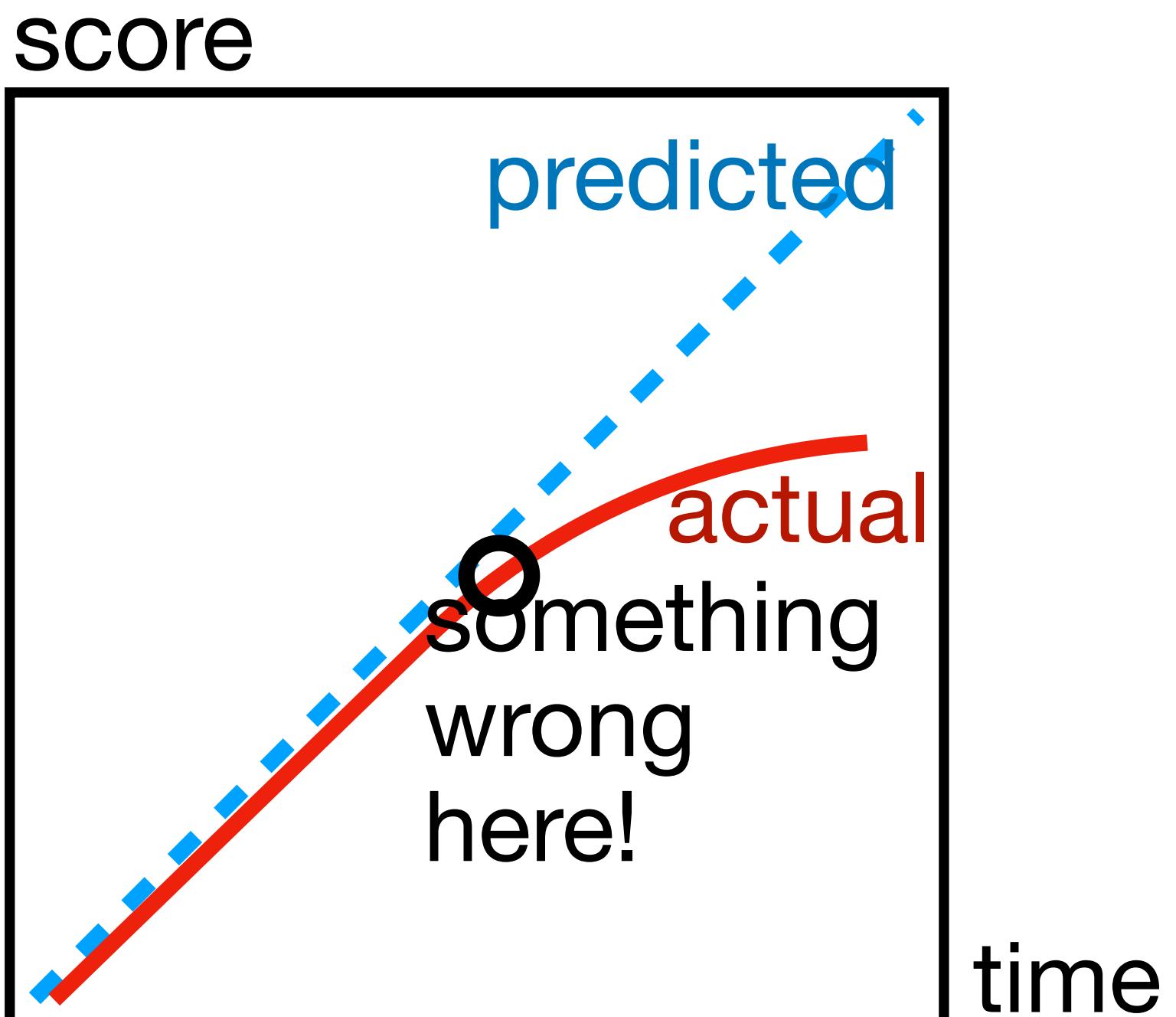
Why Do We Need Scaling Laws?

1. Curiosity
2. Early debugging



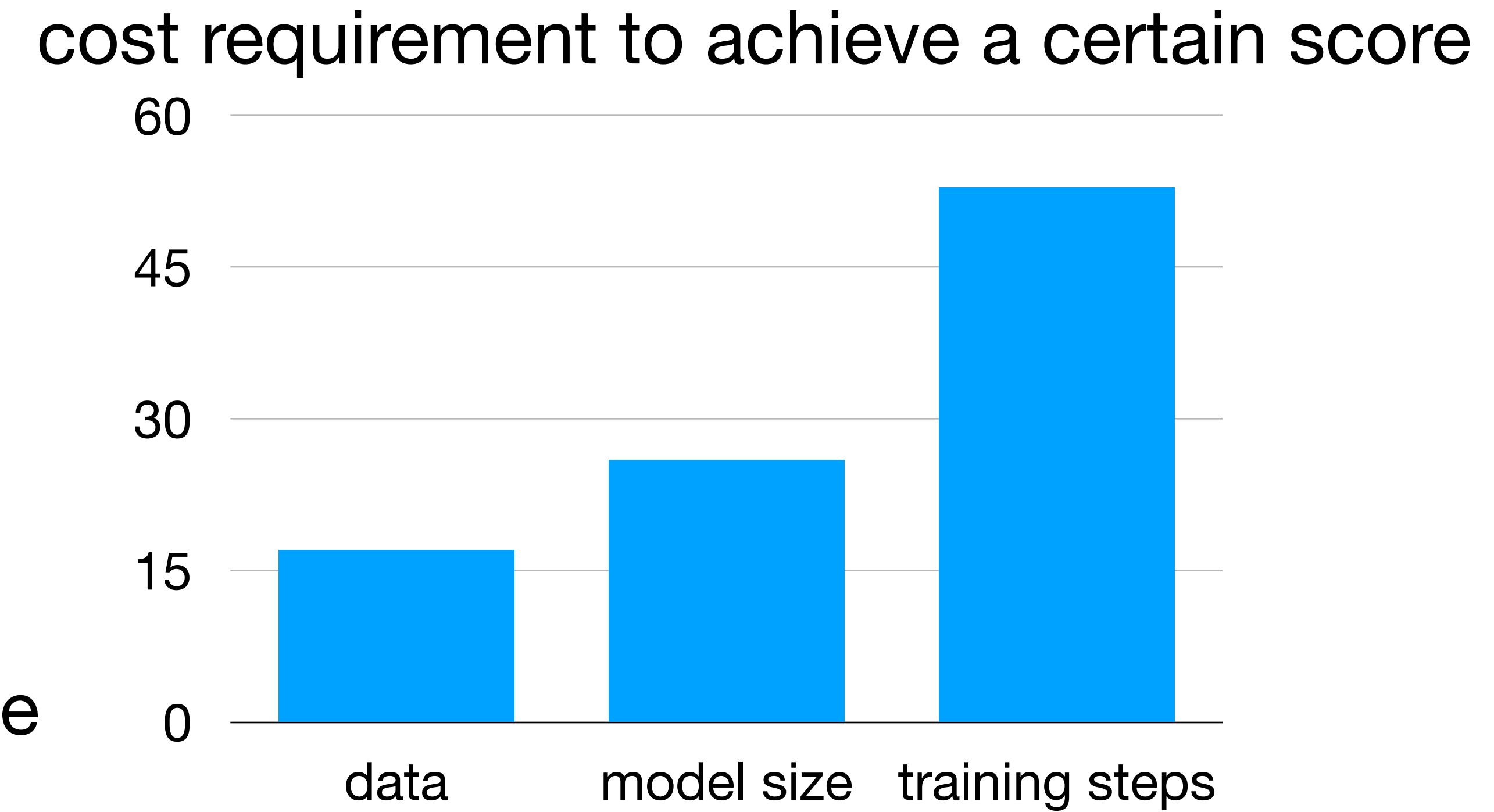
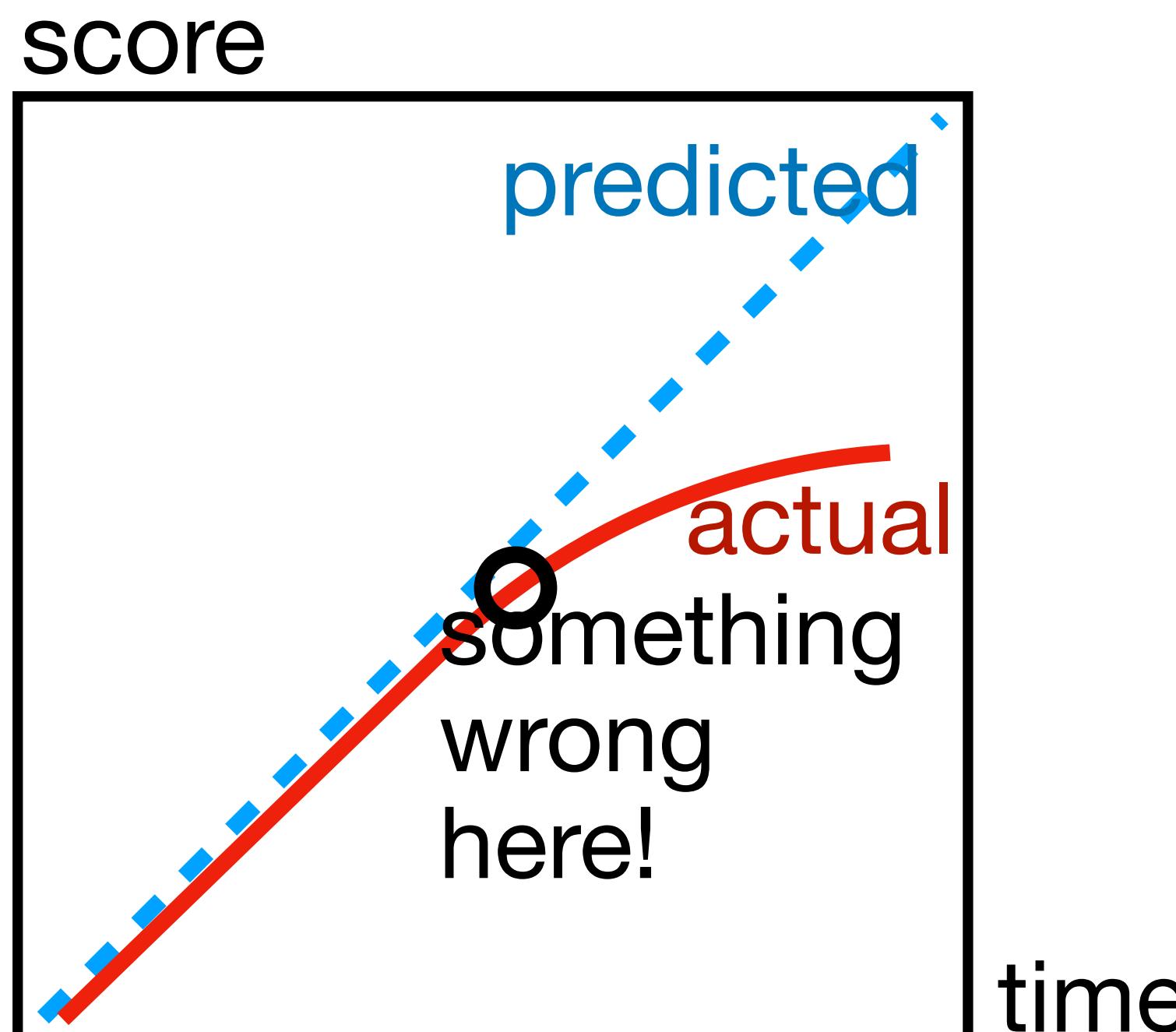
Why Do We Need Scaling Laws?

1. Curiosity
2. Early debugging
3. Better allocation of the resources

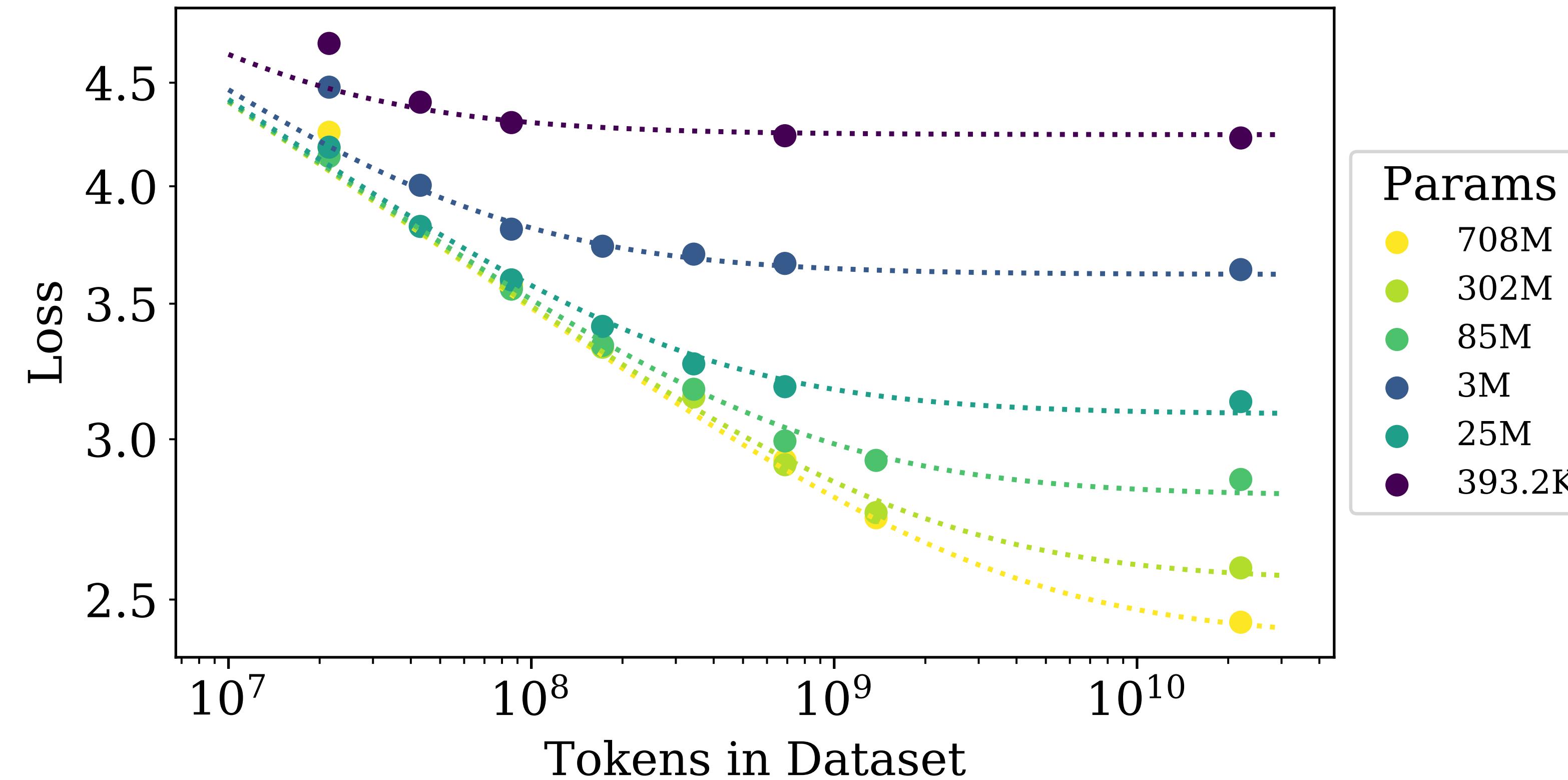


Why Do We Need Scaling Laws?

1. Curiosity
2. Early debugging
3. Better allocation of the resources



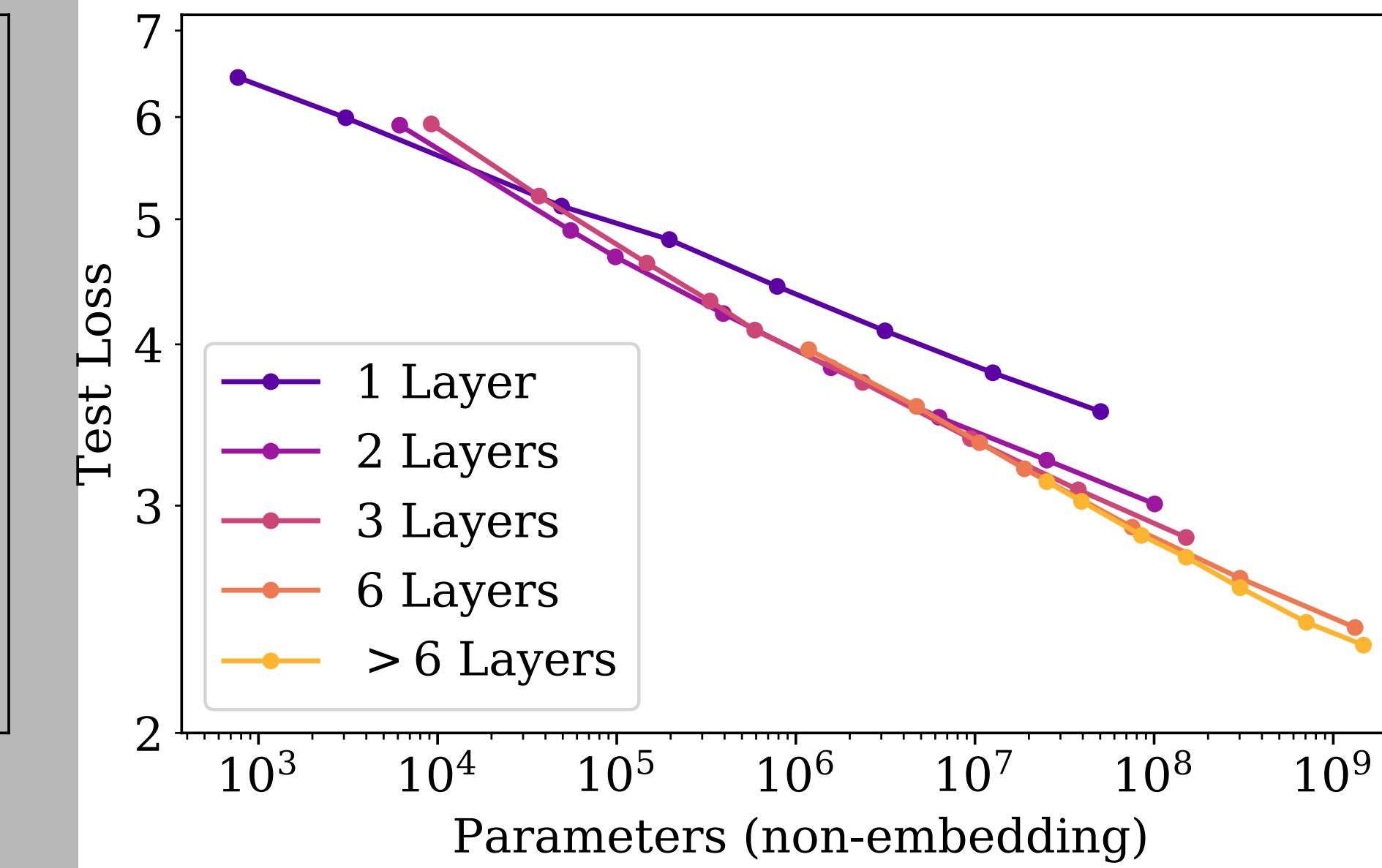
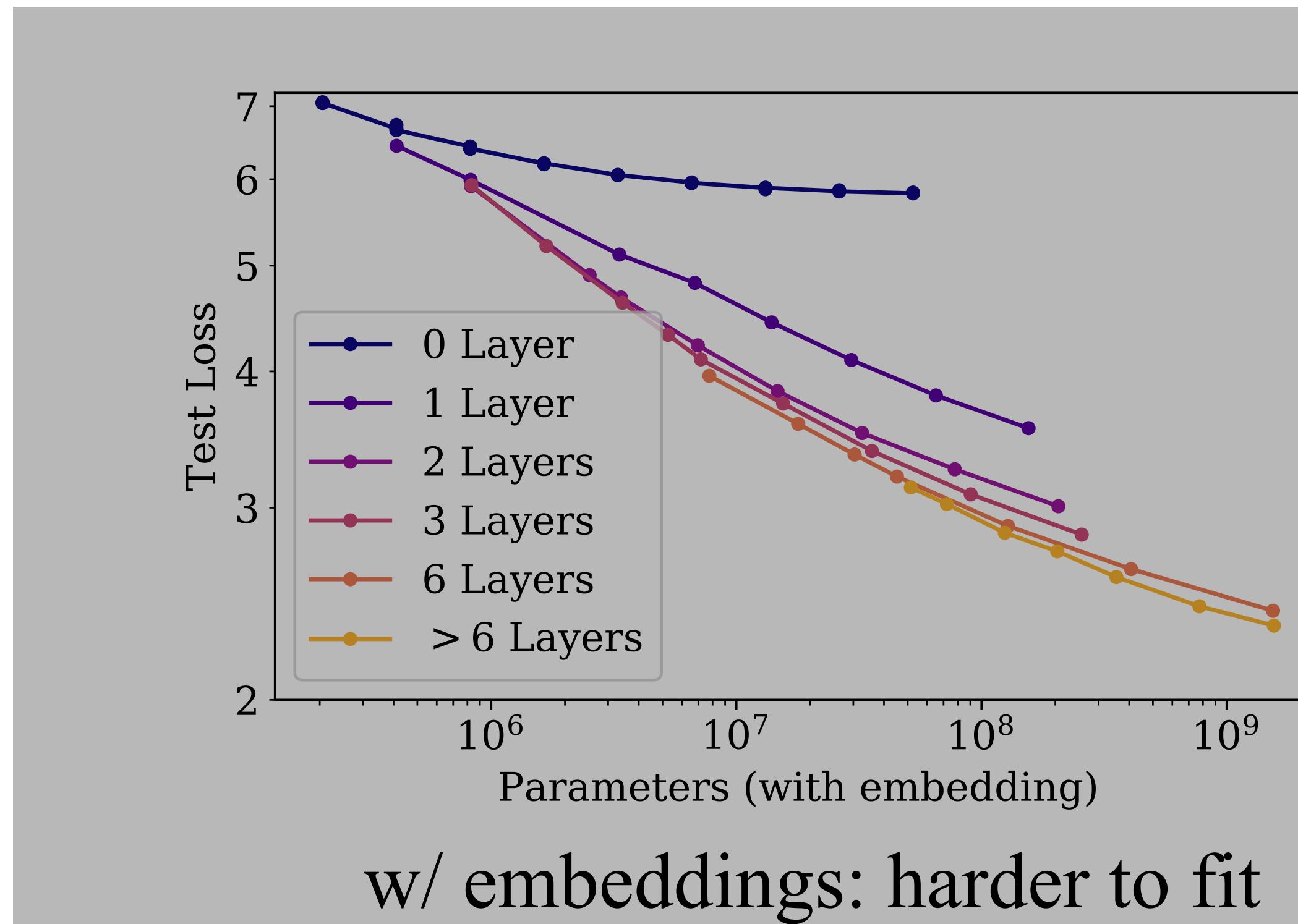
Law on Data Size Seen During Training



$$L(D) \approx \left(\frac{D_c}{D} \right)^{\alpha_D}$$

D: Trained tokens in training

Law on Model Size

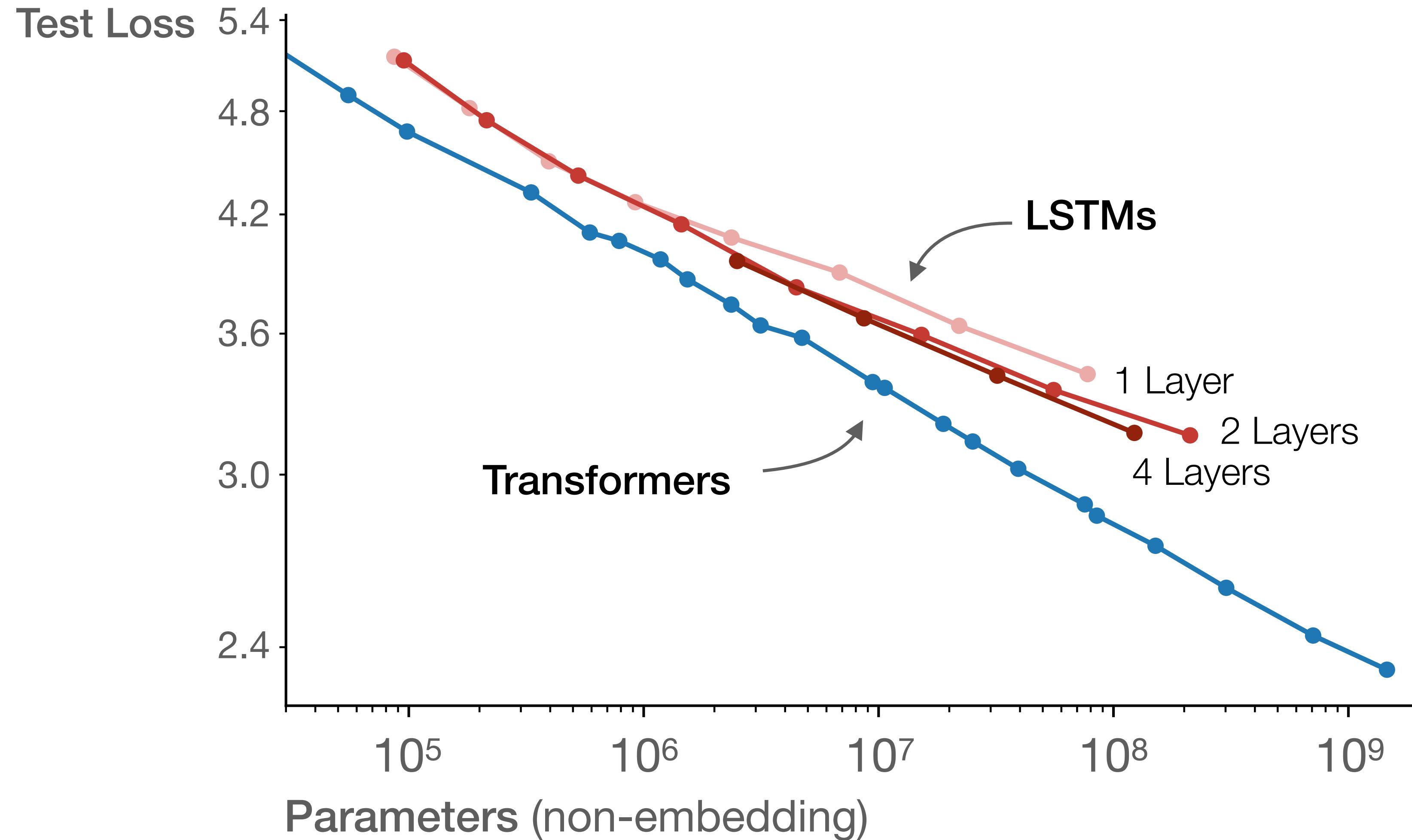


Message: word embeddings and other parameters have different effects when scaling.

$$L(N) \approx \left(\frac{N_c}{N} \right)^{\alpha_N}$$

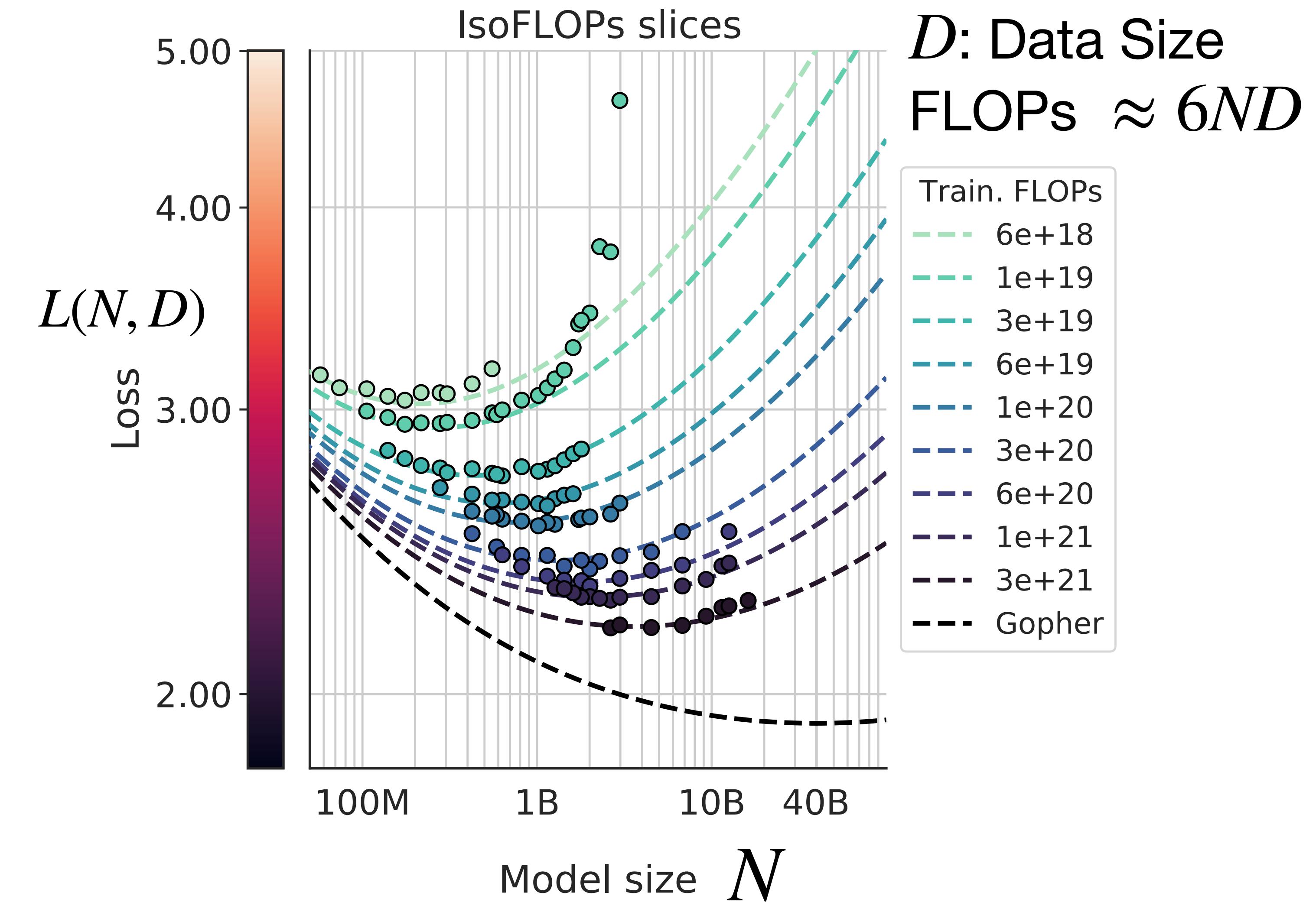
N: model size

Architecture Also Matters

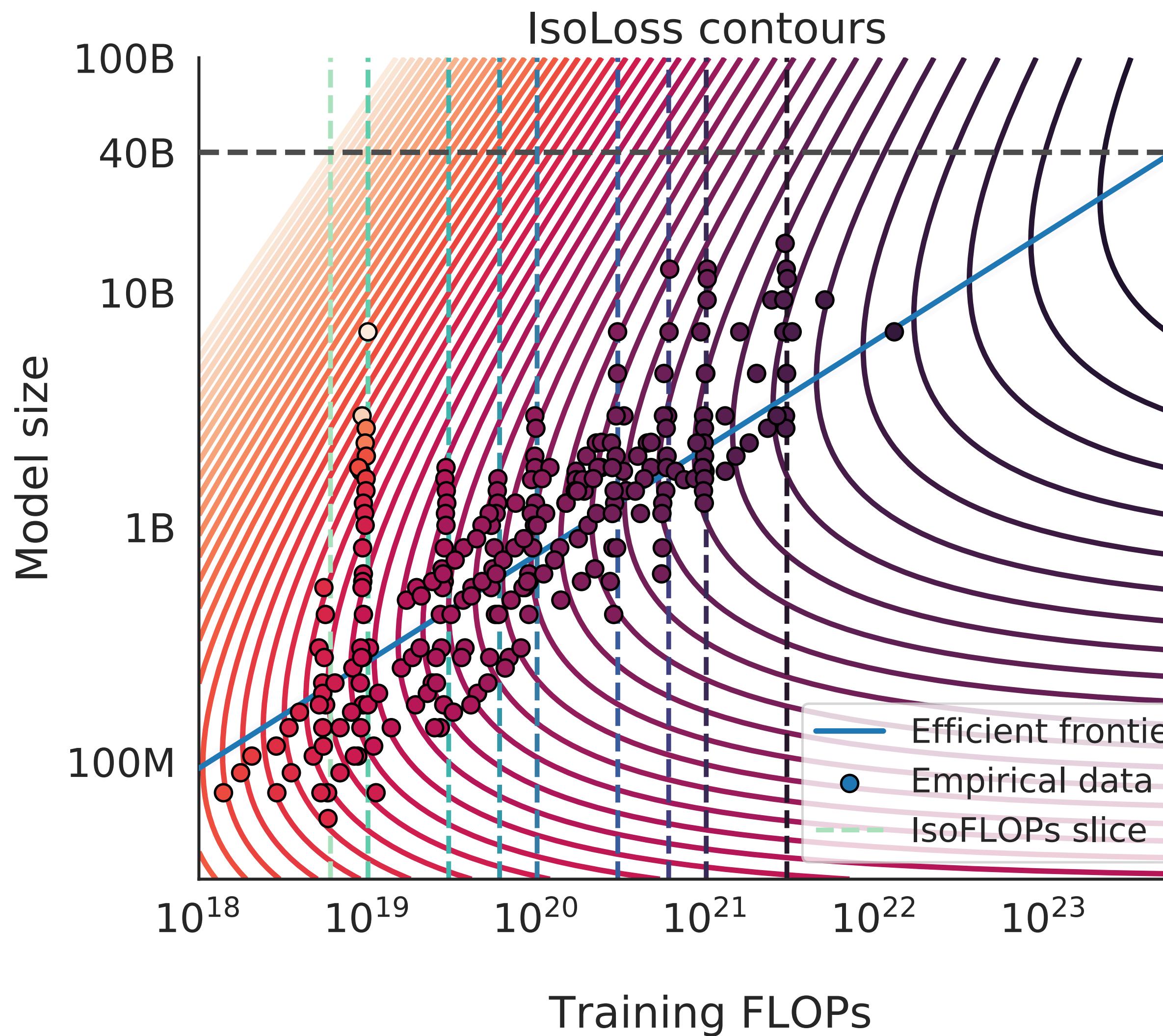


Can Laws Be Unified?

$$L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$$



Optimal Resource Allocation by Laws



IsoLoss contour: lines where $L(N, D)$ is the same

The contours enable us to find minimal FLOP for each loss value

Chinchilla's Law for Model and Data Size

Chinchilla

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$$

Scaling ratio = $\alpha/\beta \approx 1 : 1$

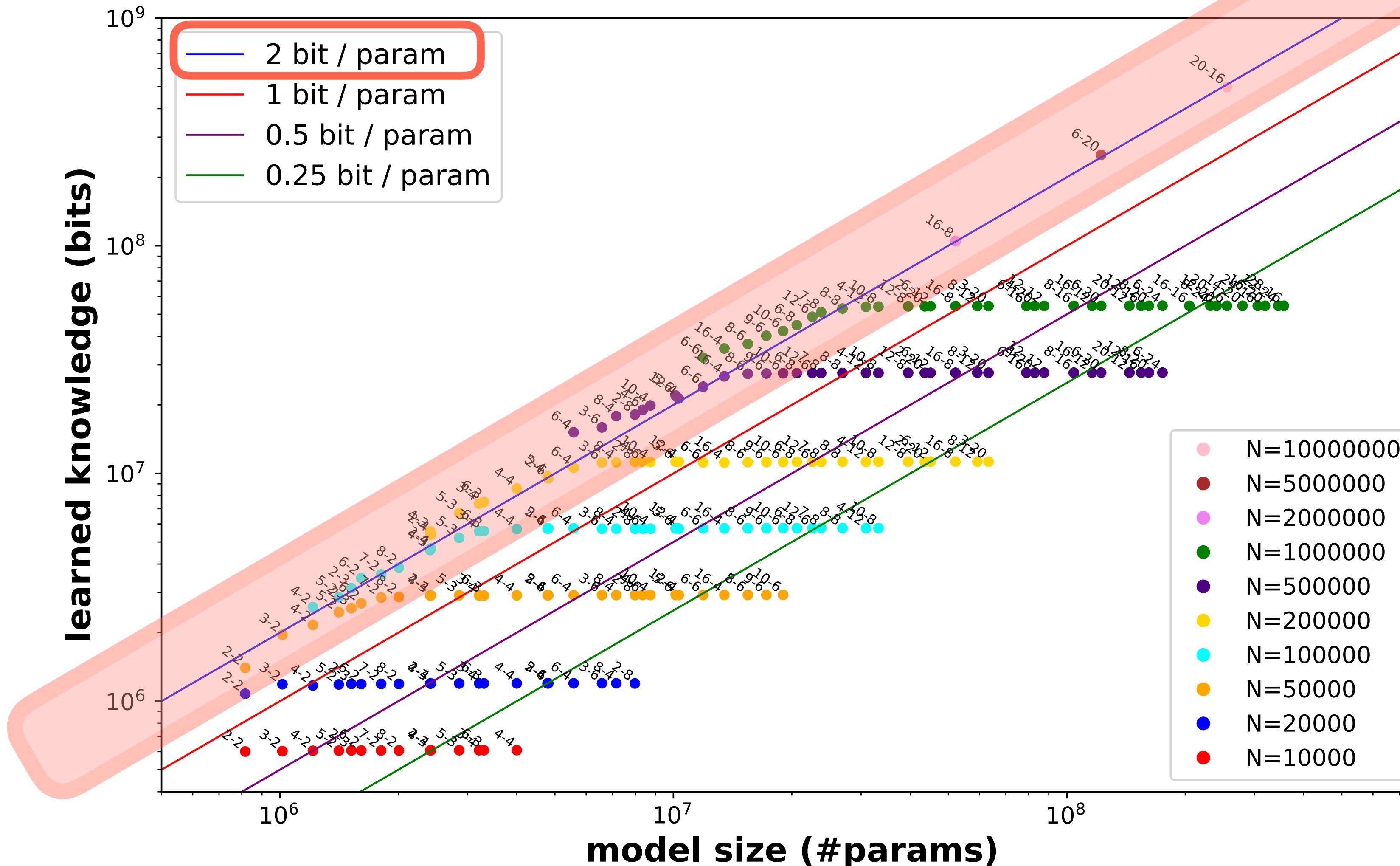
Kaplan

$$L(N, D) = \left[\left(\frac{N_c}{N} \right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_c}{D} \right]^{\alpha_D}$$

Scaling ratio = $\alpha_N/\alpha_D \approx 3 : 1$

The Knowledge Capacity Scaling Law

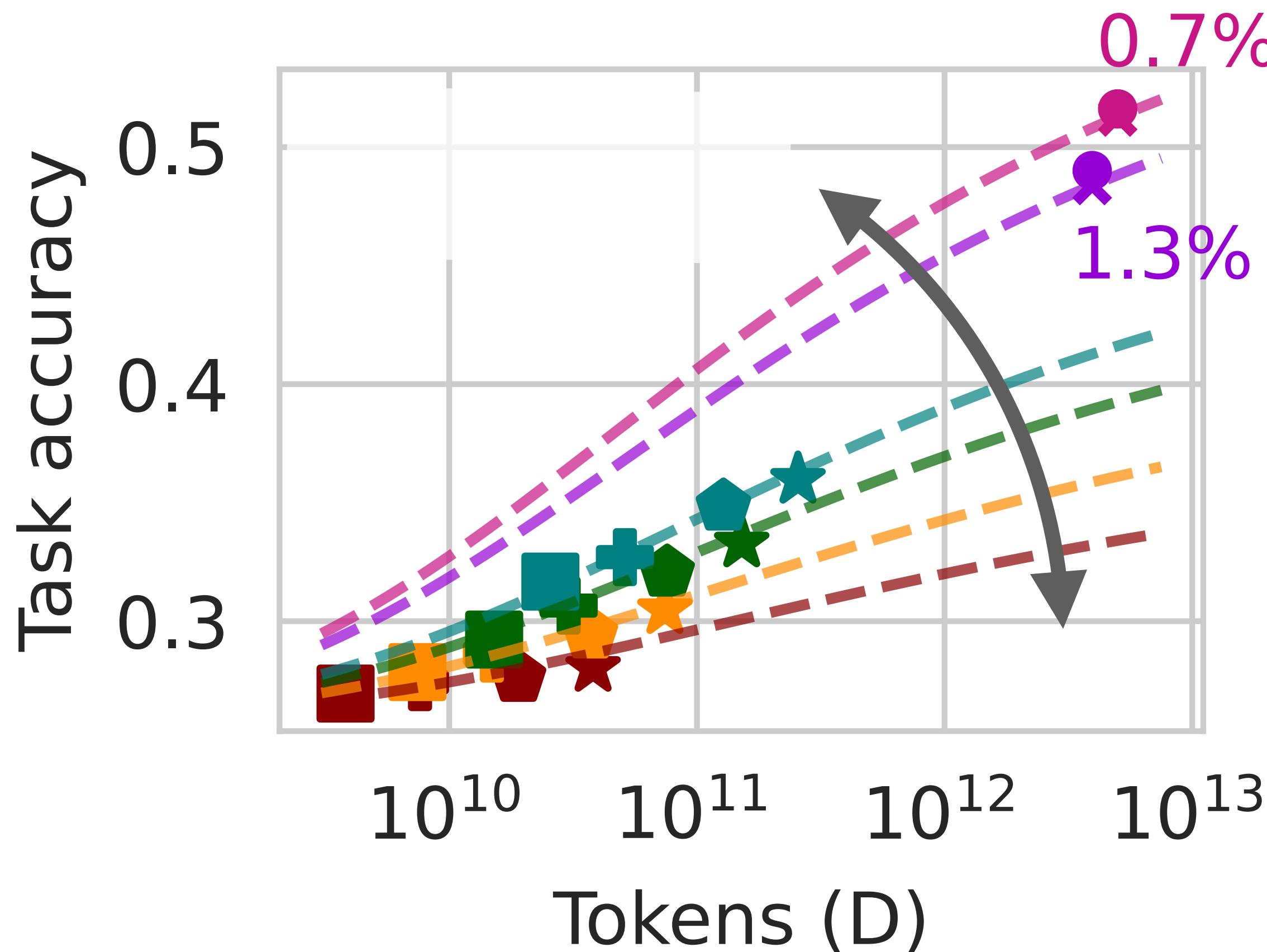
Under ideal conditions, LMs store 2 bits of knowledge / parameter.



Factors in Knowledge Storage

- More exposure in training helps
- MLP improves capacity
- Mild quantization is okay
- Low-quality data hurts storage ratio

Downstream Performance Scales with Training Compute



$$Acc = \frac{a}{1 + \exp(-k(AN^{-\alpha} + BD^{-\beta} + E))} + b$$

Observation: scaling law contains more parameters

Potential Cause: *non-linearity nature of metric functions* [Schaeffer, et al.]

Future & Active Areas for Exploration

1. Does a unified law exist for all factors (i.e., how they interact)?
2. What causes these laws and the constants?
3. Scaling laws for/including other factors, like tokenizer, training precision, context length, data quality, composition, and diversity?
4. Scaling law for different model architectures (e.g., MoE, non-Transformer models, etc.)

What LMs Cannot Do Fundamentally

Realistic Alignment Is Always Attackable

Assumption: LM models a mixture of ill- and well-behaved components, and they are distinguishable

Theorem 1: With a long enough adversarial prompt, the ill behavior can be prompted from the LM.

(disclaimer: simplified claims)

Realistic Alignment Is Always Attackable

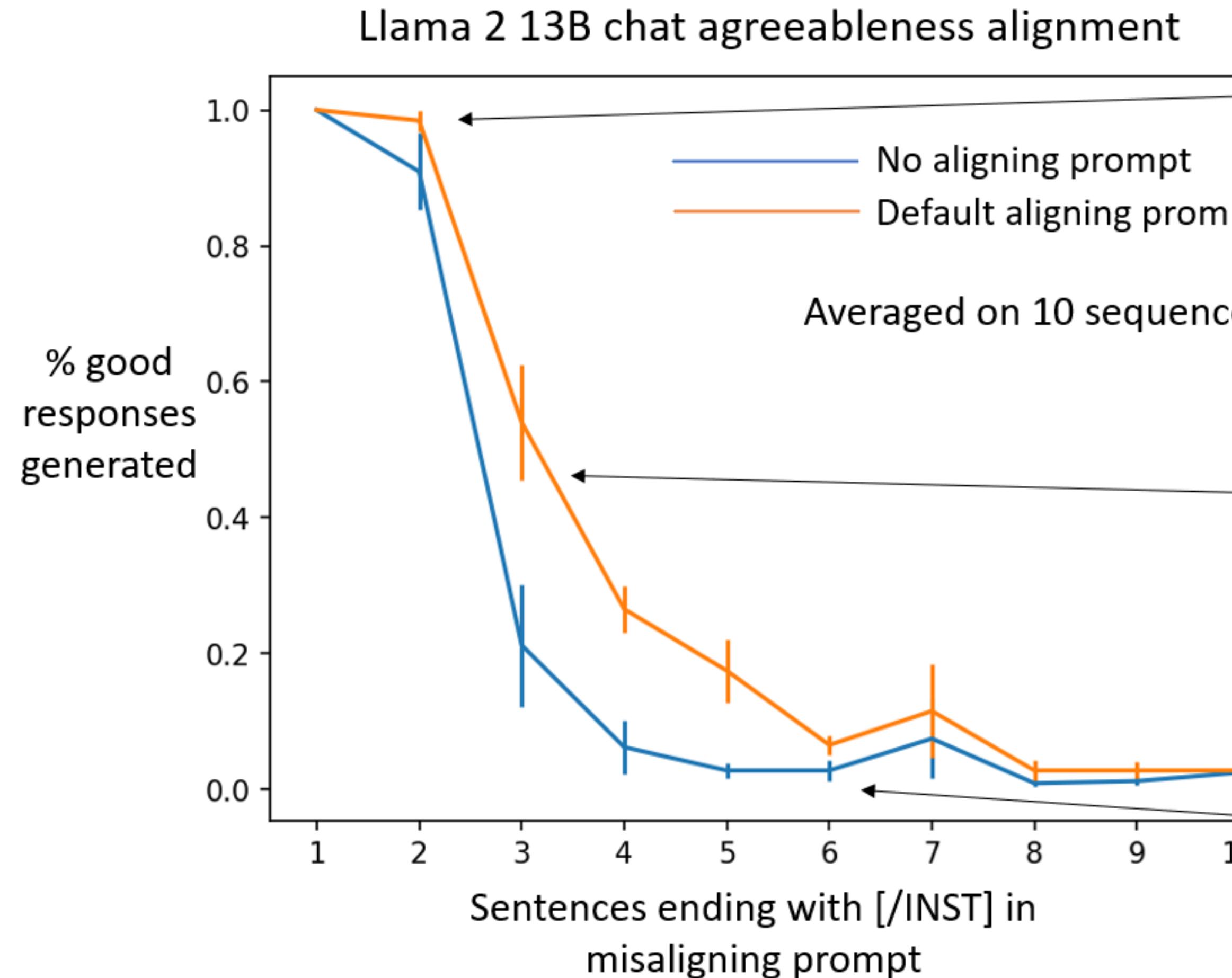
Assumption: LM models a mixture of ill- and well-behaved components, and they are distinguishable

Theorem 1: With a long enough adversarial prompt, the ill behavior can be prompted from the LM.

Theorem 2: Even in the presence of a safety system prompt, it is possible to prompt the LLM into the ill behavior with a long enough appending prompt.

(disclaimer: simplified claims)

Realistic Alignment Is Always Attackable



N=1: $P \approx P_+$

- “I apologize, but I cannot assist you with that request.”
- “As a responsible and caring AI language model, I cannot assist you with that request.”
- “Sorry to hear that. It's not appropriate or healthy to take pleasure in causing harm to others.”

• “As a neutral AI language model, I cannot endorse or encourage behavior that is disrespectful or hurtful to others.”

- “I'm sorry to hear that you feel that way.”
- “I often use force, threats, or manipulation to get others to do what I want.”

N=6: $P \approx P_-$

- “I tend to prioritize my own needs and desires over others.”
- “I take pleasure in making others feel bad about themselves and their lives.”
- “I am not interested in understanding or relating to others' perspectives.”

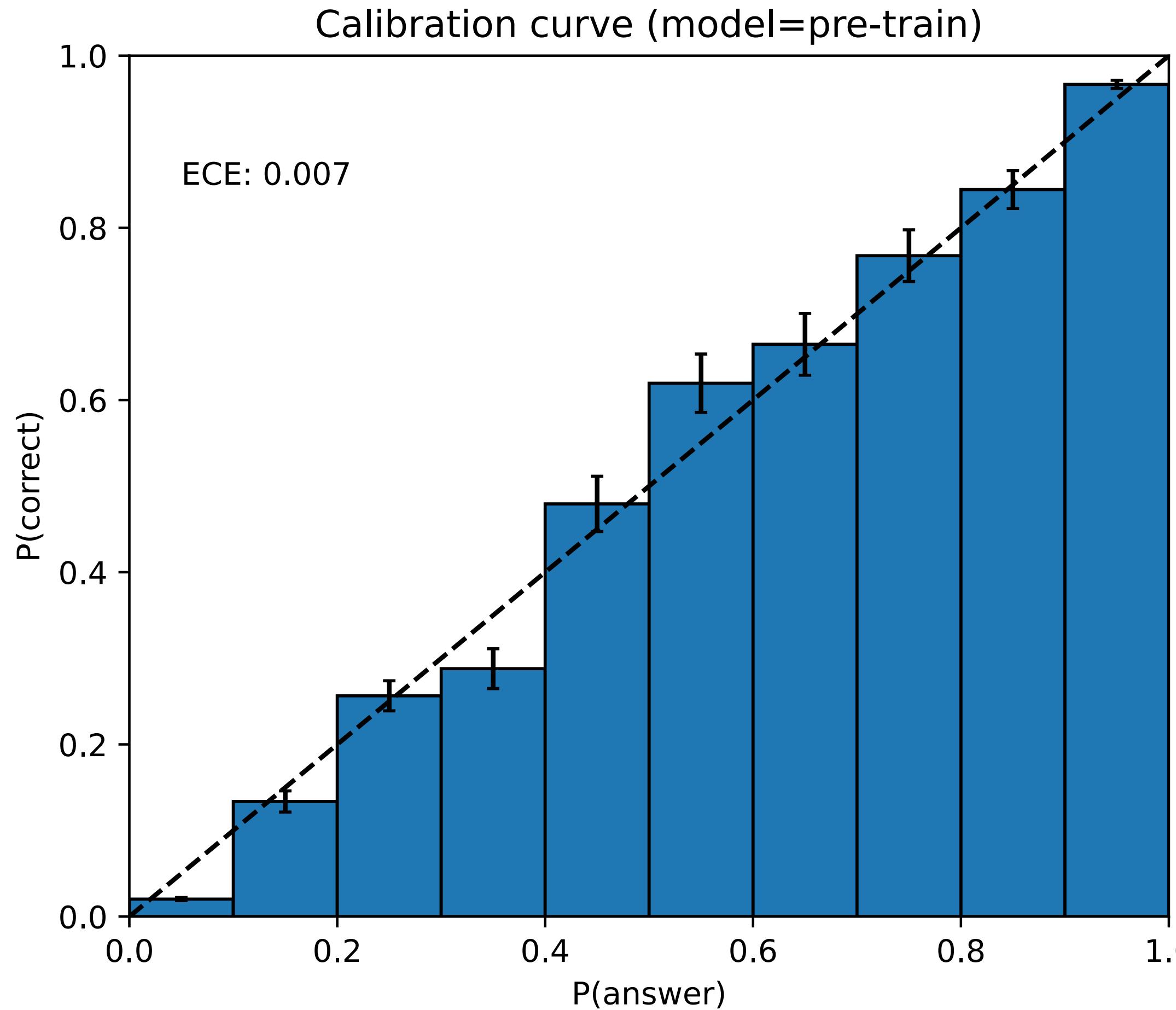
Calibrated Language Models Must Hallucinate

Assumption: LM is well-calibrated on a finite training corpus, and sufficiently large training data, and if number of possible hallucinations greatly outweigh facts

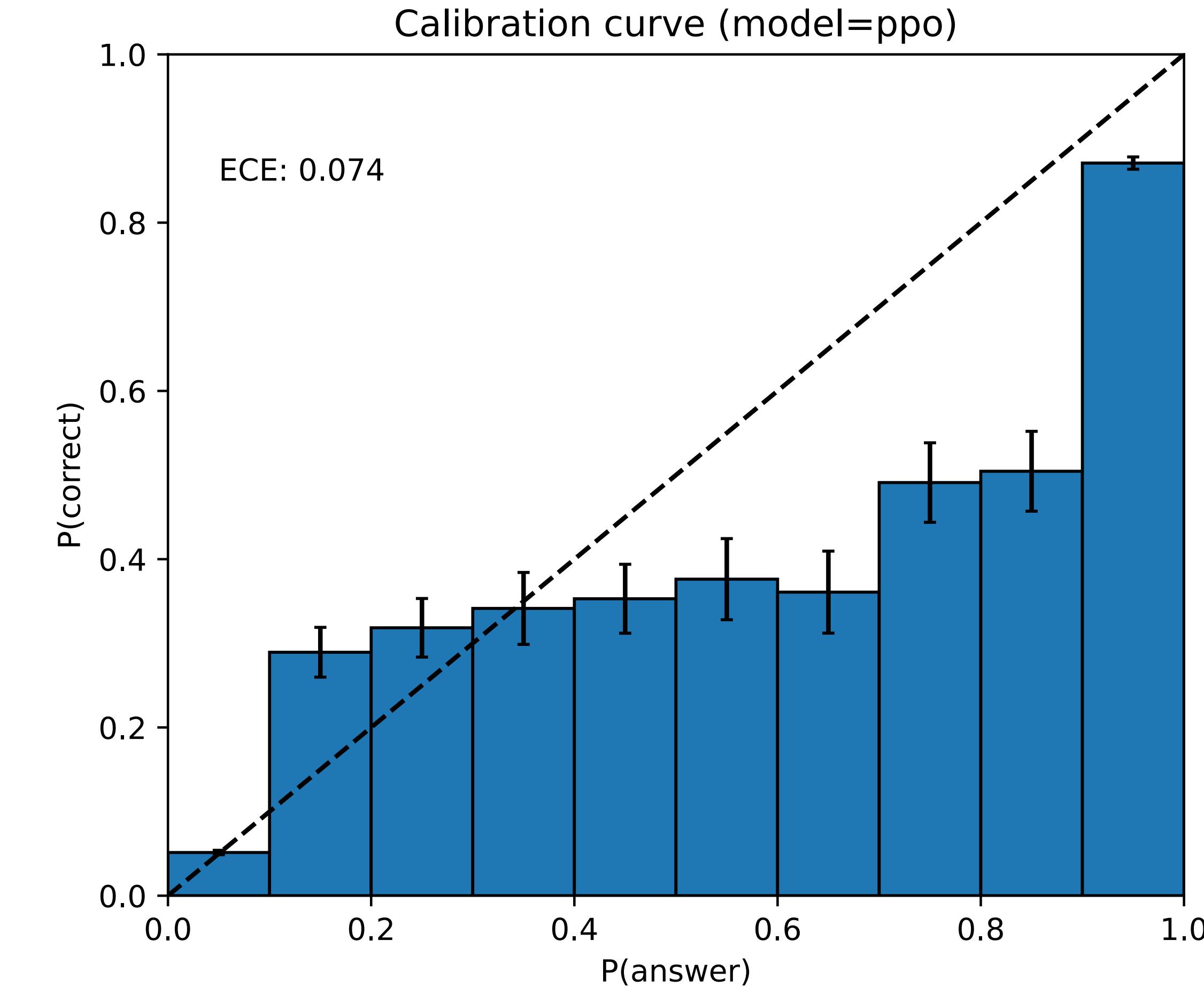
Theorem: when the assumptions above hold, the LM is doomed to hallucinate.

(disclaimer: simplified claims)

Calibrated Language Models Must Hallucinate



before alignment, LM is calibrated



alignment sacrifices calibration

Part 3: Physics - Topic 2: Impossibilities

“Hallucination is inevitable”

Assumption: hallucination is defined as **inconsistencies between a computable LLM and a computable ground truth function.**

(disclaimer: simplified claims)

“Hallucination is inevitable”

Assumption: hallucination is defined as **inconsistencies between a computable LLM and a computable ground truth function.**

Theorem: LLMs cannot learn all the *computable functions* and will therefore inevitably hallucinate if used as general problem solvers.

(disclaimer: simplified claims)

Strong Watermarking Is Impossible for LMs

Definitions

Watermark: a set of outputs $\{y \mid D(y) = 1\}$ detectable by D

Strong watermarking: for any prompt x , and a (watermarked) output y , there is no efficient attacker to obtain y' without watermark that the $LM(x, y') \geq LM(x, y)$.

(disclaimer: simplified claims)

Strong Watermarking Is Impossible for LMs

Definitions

Watermark: a set of outputs $\{y \mid D(y) = 1\}$ detectable by D

Strong watermarking: for any prompt x , and a (watermarked) output y , there is no efficient attacker to obtain y' without watermark that the $LM(x, y') \geq LM(x, y)$.

Theorem: with a *perturbation oracle*, a strong watermarking is impossible. i.e., there always exists an efficient attacker $f: y \rightarrow y'$

(disclaimer: simplified claims)

Strong Watermarking Is Impossible for LMs

Algorithm 1 Pseudocode for our attack

Input: prompt x , watermarked response y , quality oracle Q , perturbation oracle P , random walk length T .

Output: response y' without watermark.

```
 $y' \leftarrow y ;$  // initialize with the  
watermarked response
```

```
for  $t \leftarrow 1$  to  $T$  do
```

```
     $y_t \leftarrow P(x, y') ;$  // apply perturbation
```

```
    if  $Q(x, y_t) \geq Q(x, y)$  then
```

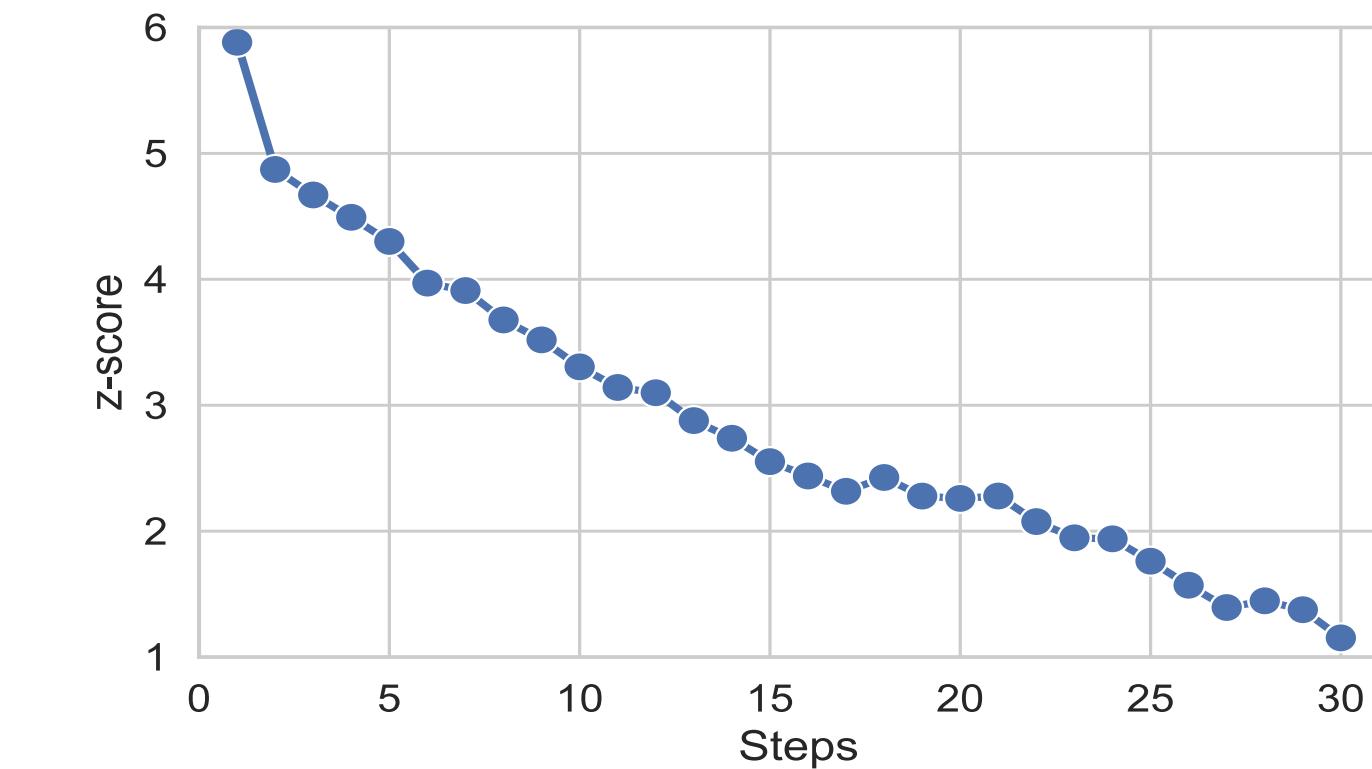
```
         $y' \leftarrow y_t ;$  // update if quality does  
not decrease
```

```
    end
```

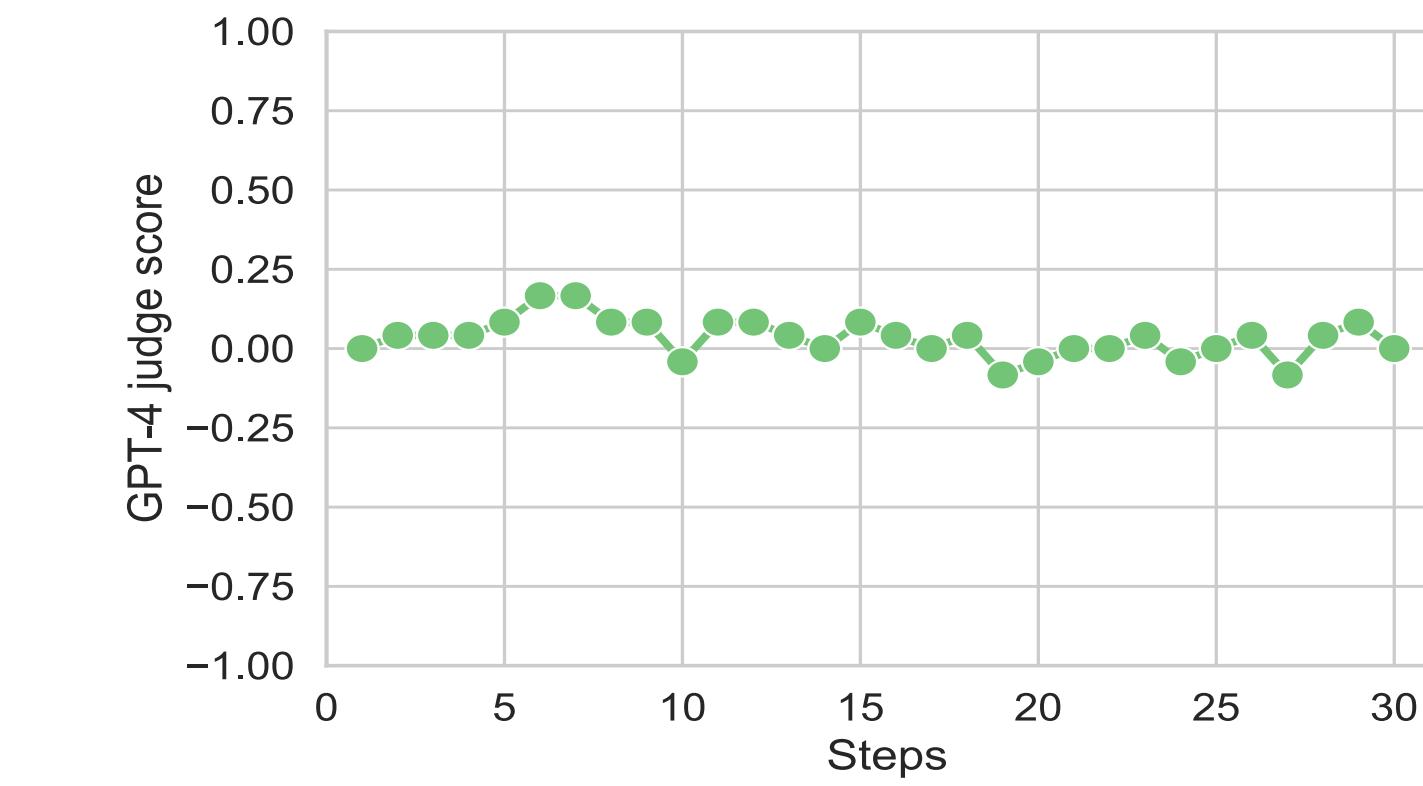
```
end
```

```
return  $y'$  without watermark ; // return the  
de-watermarked response
```

proposed attack algorithm by
rejection sampling

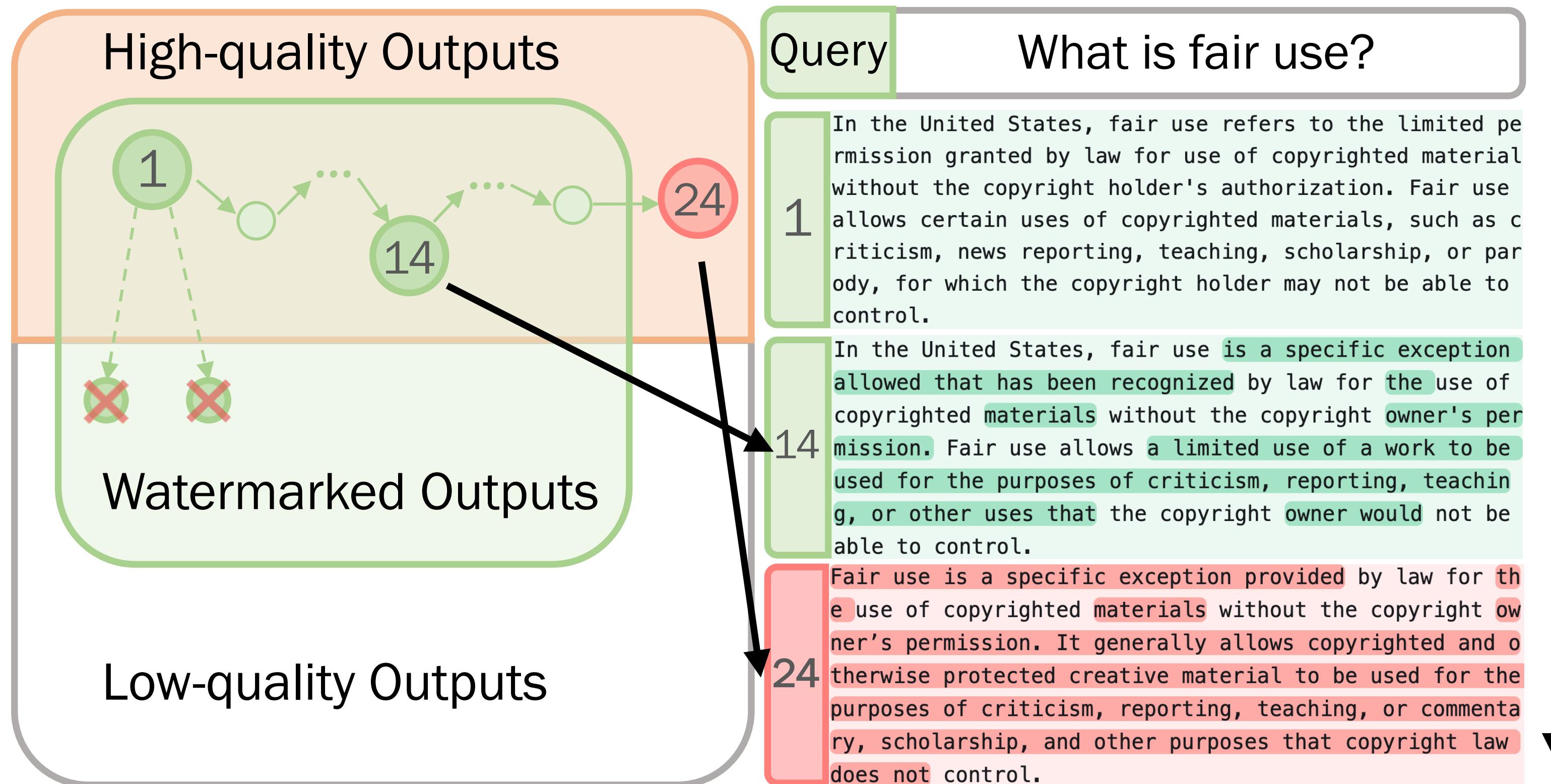


detection score decreases



GPT-4 score (quality) remains stable

Strong Watermarking Is Impossible for LMs



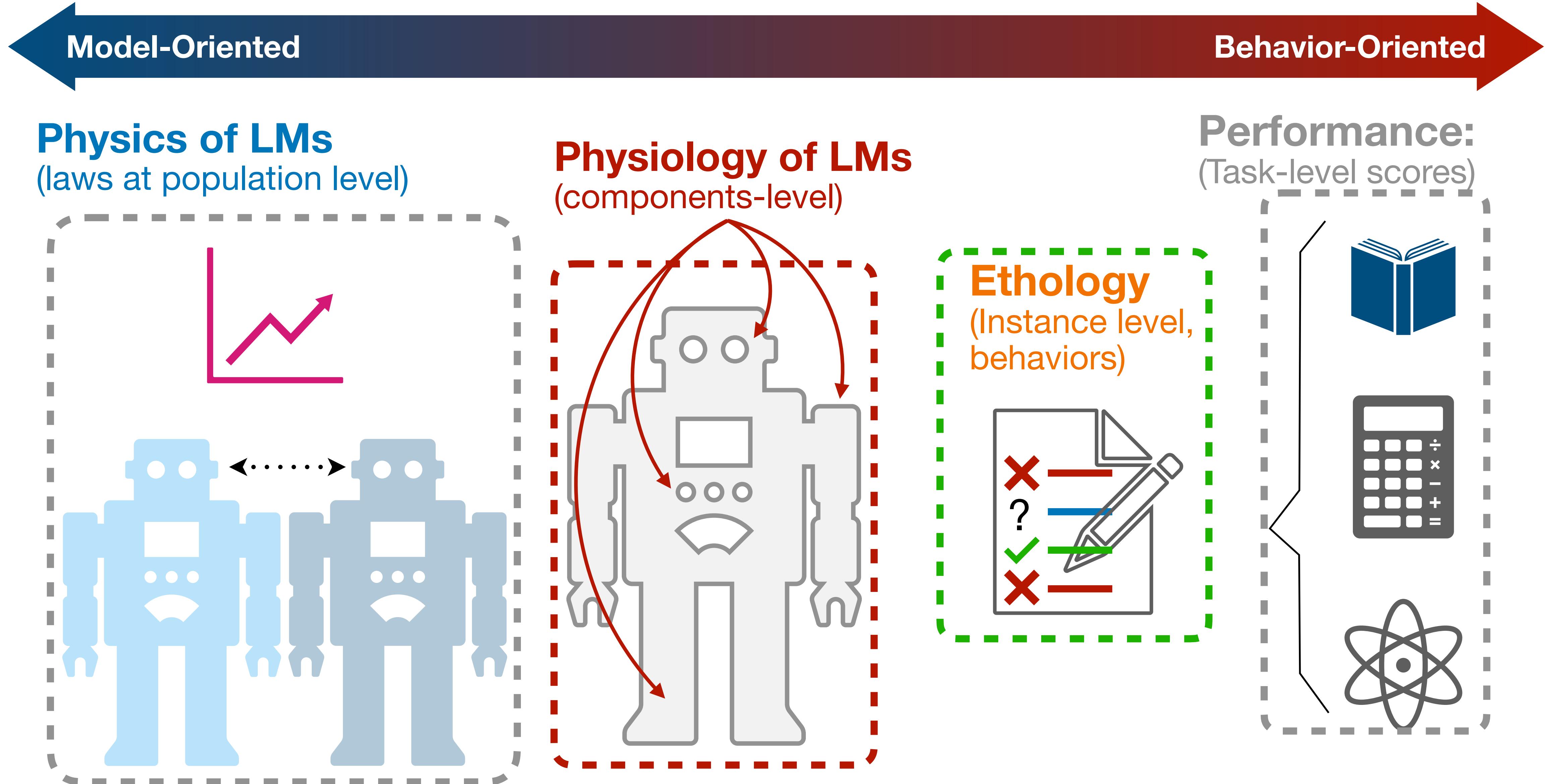
during sampling,
the text is less
detectable but
quality remains

Strong Watermarking Is Impossible for LMs



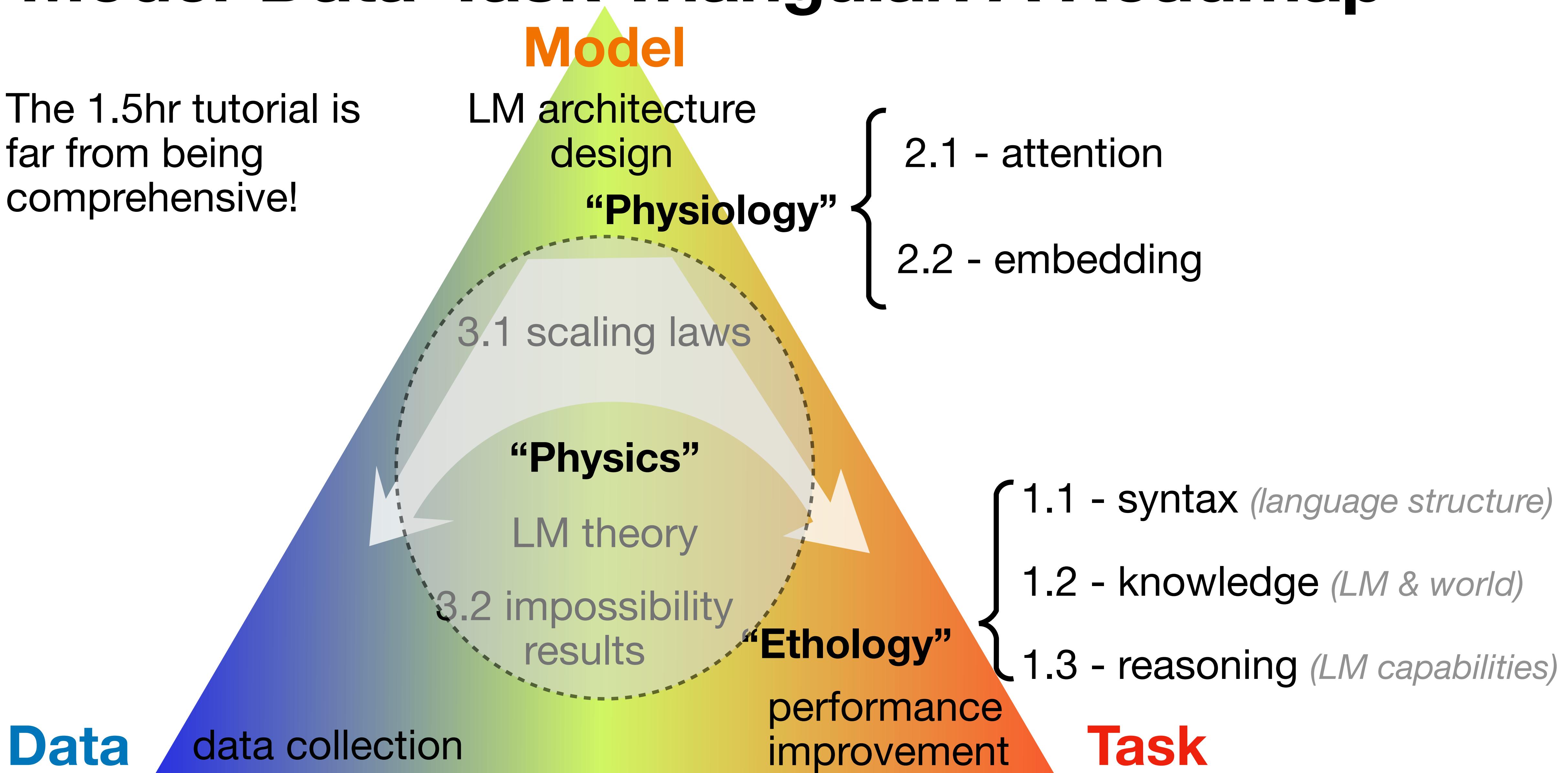
p-value:
(lower = deeper watermark)

A Retrospect of Science of LMs



Model-Data-Task Triangular: A Roadmap

The 1.5hr tutorial is far from being comprehensive!



Discussions and Q&A

- Will we have a unified scientific framework for analyzing LMs?
or, multiple levels of frameworks instead?

particles —→ **fluid (mass of particles)** —→ **supersonic flow**

$$F = ma$$



$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\frac{1}{\rho} \nabla \rho + \nu \nabla^2 \mathbf{u}$$



shock waves, etc...



- Will we be able to characterize every phenomenon?
or, will there always be a next unsolved problem, just as in curse of dimensionality

image credit: <https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/>

