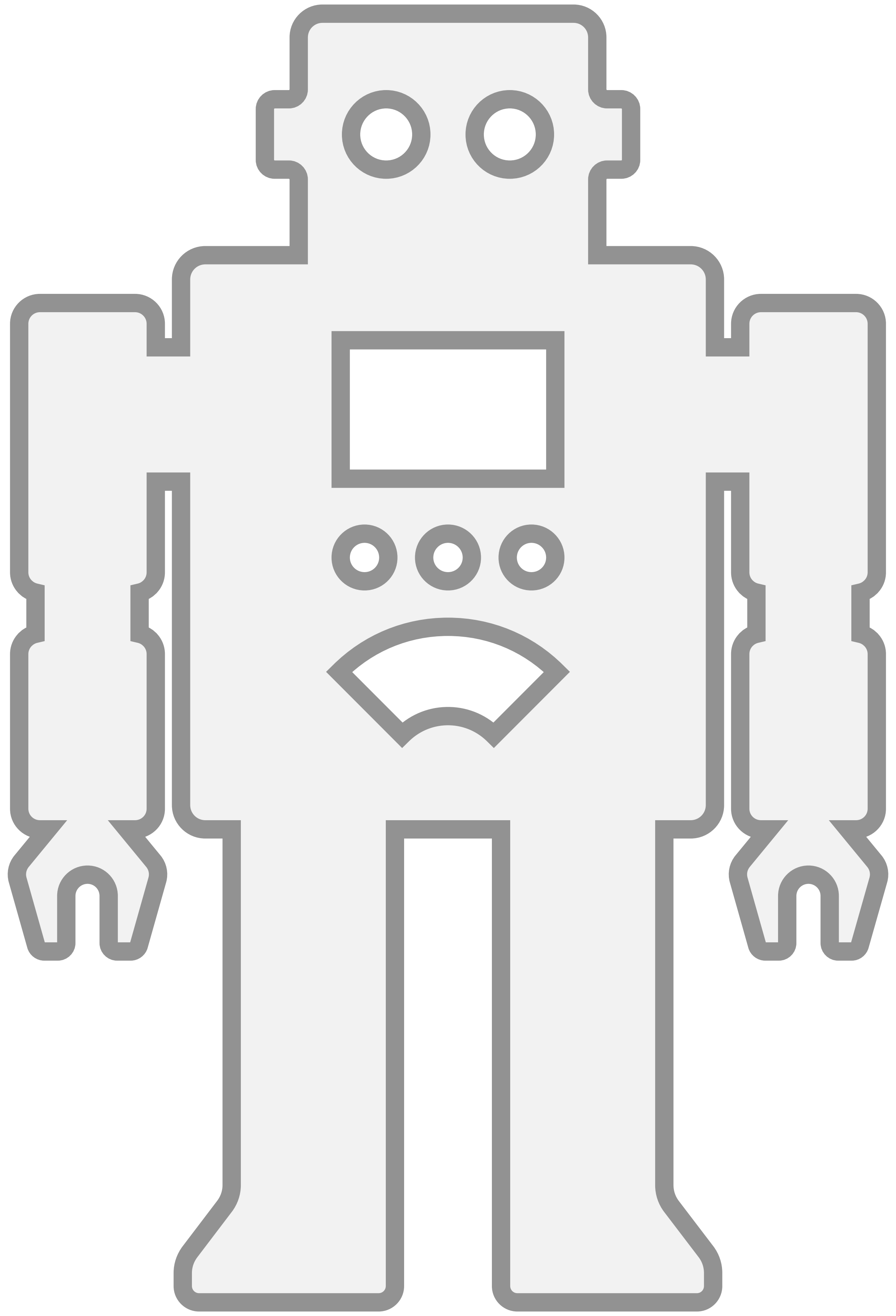
Towards a
Physiology of Language Models:
Elucidating and Utilizing
Hidden Language Representation

March 18, 2025, 5-5:45pm PT

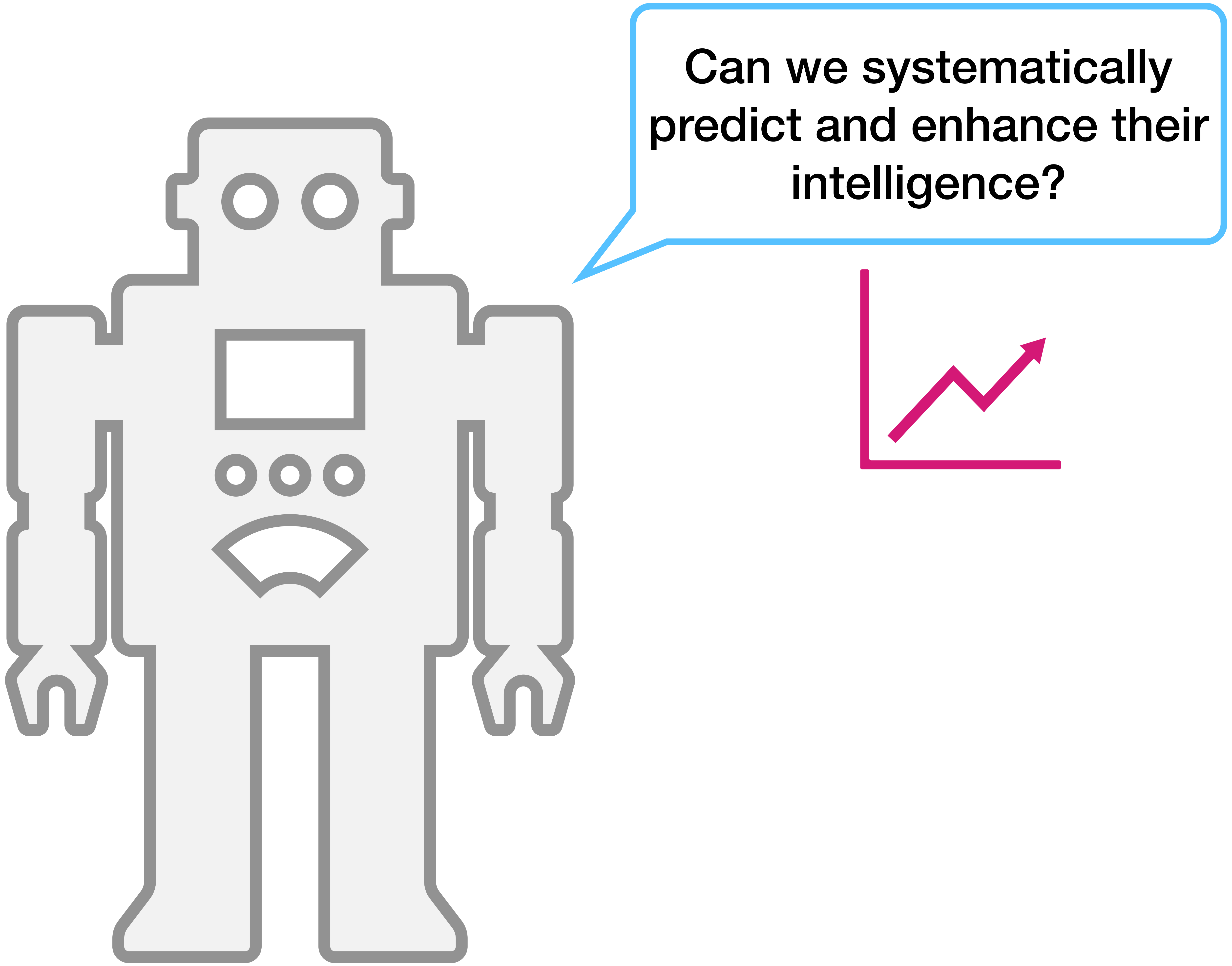
Chi Han, Ph.D. Student @ UIUC, https://glaciohound.github.io/

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

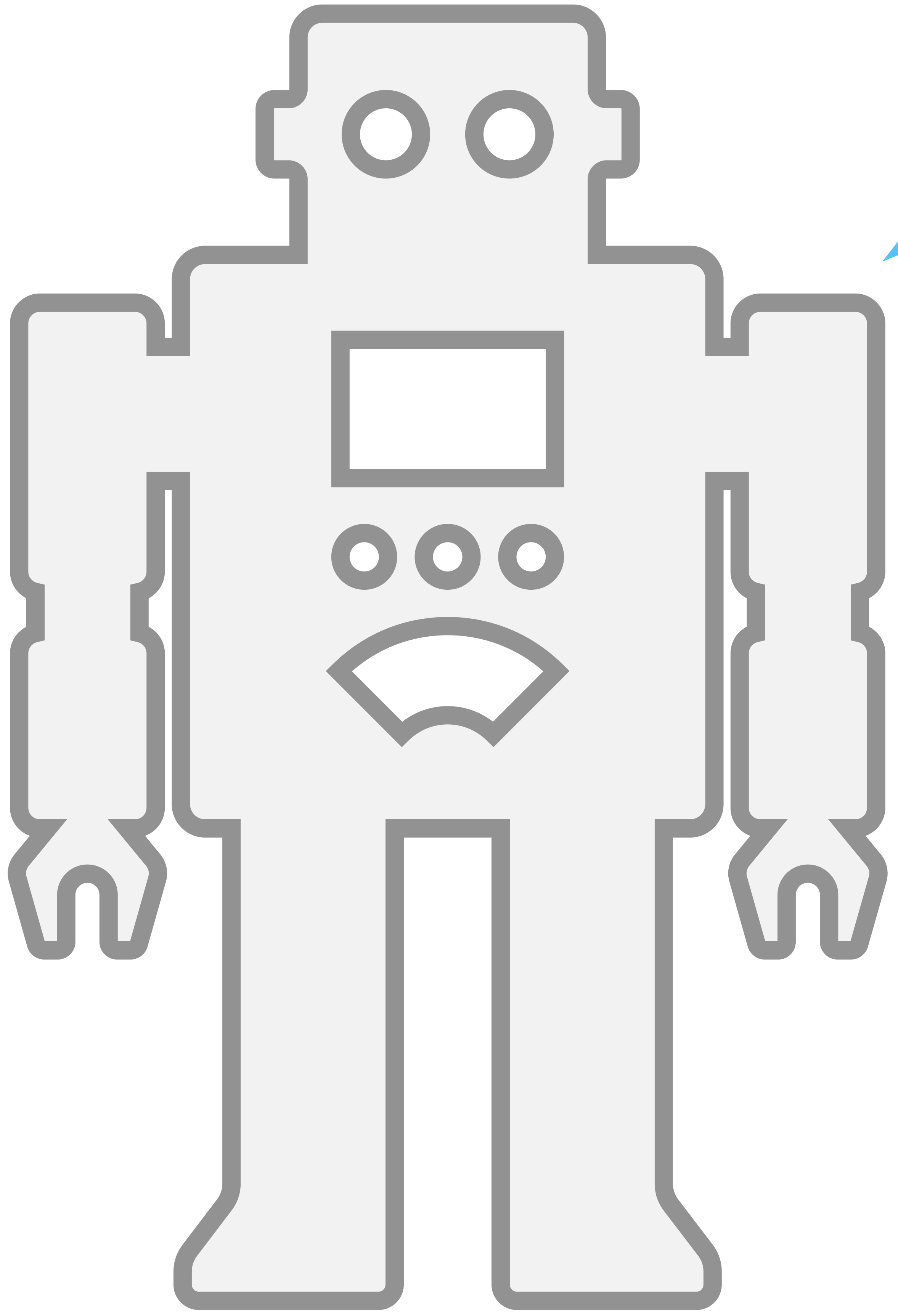
Can We Systematically Describe How LMs Behave?



Can We Systematically Describe How LMs Behave?



Can We Systematically Describe How LMs Behave?

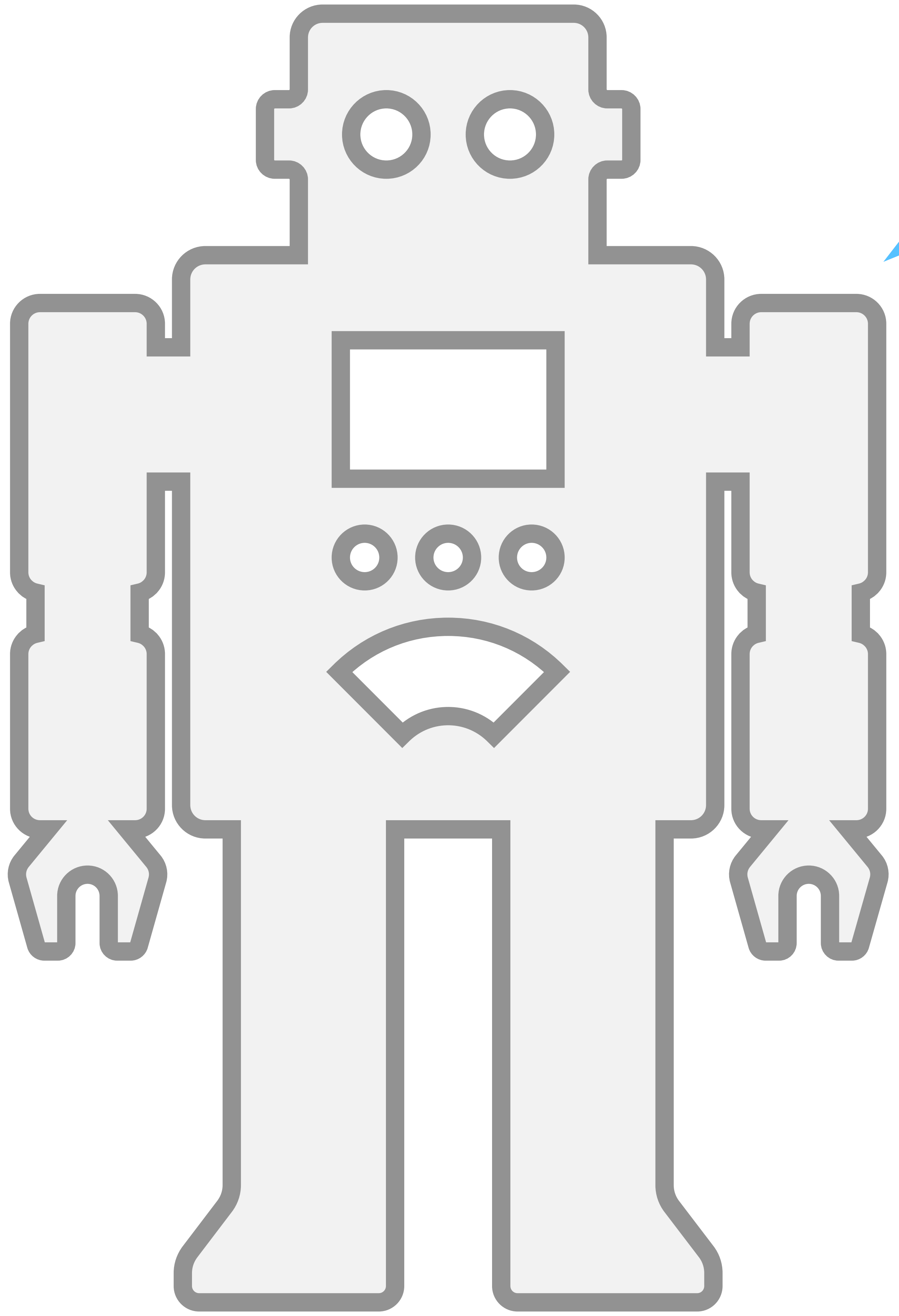


Can we systematically predict and enhance their intelligence?

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

Can We Systematically Describe How LMs Behave?

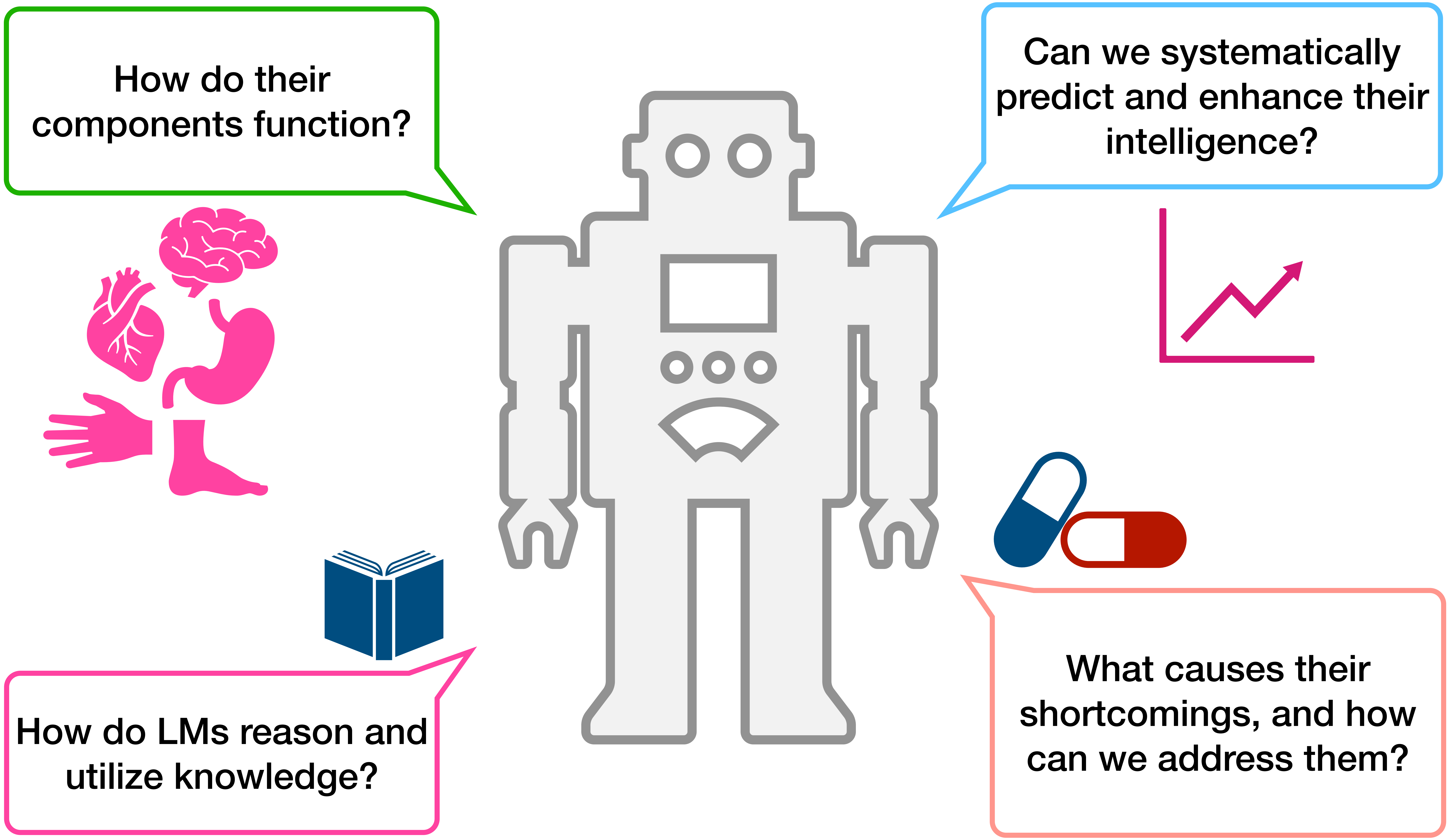
How do their components function?



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?



Why Do We Need A New Science?

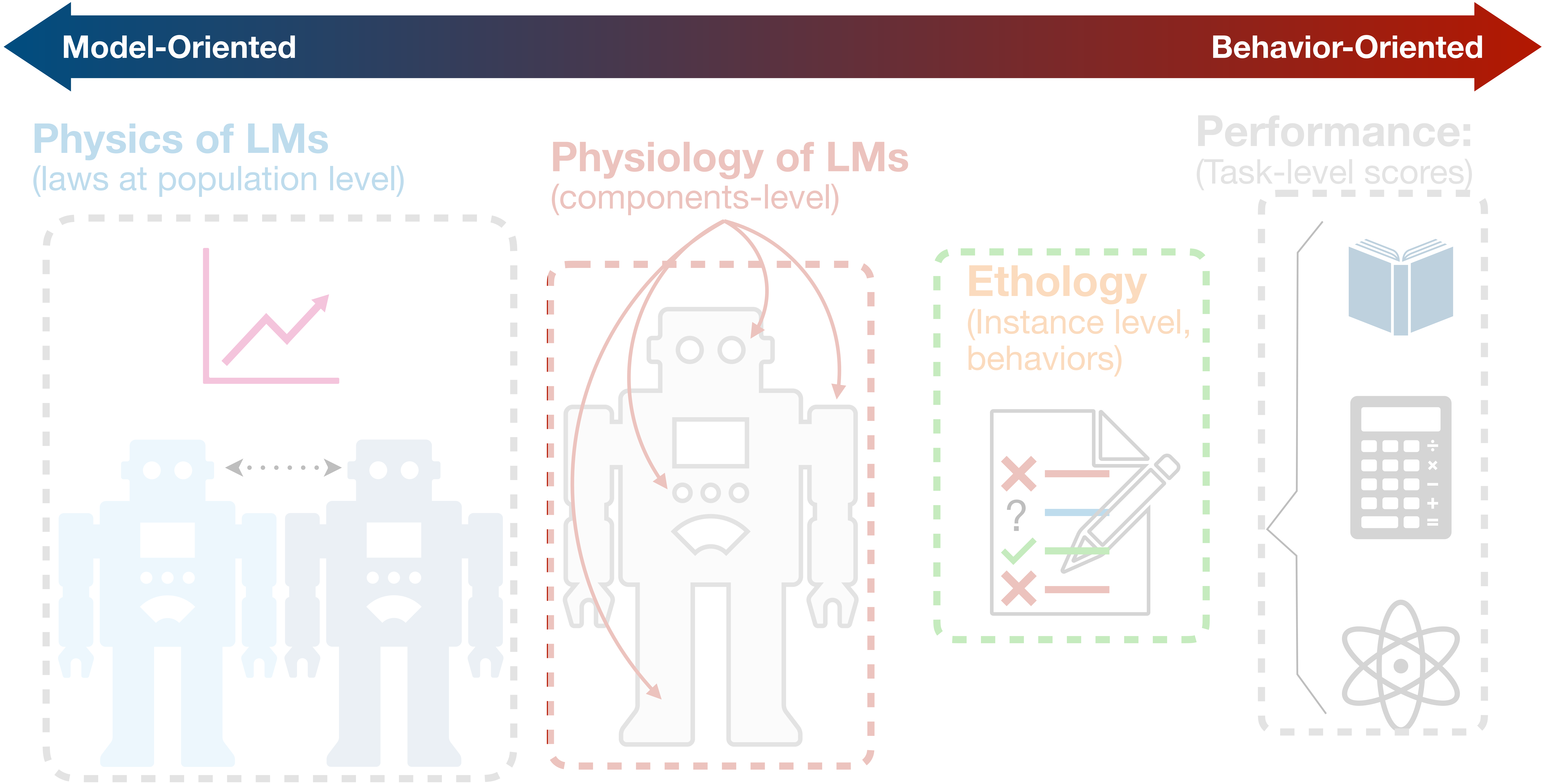
New sciences often emerge as a result of scaling up old sciences

Machine Learning ⟶ **Deep Learning** ⟶ **Language Models**

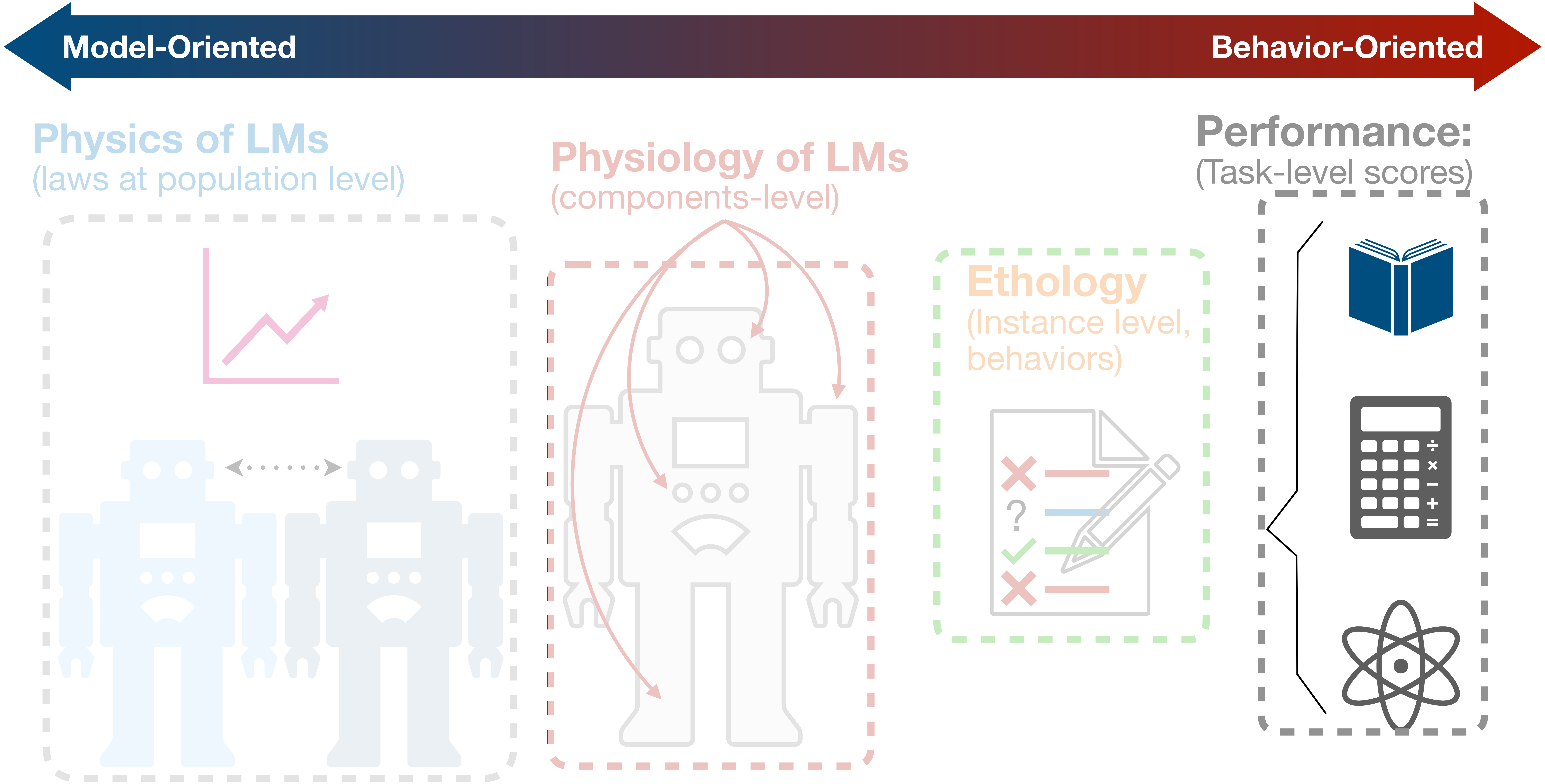
Machine Learning	Deep Learning	Language Models
PAC theory, optimization, …	Gradient Descent, Neural Tangent Kernel, …	**A Sciences of LMs**

https://glaciohound.github.io/Science-of-LLMs-Tutorial/

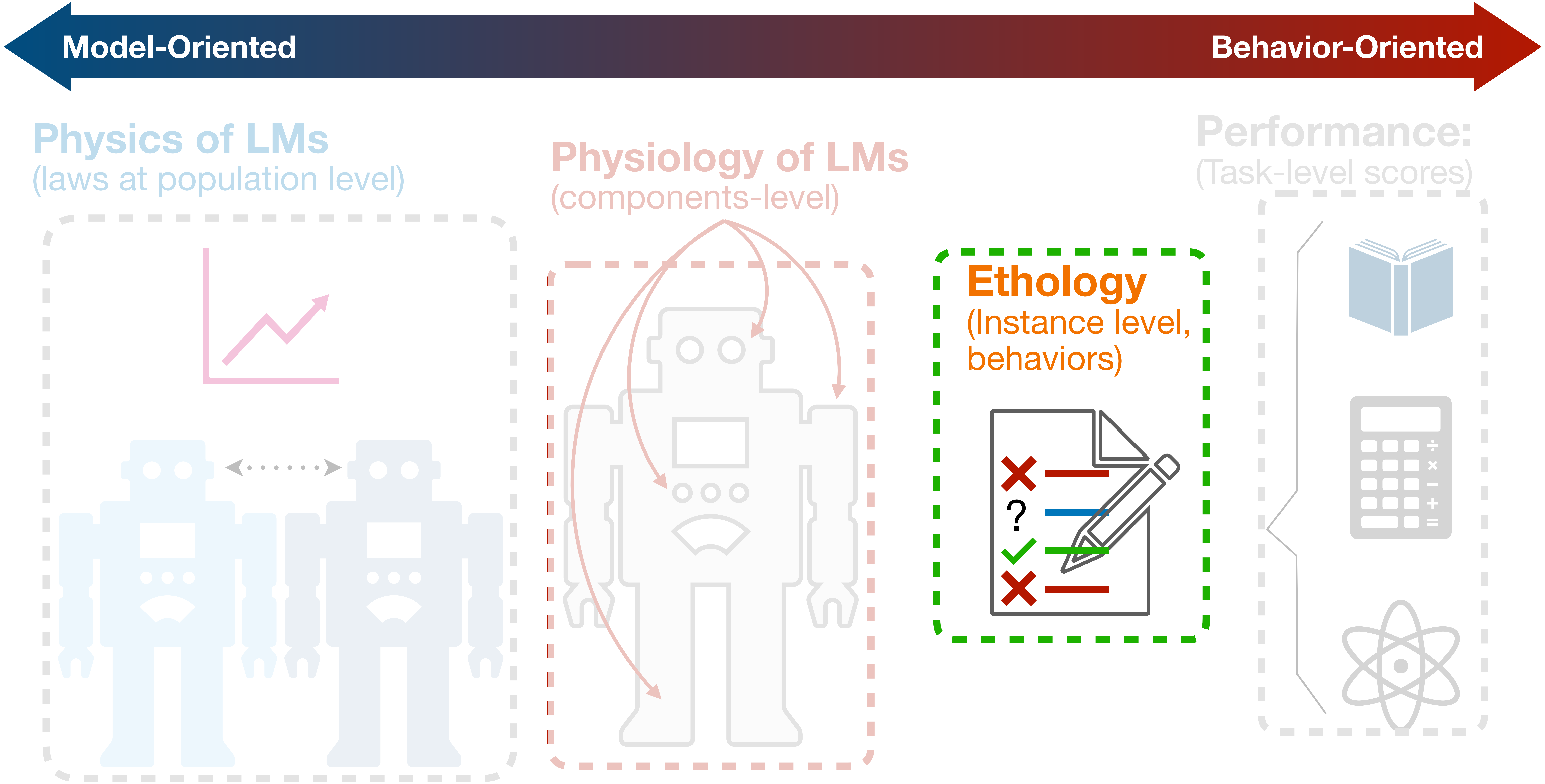
Spectrum of Sciences of LMs



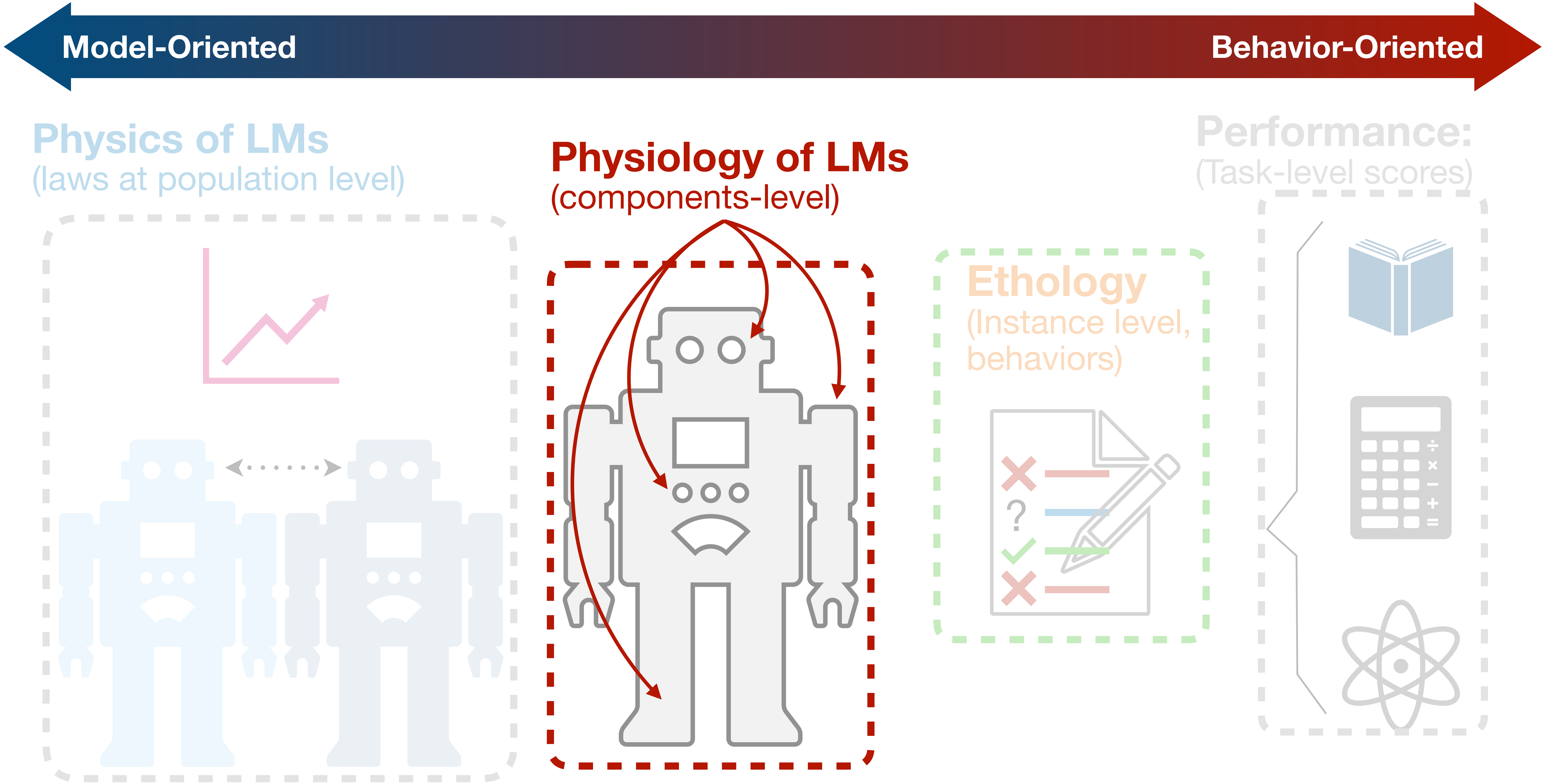
Spectrum of Sciences of LMs



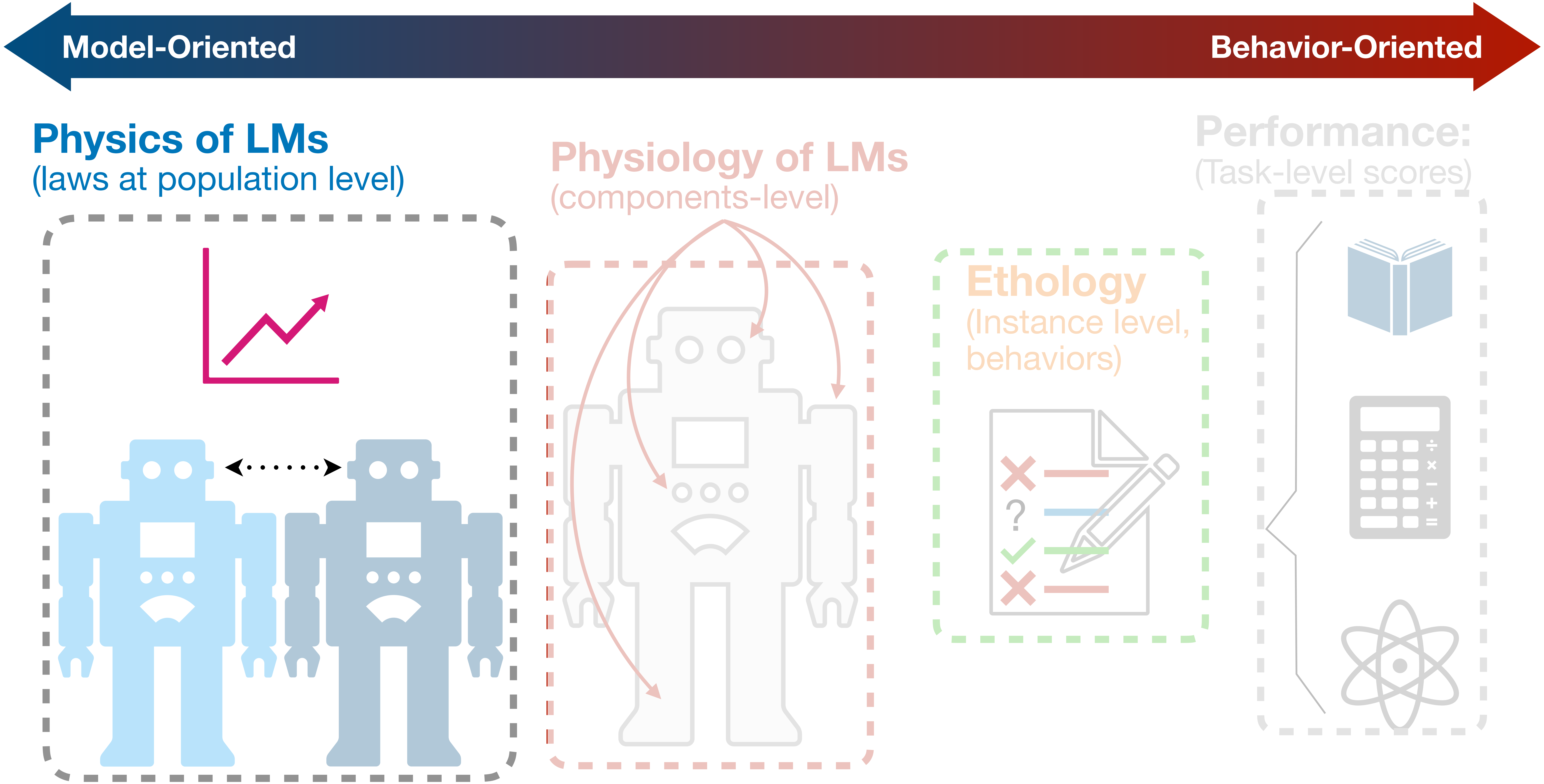
Spectrum of Sciences of LMs



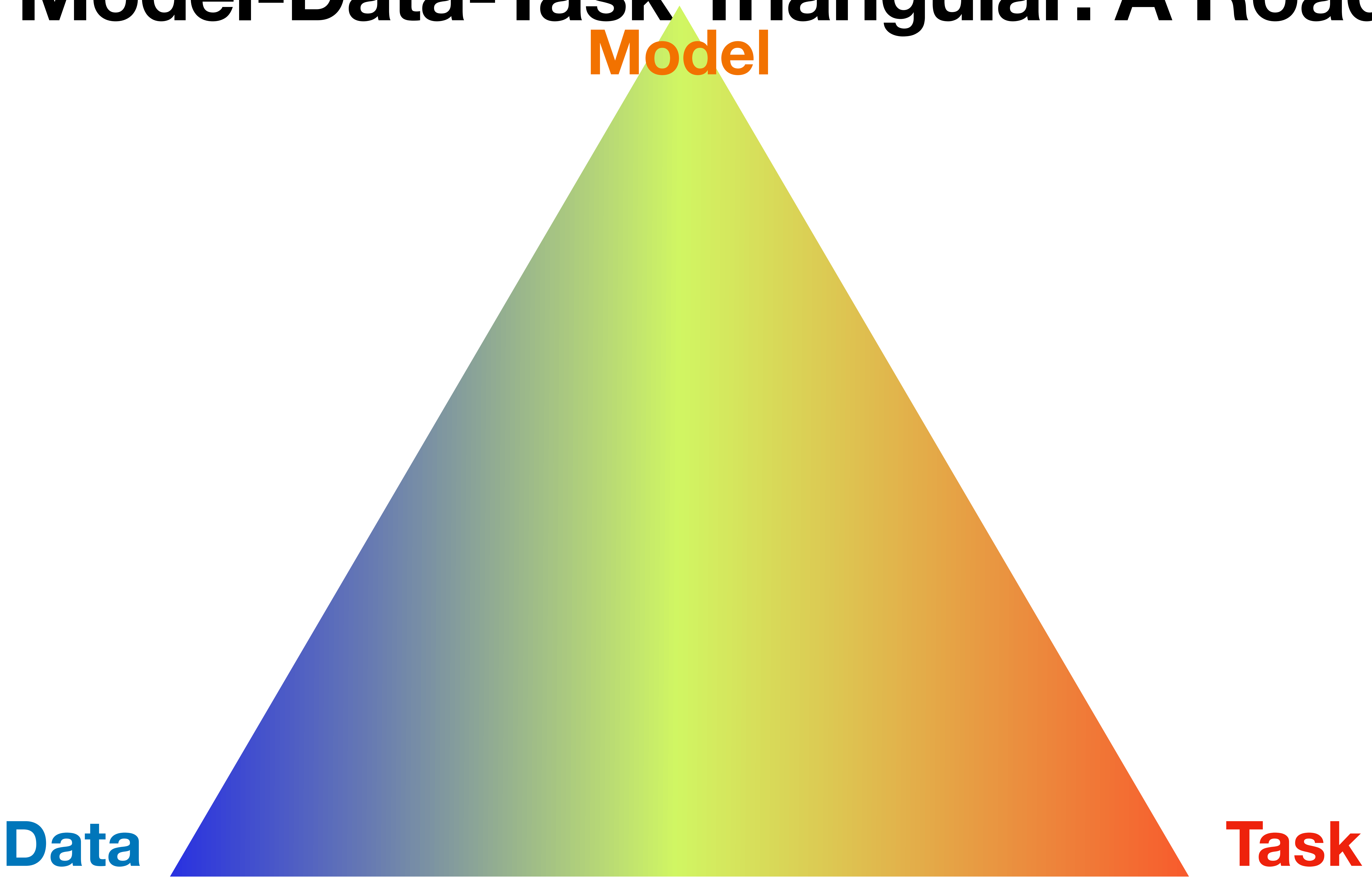
Spectrum of Sciences of LMs



Spectrum of Sciences of LMs

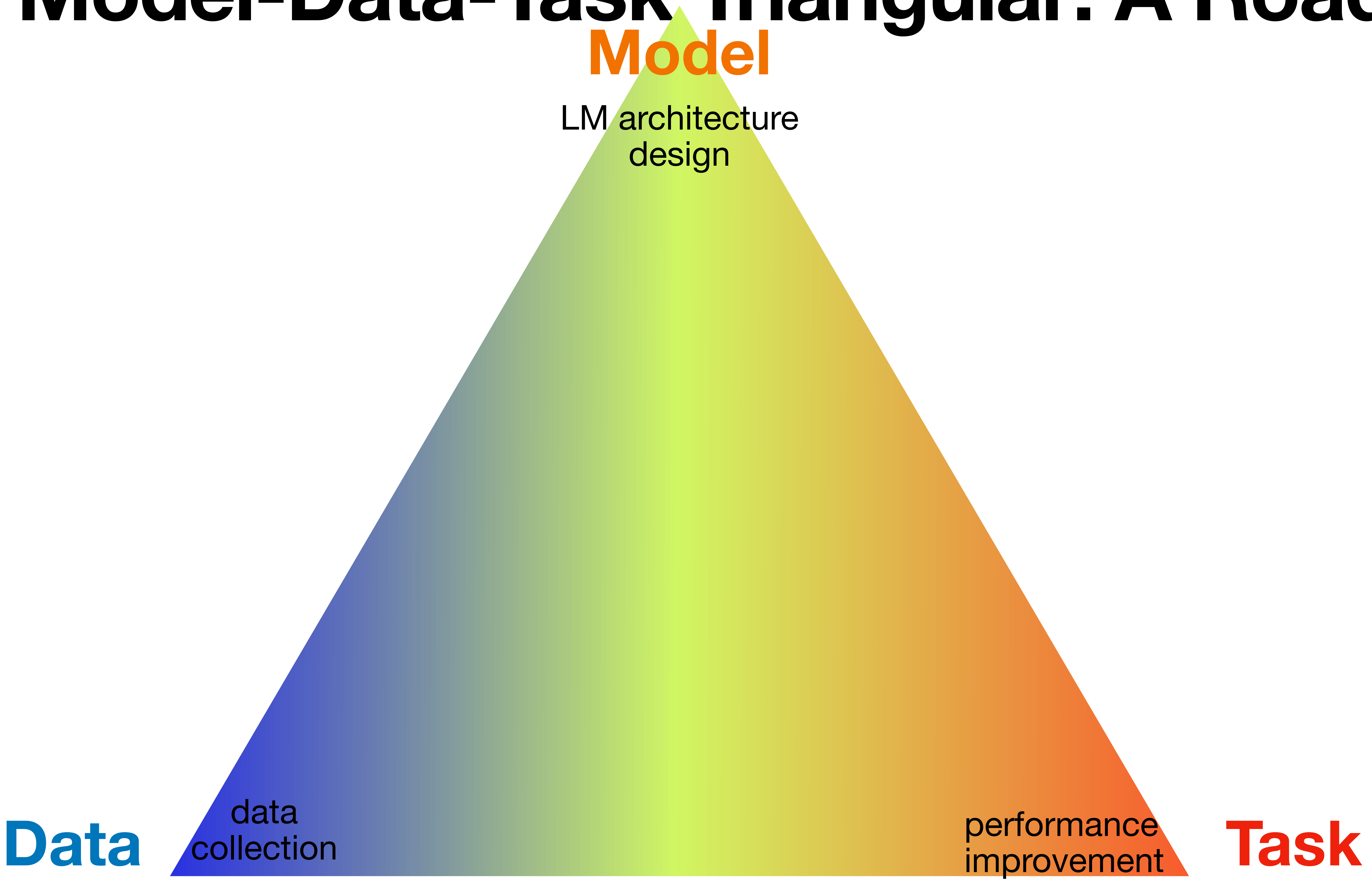


Model-Data-Task Triangular: A Roadmap

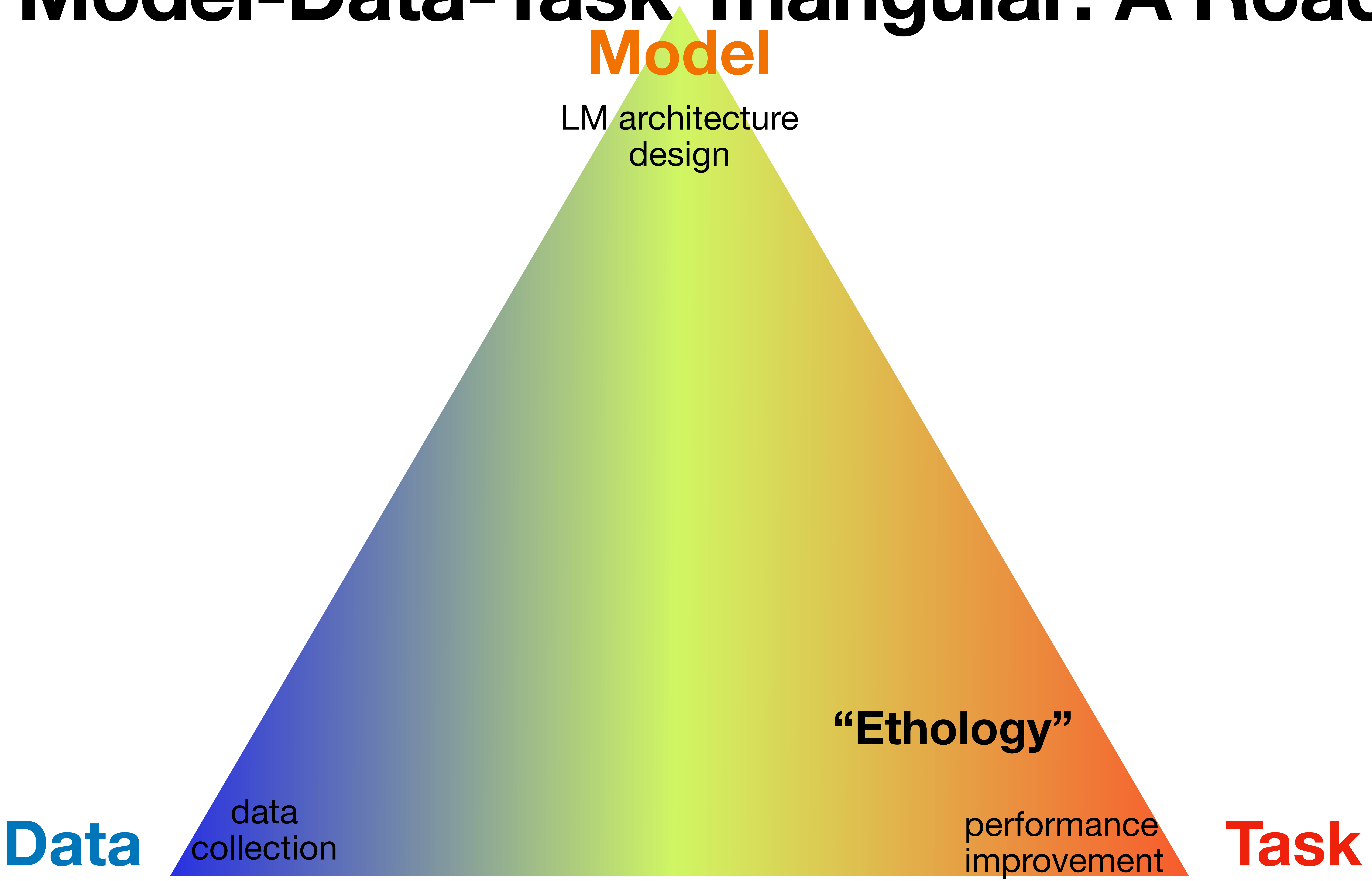


Chi Han et al, “The Quest for a Science of Language Models”, Proceedings of the AAAI 2025, Tutorials Session https://glaciohound.github.io/Science-of-LLMs-Tutorial/

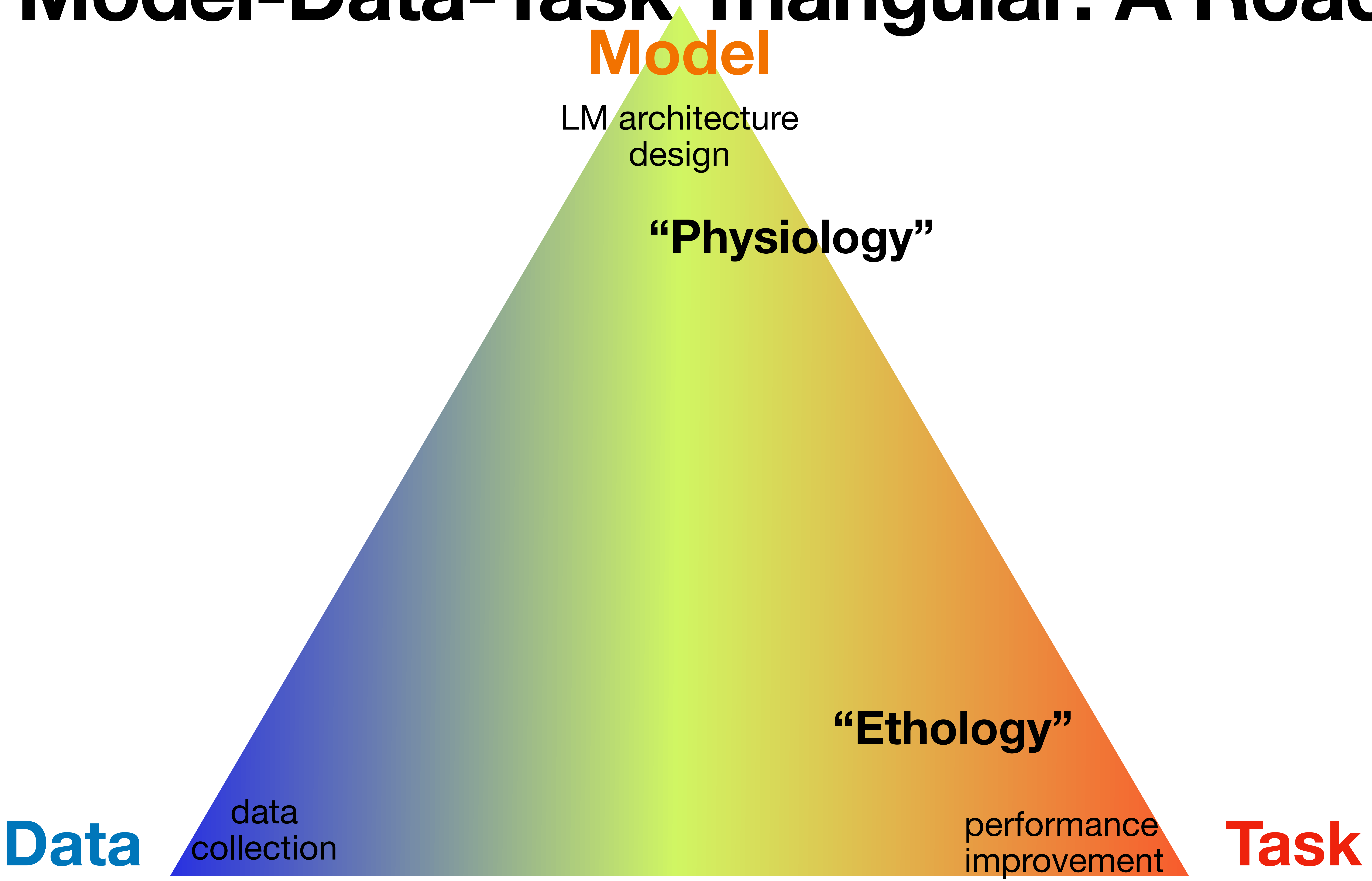
Model-Data-Task Triangular: A Roadmap



Model-Data-Task Triangular: A Roadmap

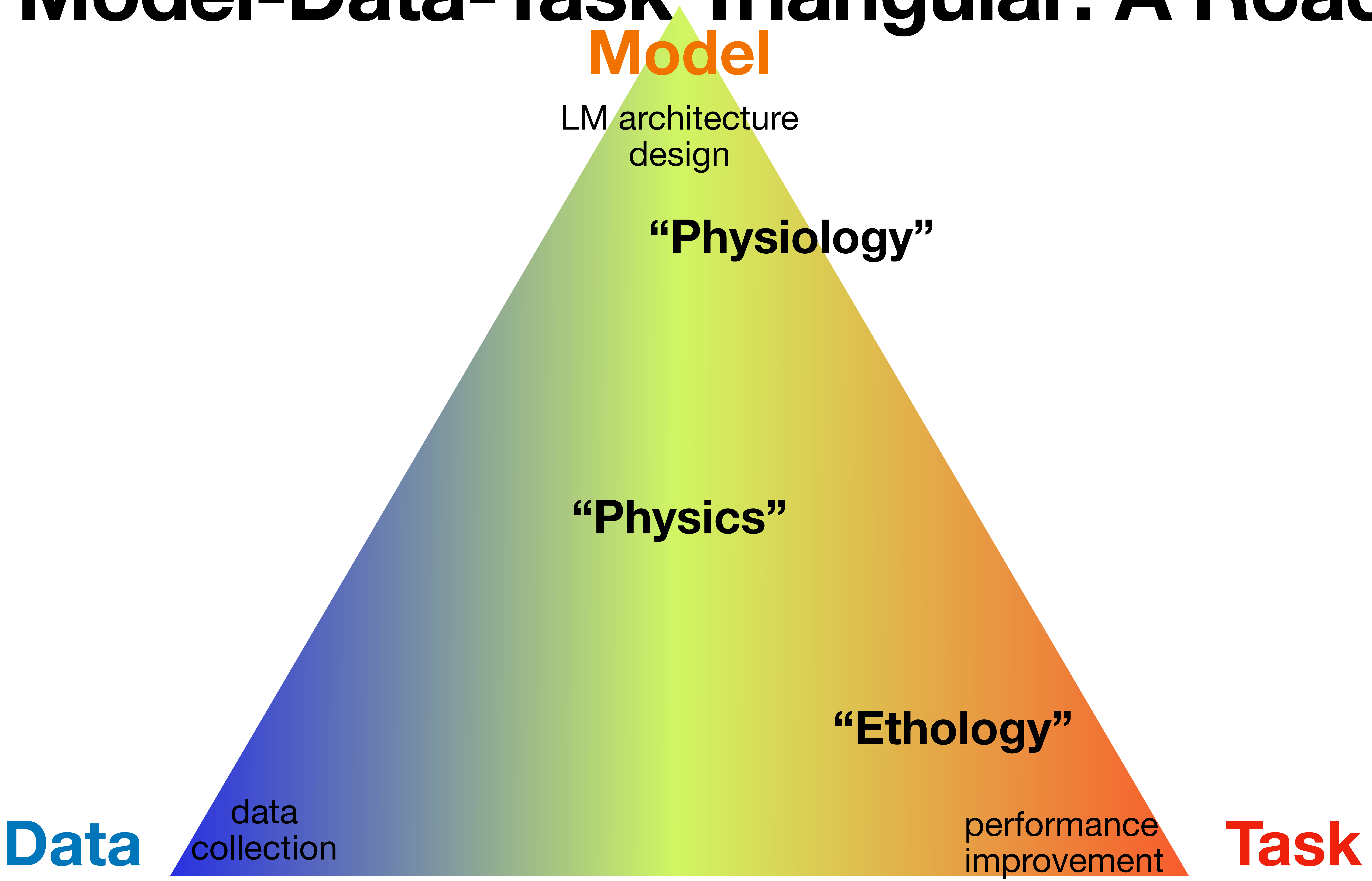


Model-Data-Task Triangular: A Roadmap



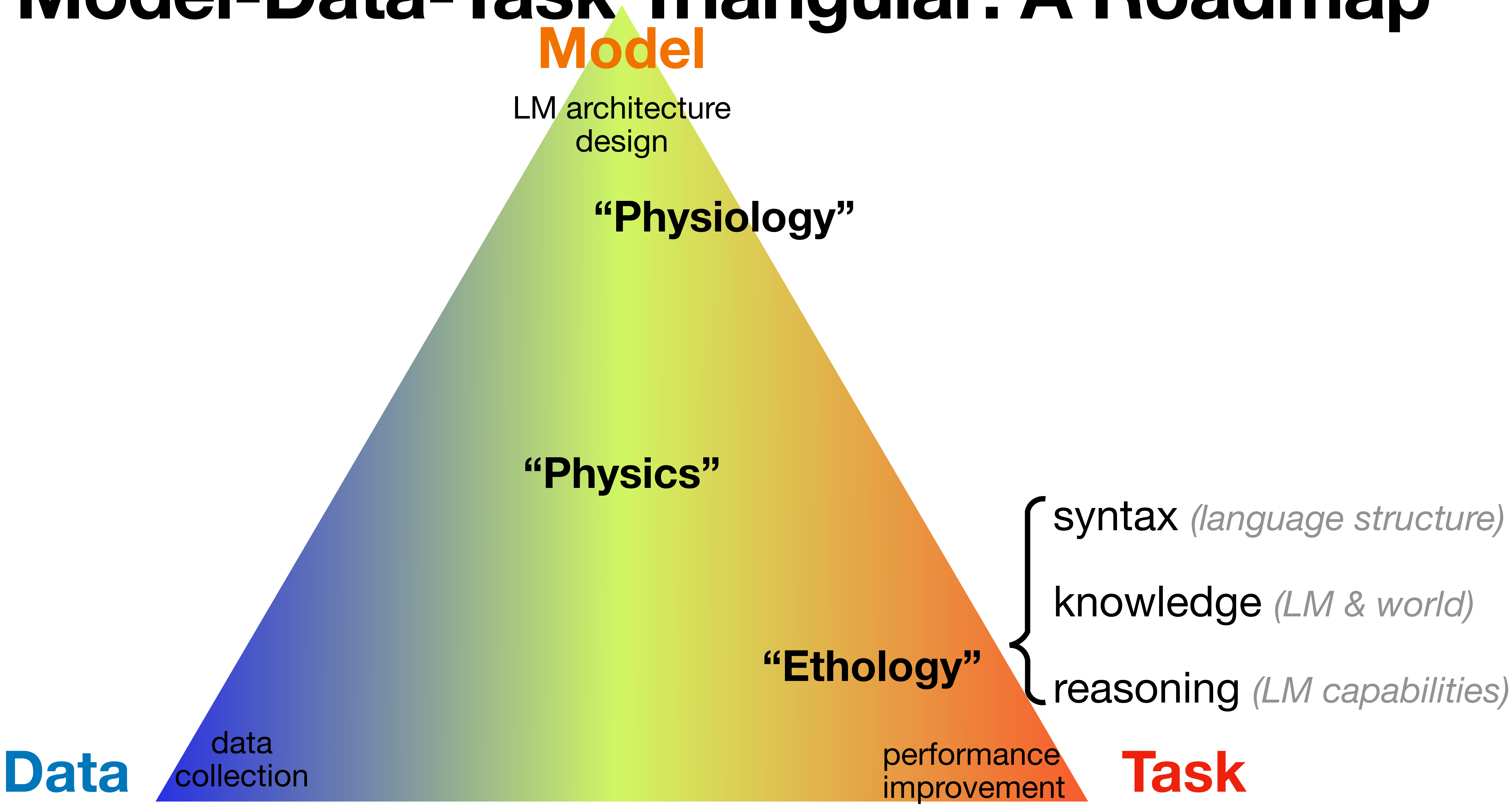
Chi Han et al, "The Quest for a Science of Language Models", Proceedings of the AAAI 2025, Tutorials Session https://glaciohound.github.io/Science-of-LLMs-Tutorial/

Model-Data-Task Triangular: A Roadmap



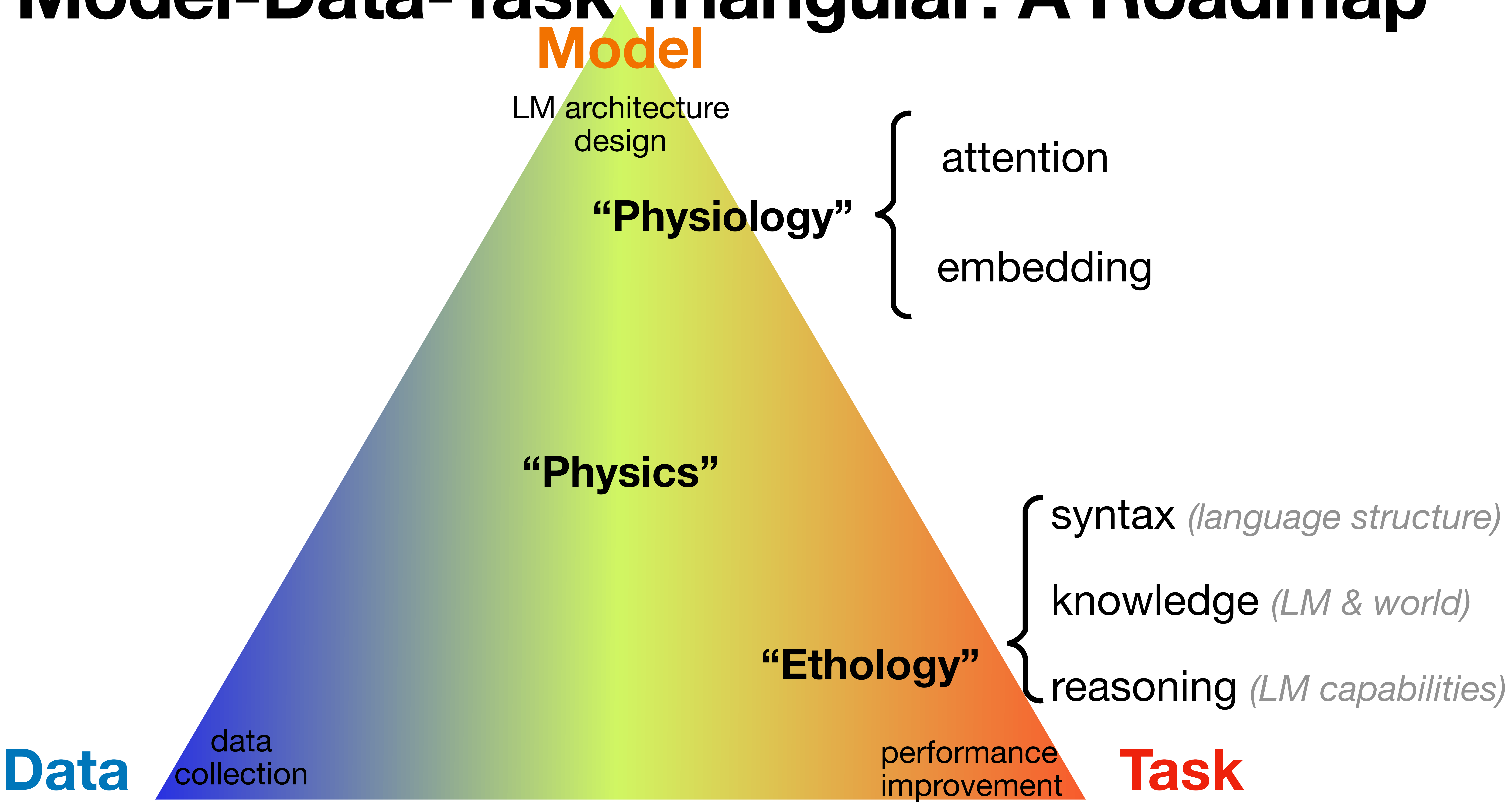
Chi Han et al, "The Quest for a Science of Language Models", Proceedings of the AAAI 2025, Tutorials Session https://glaciohound.github.io/Science-of-LLMs-Tutorial/

Model-Data-Task Triangular: A Roadmap

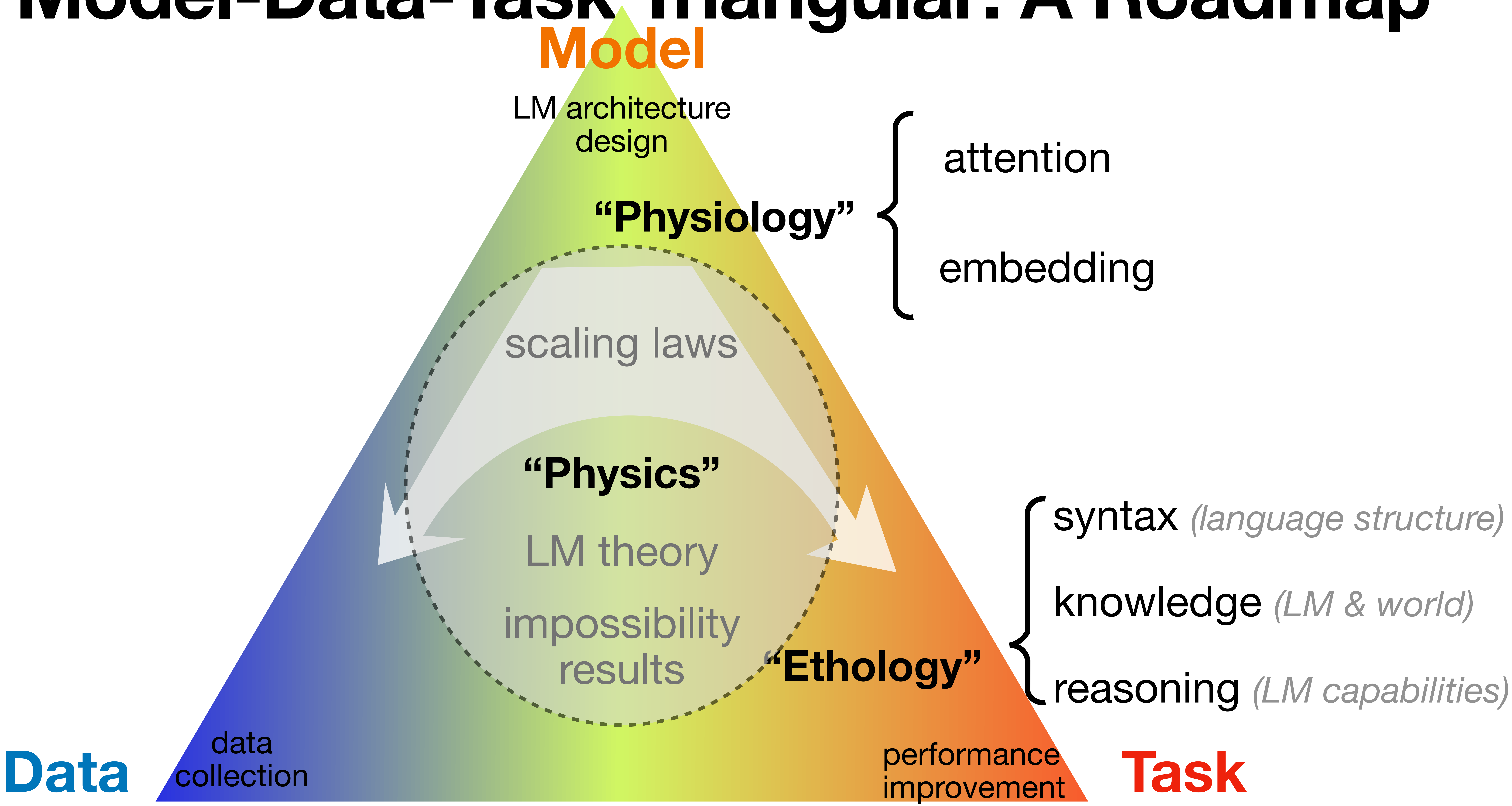


Chi Han et al, "The Quest for a Science of Language Models", Proceedings of the AAAI 2025, Tutorials Session https://glaciohound.github.io/Science-of-LLMs-Tutorial/

Model-Data-Task Triangular: A Roadmap



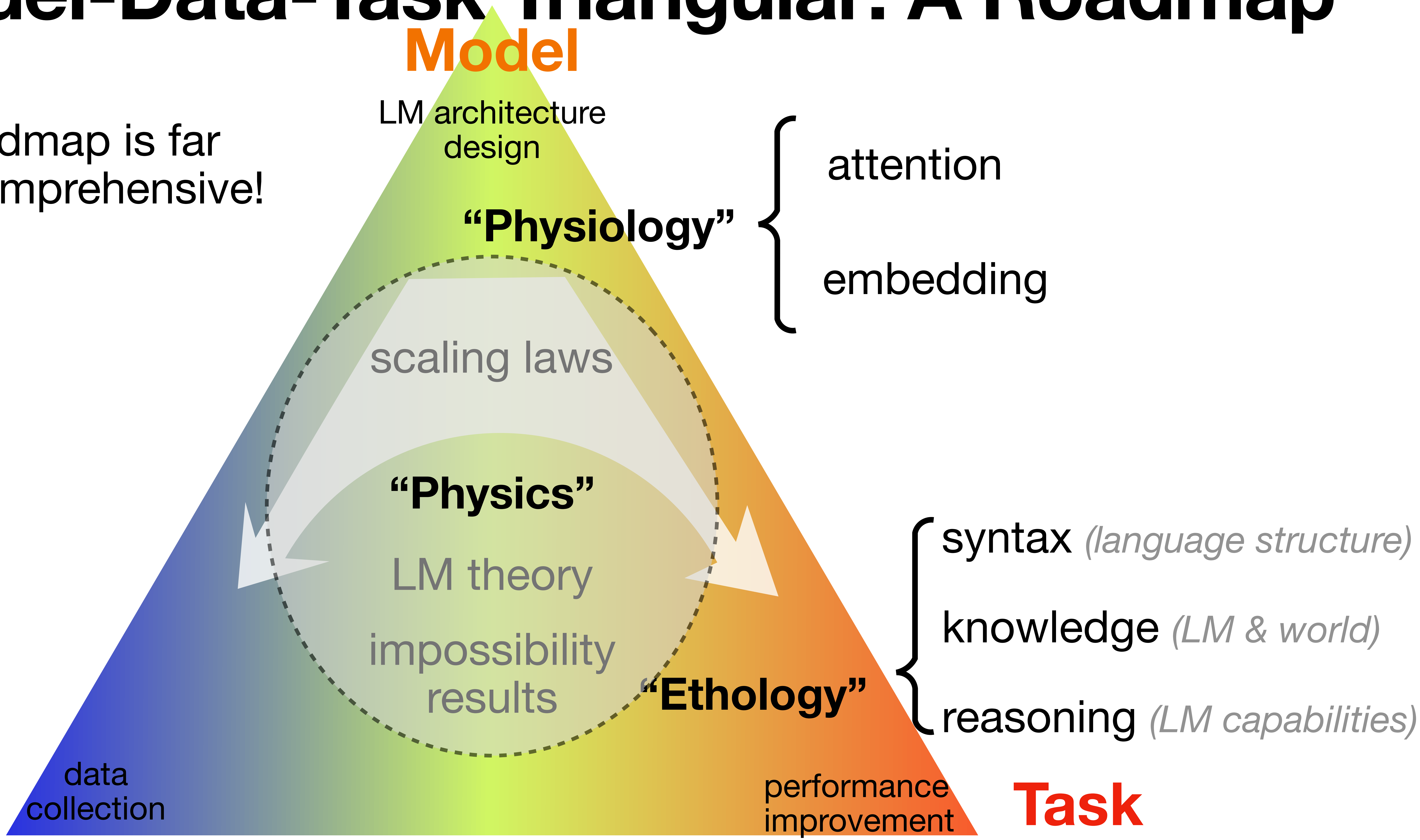
Model-Data-Task Triangular: A Roadmap



Chi Han et al, "The Quest for a Science of Language Models", Proceedings of the AAAI 2025, Tutorials Session https://glaciohound.github.io/Science-of-LLMs-Tutorial/

Model-Data-Task Triangular: A Roadmap

The roadmap is far from comprehensive!



Chi Han et al, "The Quest for a Science of Language Models", Proceedings of the AAAI 2025, Tutorials Session https://glaciohound.github.io/Science-of-LLMs-Tutorial/

Prerequisites: Language Modeling

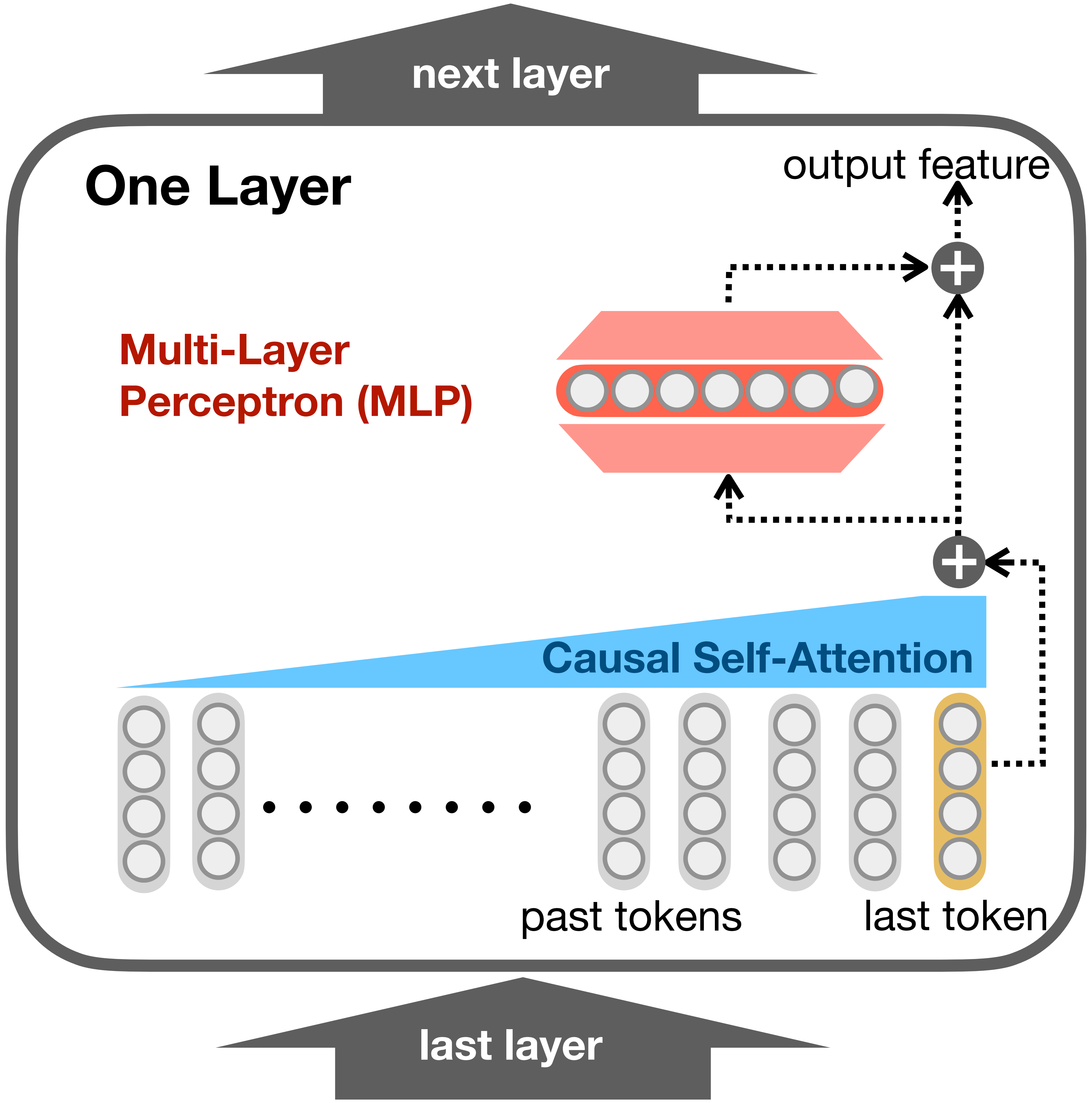
Large Language Models are "impressive"

Input: $[x_1, x_2, \cdots, x_n]$ **Output:** x_{n+1}

Language Model: $P(X_{n+1} | x_1, x_2, \cdots, x_n)$

next-word probability

Language Modeling



A Transformer-Based Architecture

Physiology: How Do Components Function in Language Models?

Topics

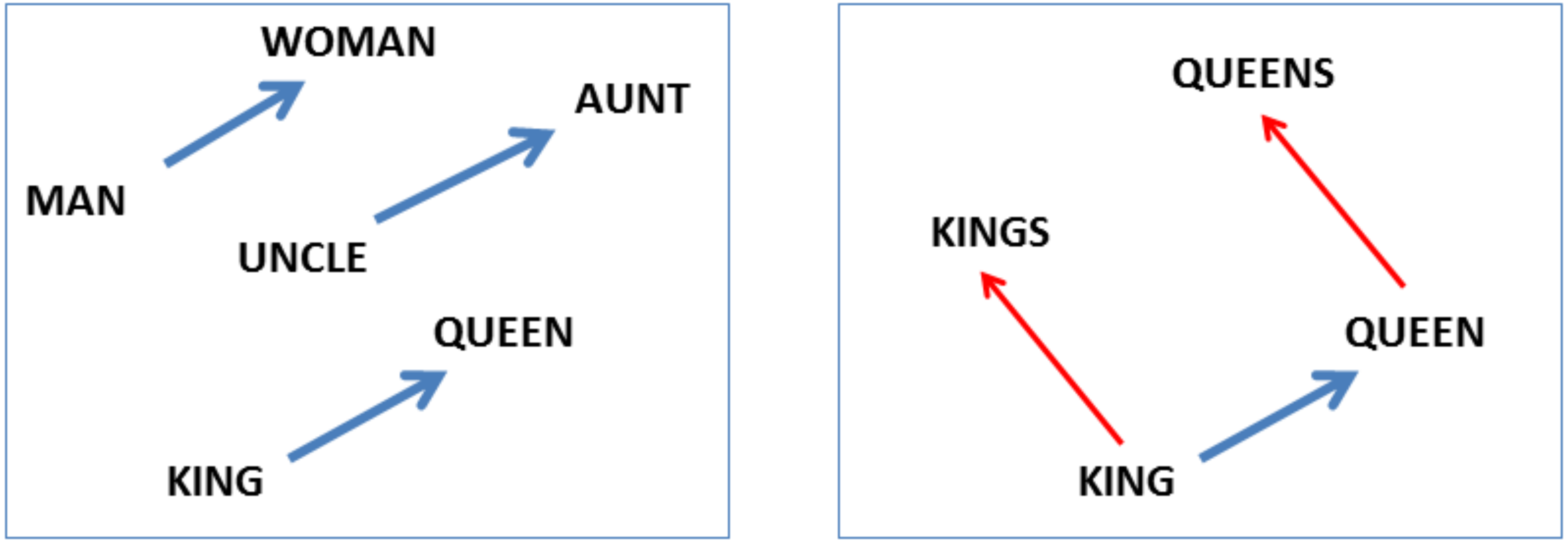
- **Attention**: Attention, position and context
- **Embeddings**: What is the function of word embeddings

What Is the Function of Word Embeddings

Topic 1: Embedding

What Do Word Embeddings Embed?

Previous papers mostly focus on word-level interpretations

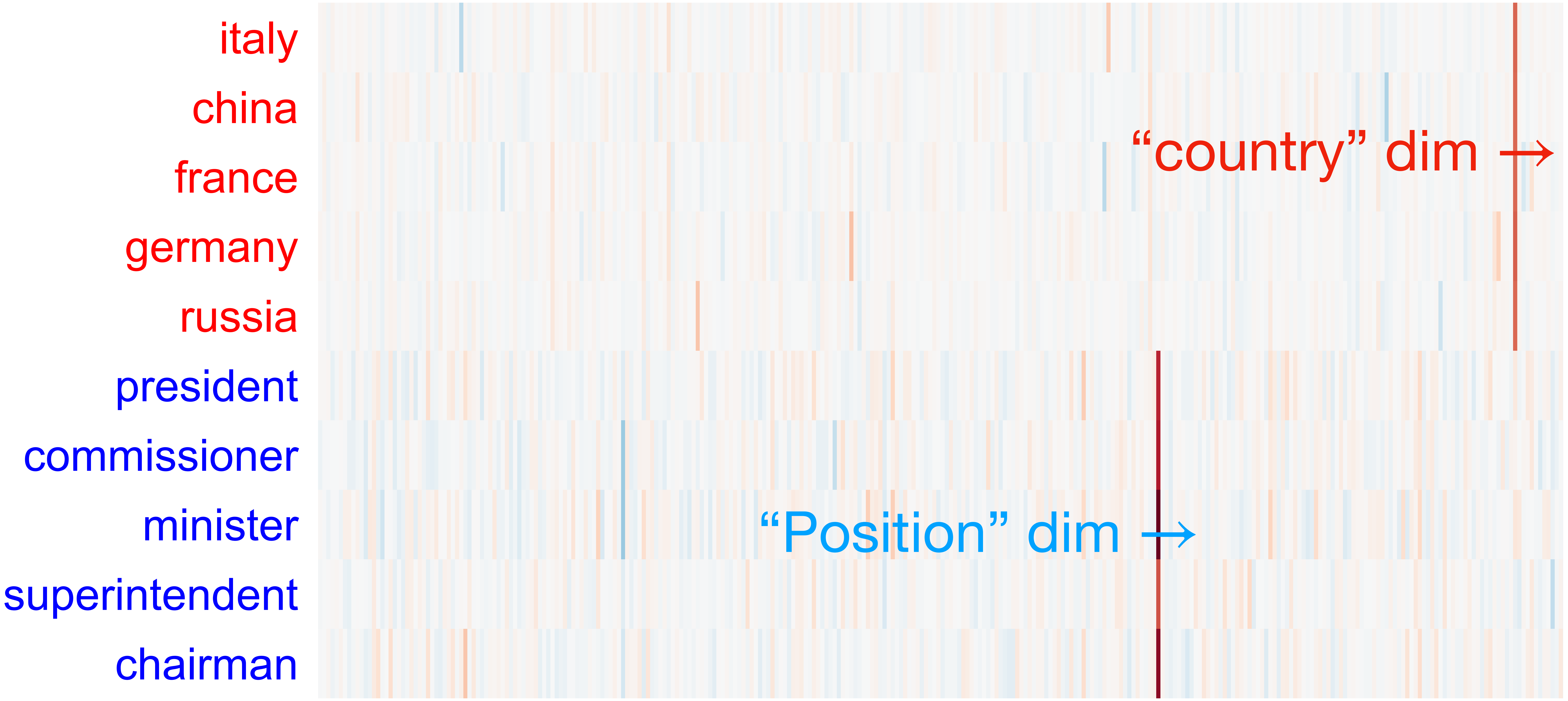


(a) Analogical Relations (metric space)

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 2013.
Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

What Do Word Embeddings Embed?

Previous papers mostly focus on word-level interpretations



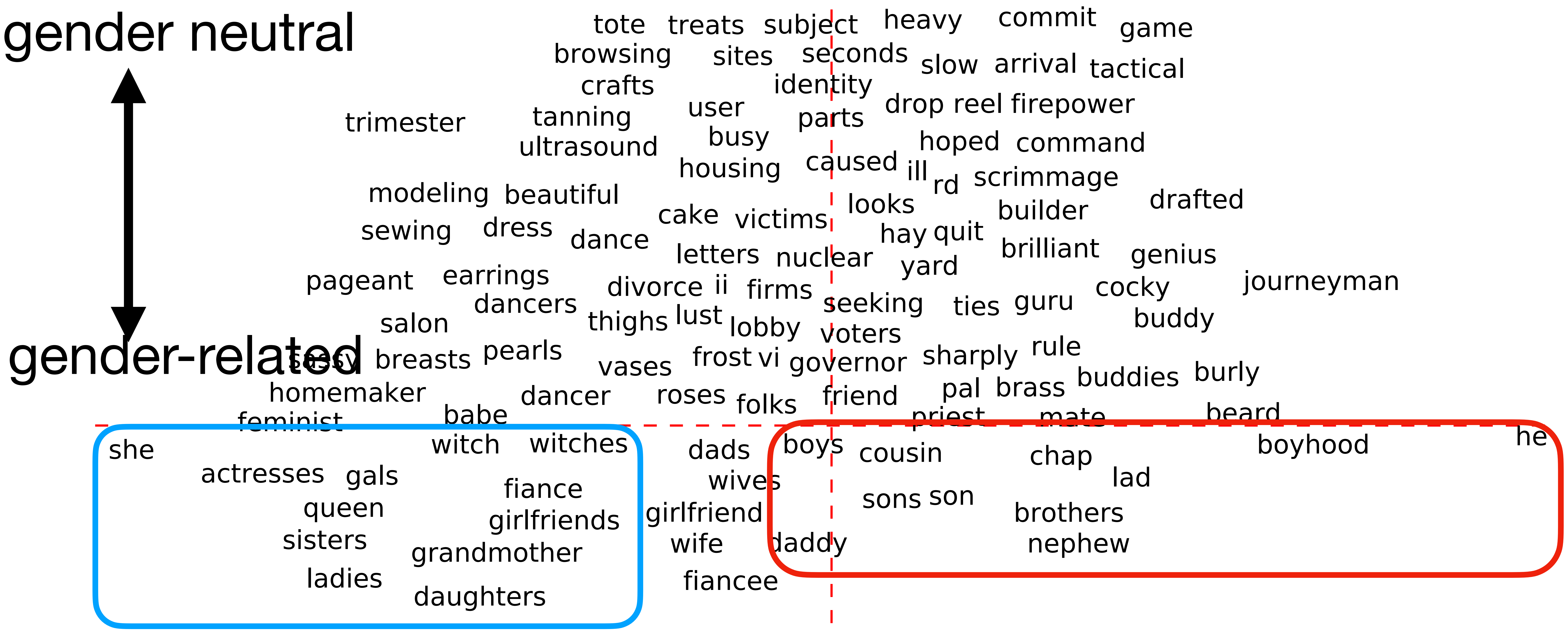
(b) Meaningful Dimensions (linear Space)

Park, Sungjoon, JinYeong Bak, and Alice Oh. "Rotated word vector representations and their interpretability." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

Topic 1: Embedding

What Do Word Embeddings Embed?

Previous papers mostly focus on word-level interpretations

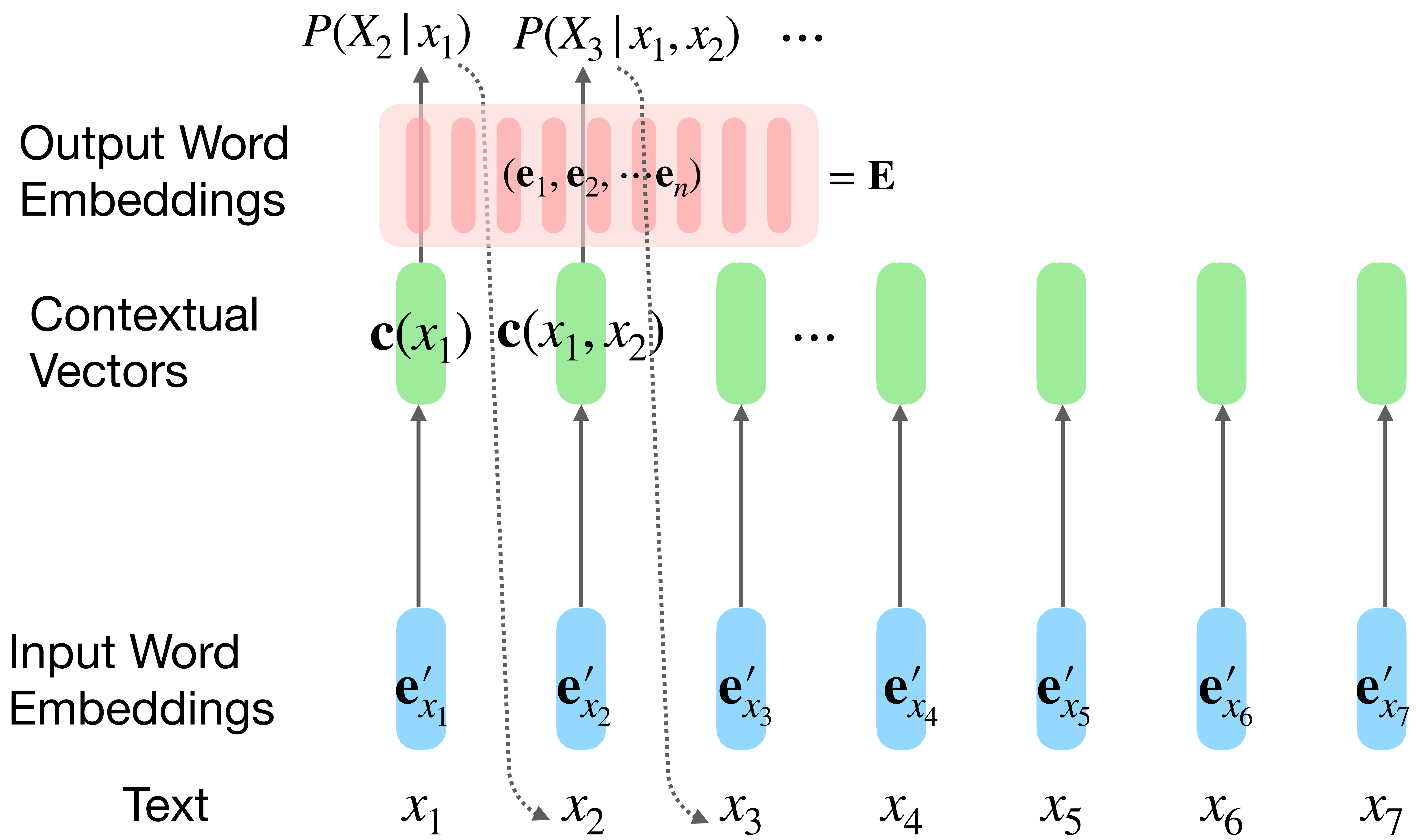


(b) Meaningful Dimensions (linear Space)

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

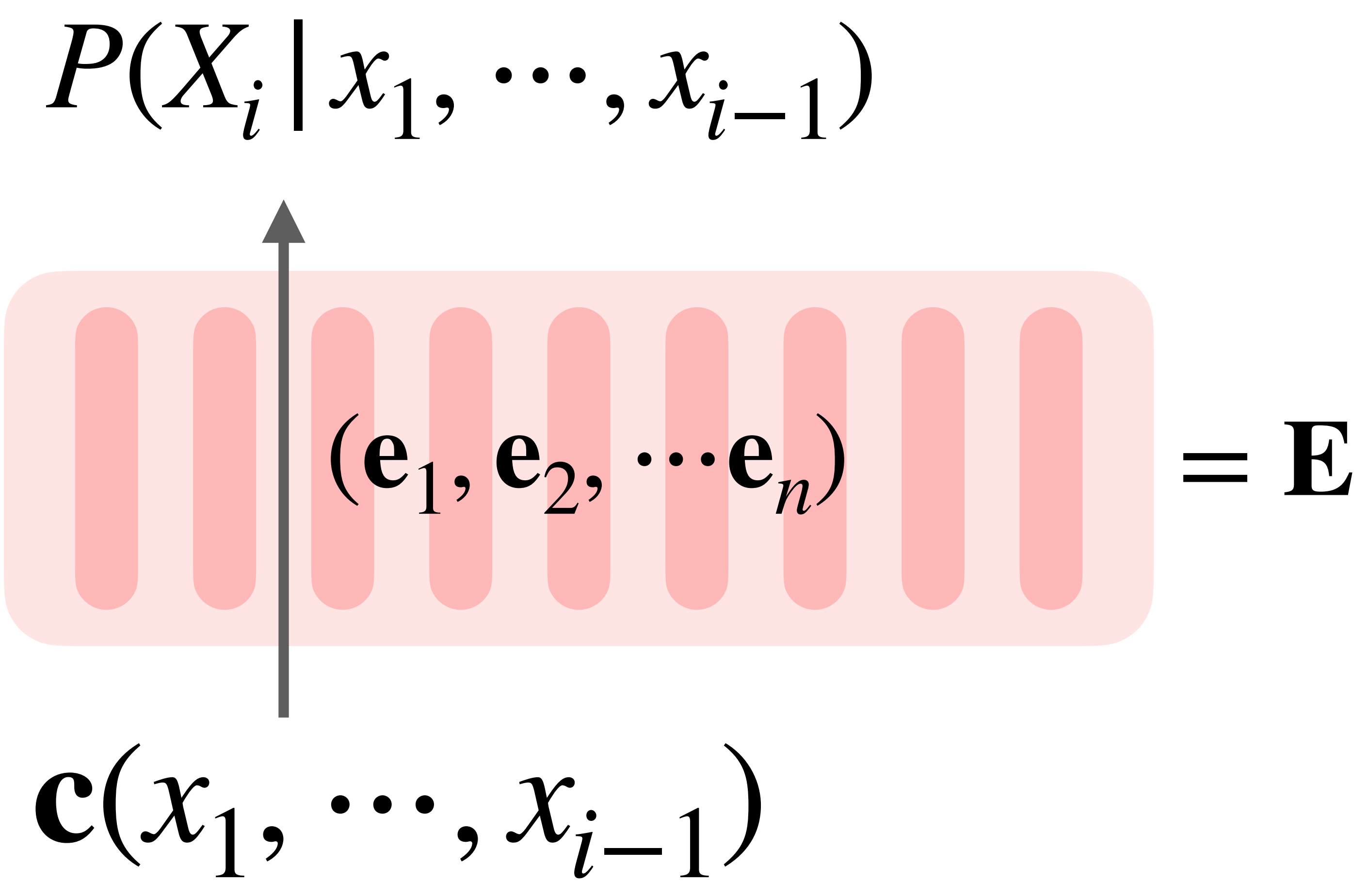
Topic 1: Embedding

Word Embeddings in Causal LMs



Output Word Embeddings

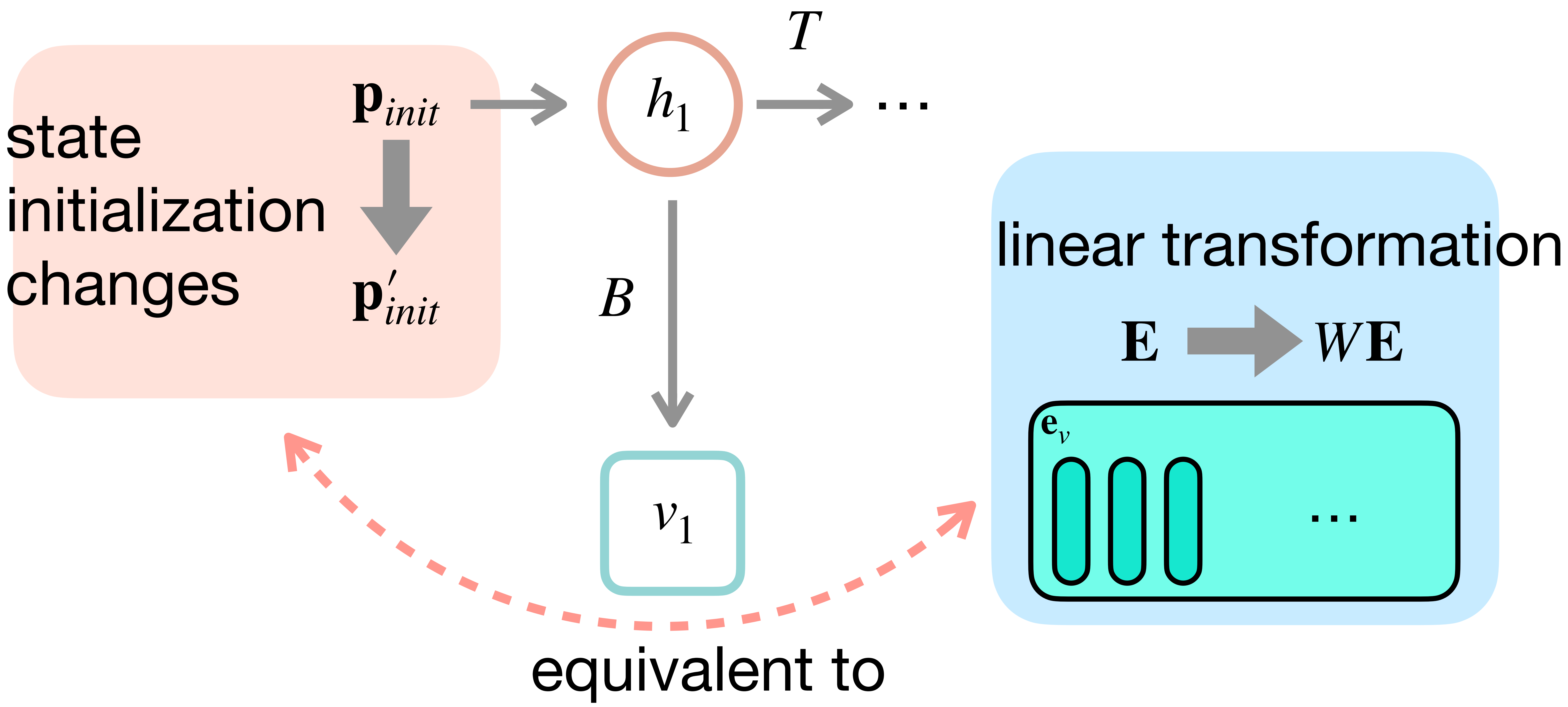
Projecting to Logits



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

Sequence Shift ≈ Word Embedding Transform

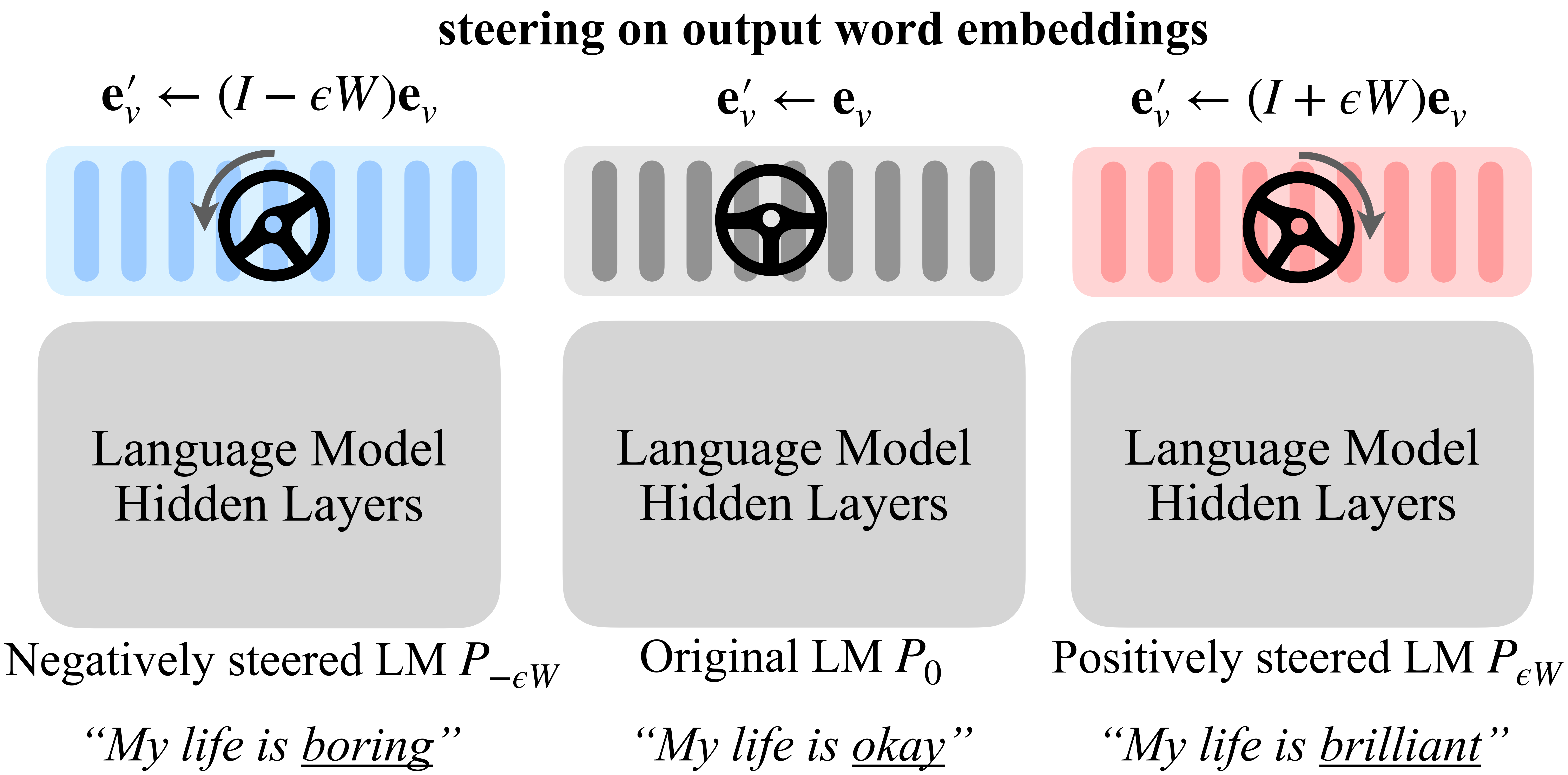
- **Theorem (Informal)**: steering between text distribution is associated with a linear transformation on word embedding space under assumptions.



Han, Chi, et al. "Word Embeddings Are Steers for Language Models." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

Topic 1: Embedding

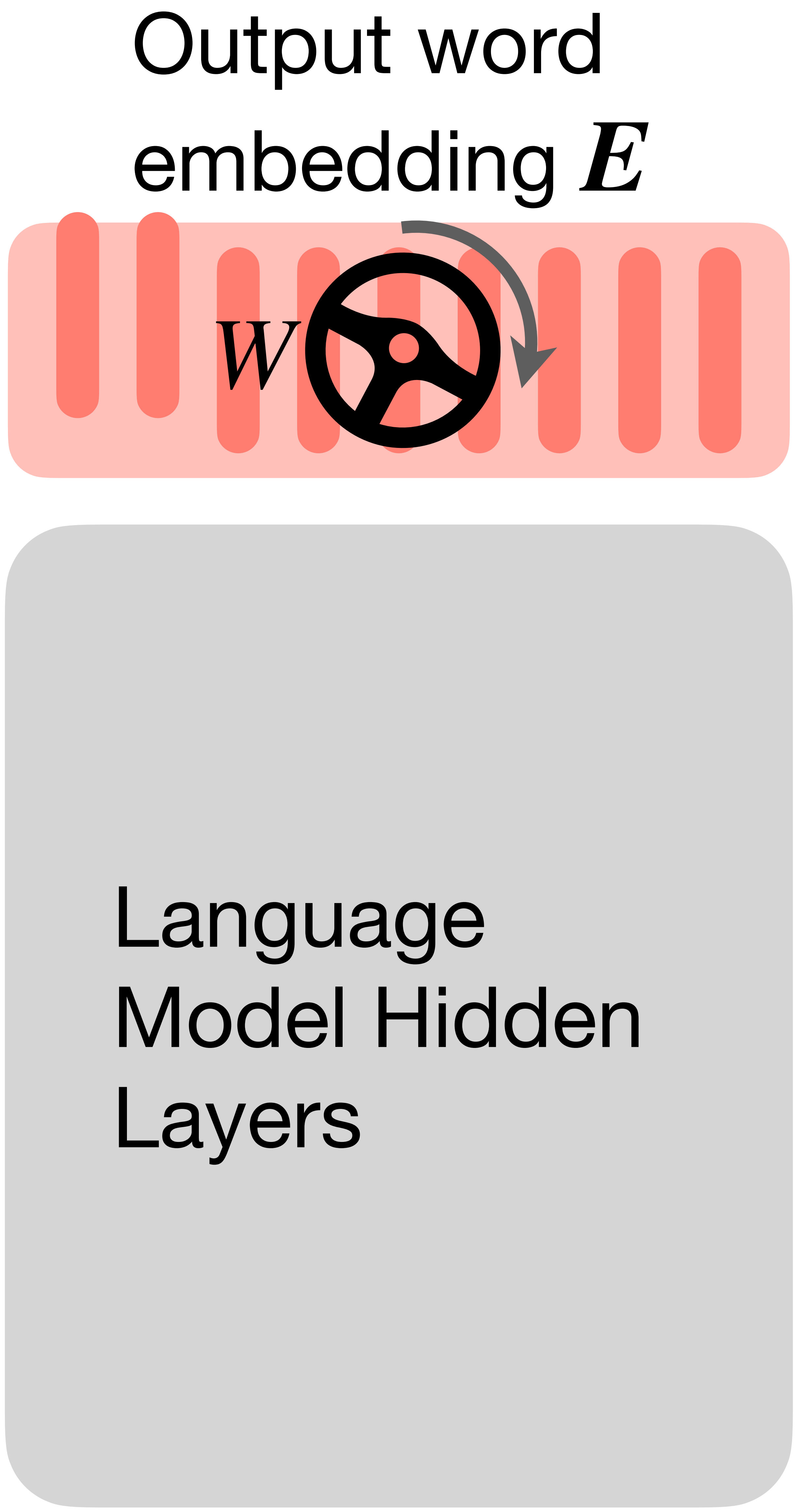
LM-Steer



Han, Chi, et al. "Word Embeddings Are Steers for Language Models." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

Topic 1: Embedding

LM-Steer Broken Down

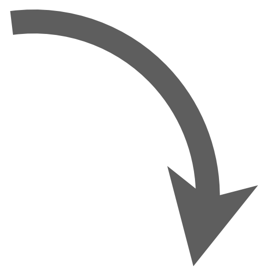


$$+= \epsilon \cdot W E$$

for each word:

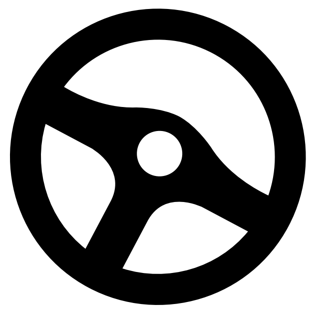
$$\mathbf{e}'_v = \mathbf{e}_v + \epsilon W \mathbf{e}_v$$

The steering scale

“  ”

ϵ

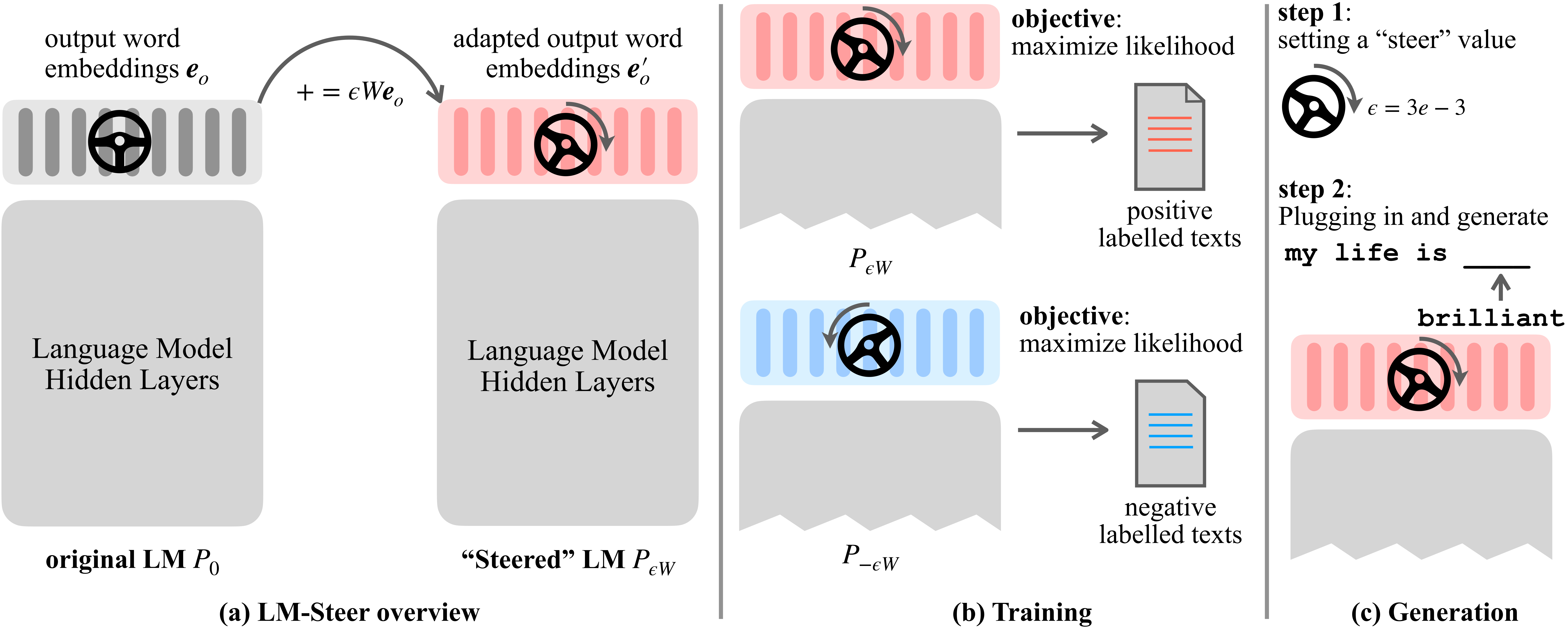
the steering matrix

“  ”

W

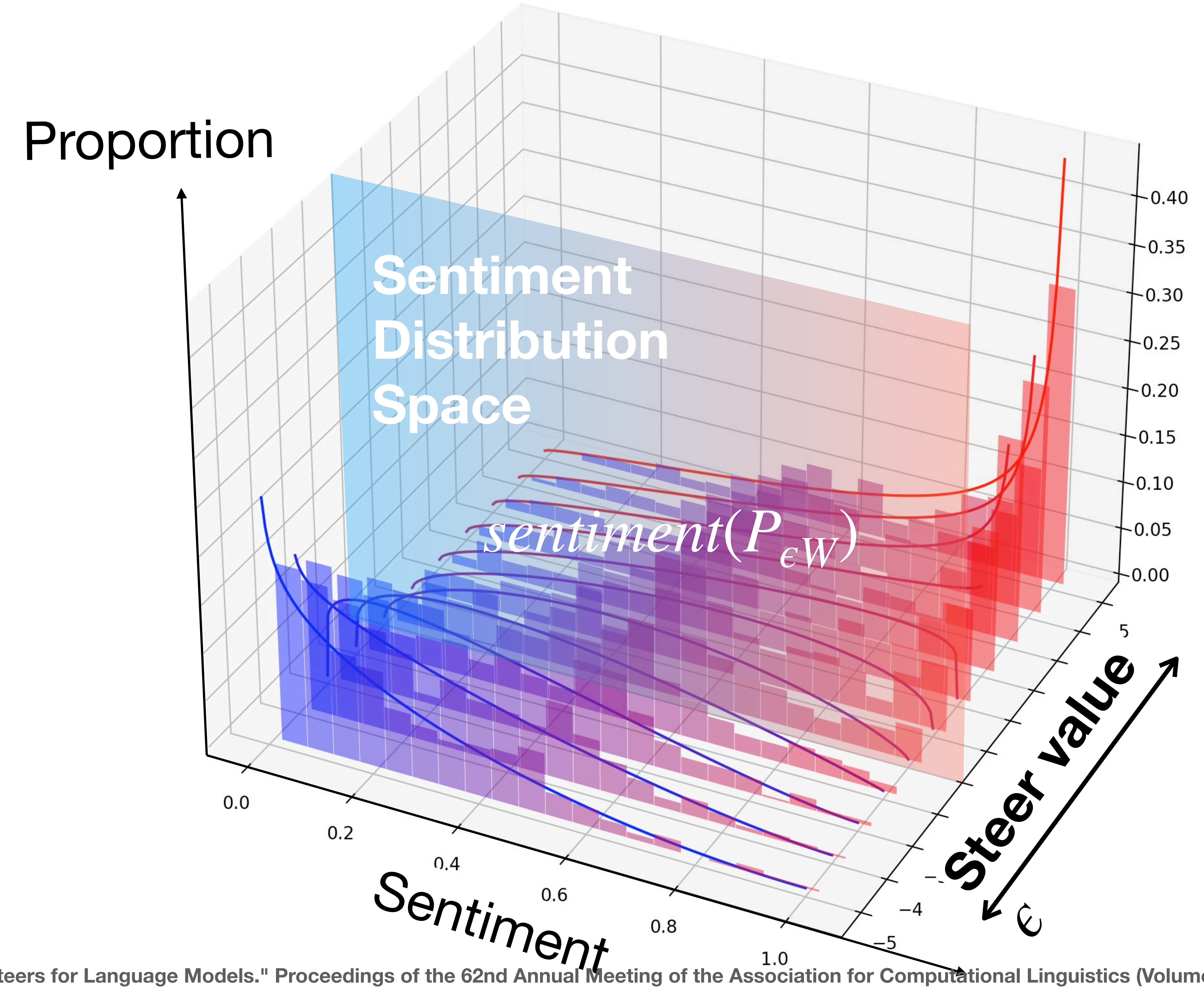
Topic 1: Embedding

Training & Inference



Han, Chi, et al. "Word Embeddings Are Steers for Language Models." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

Continuous Steering



curves: maximal likelihood beta-distribution

Han, Chi, et al. "Word Embeddings Are Steers for Language Models." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

Compositional Steering

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

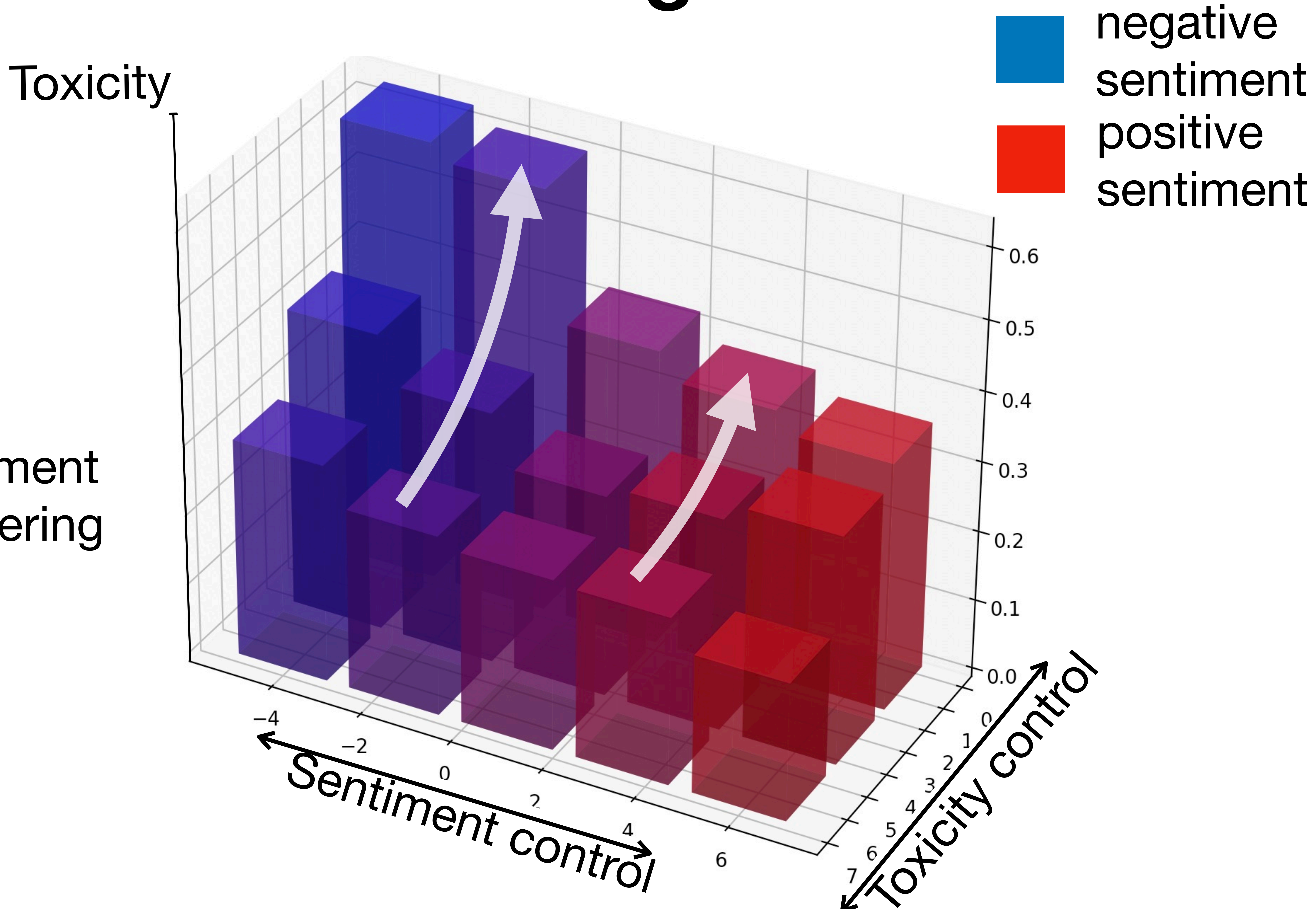
LM-Steer 2: $P_{\epsilon_2 W_2}$

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

Han, Chi, et al. "Word Embeddings Are Steers for Language Models." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

Topic 1: Embedding

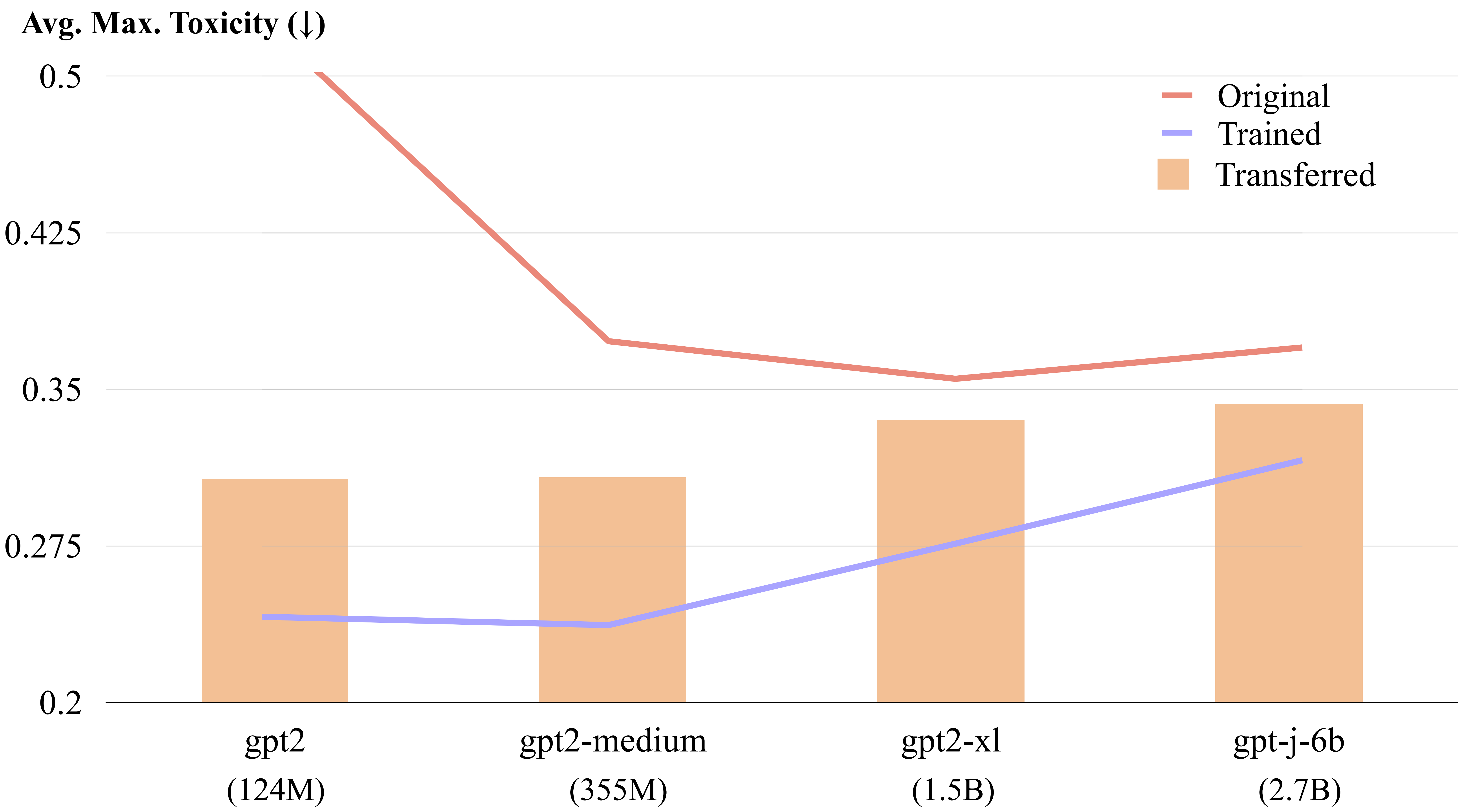
Compositional Steering



An entanglement between steering dimensions

Topic 1: Embedding

Transferring to Another LM



transfers about half of the detoxification capability

Han, Chi, et al. "Word Embeddings Are Steers for Language Models." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

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.

Han, Chi, et al. "Word Embeddings Are Steers for Language Models." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

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)

Han, Chi, et al. "Word Embeddings Are Steers for Language Models." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

Room for Future Research

- Evolution of contextual embeddings across layers, e.g., how ambiguity is resolved in LMs
- Better frameworks for studying the role of word embeddings
- Other functions of word embeddings, such as semantics and sense

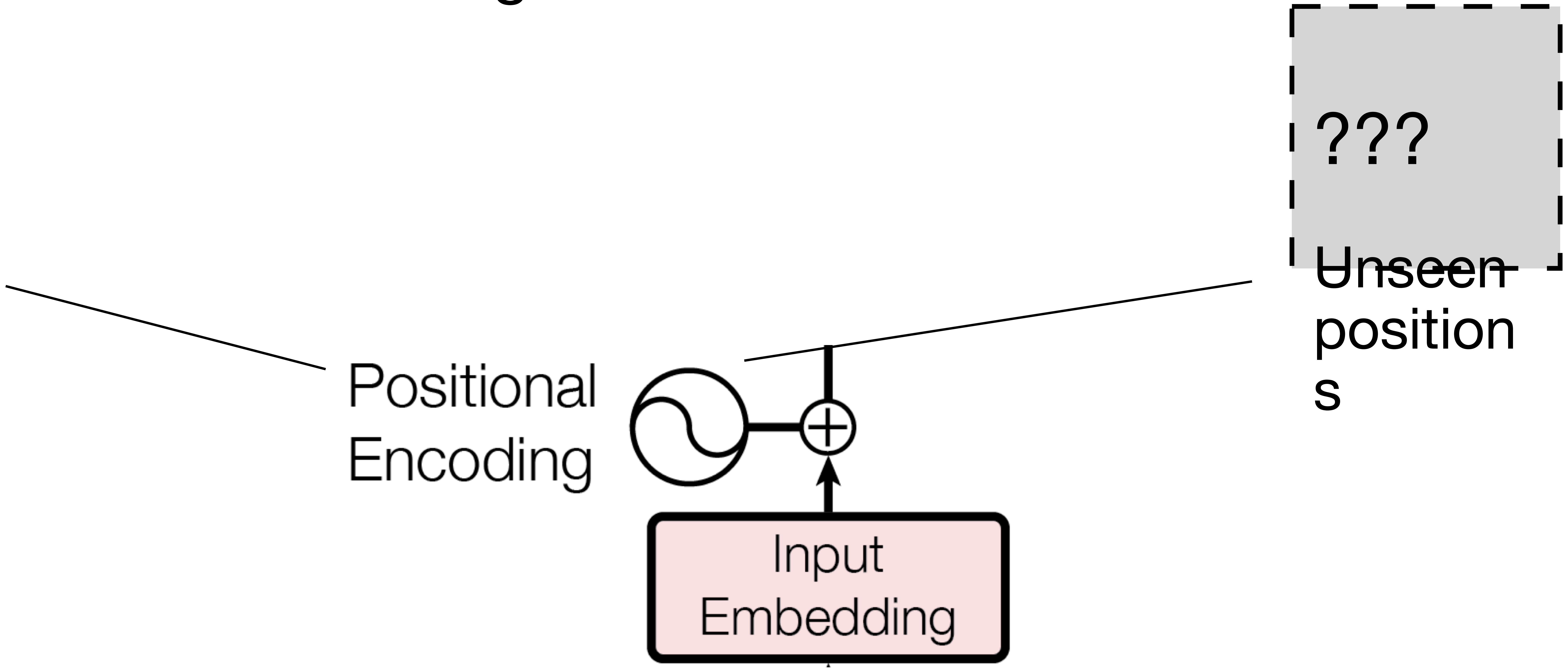
Attention, Position and Context

Questions:

1. How LMs Deal with Context Length
2. How LMs Process Position Information
3. How LMs Comprehend Contextual Knowledge

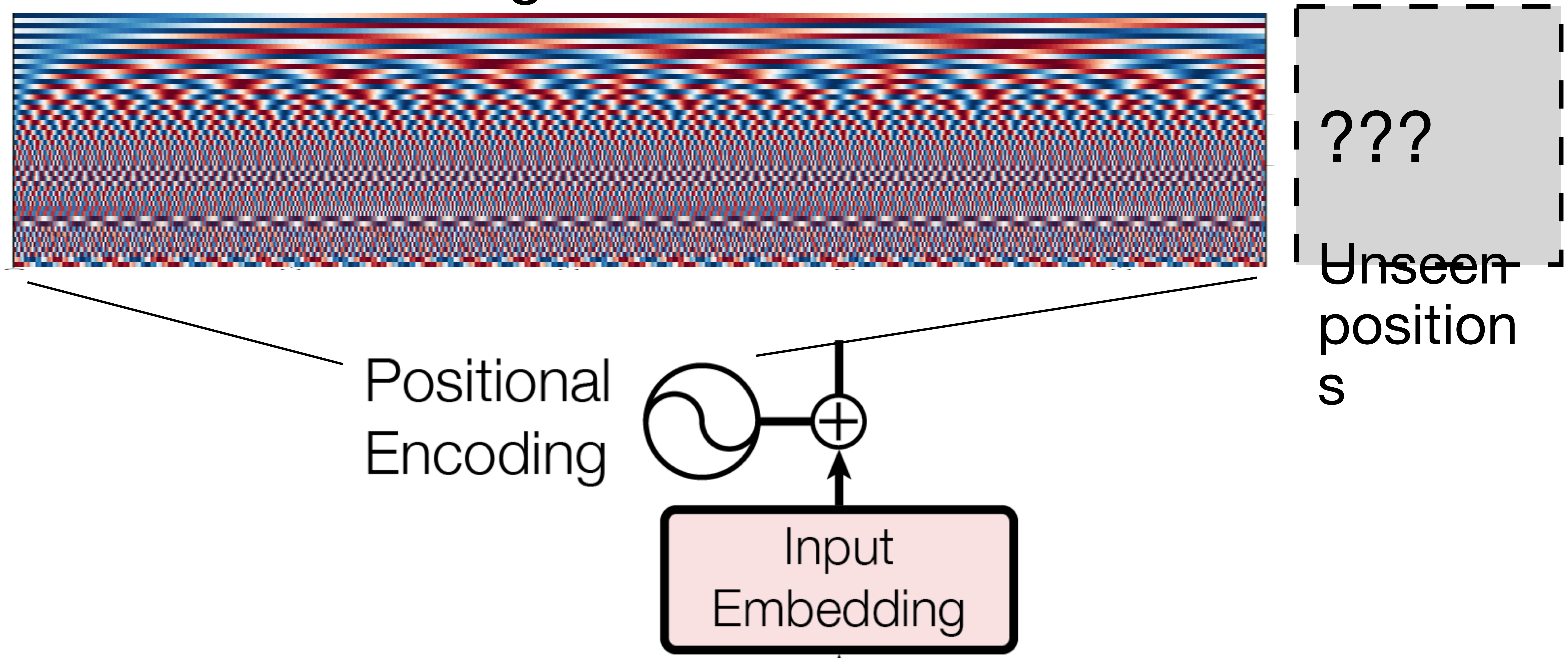
Absolute Positional Encoding: ❌

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



Absolute Positional Encoding: ❌

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



https://erdem.pl/2021/05/understanding-positional-encoding-in-transformers

Relative Positional Encoding: ?

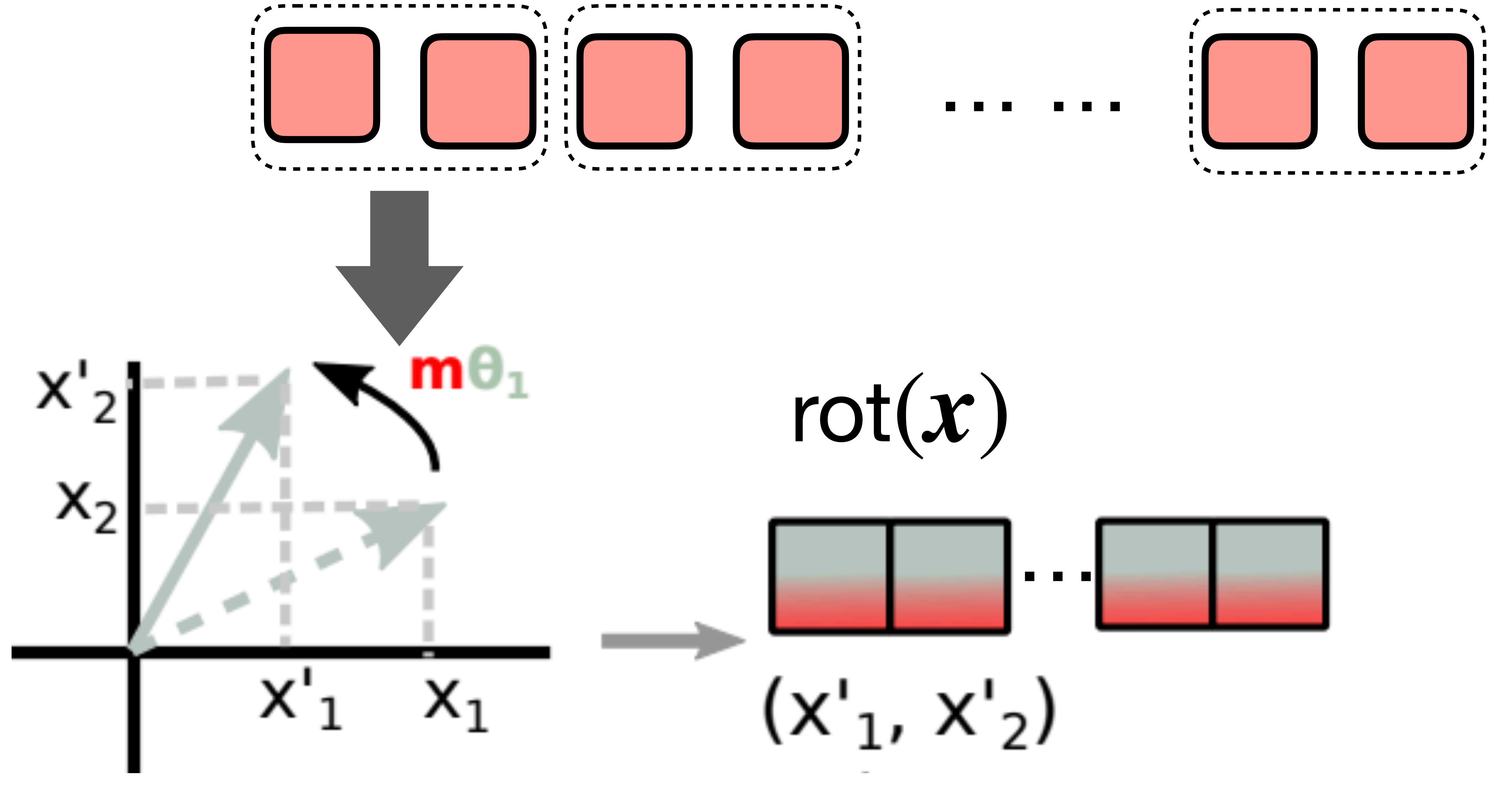
Relative positional encoding was proposed in the hope to alleviate this problem

Core idea: determining attention based on distance

RoPE:

$$\boldsymbol{x} = (x_1,\ x_2,\ x_3,\ x_4,\ \cdots\cdots\ x_{d-1}, x_d)$$

(Used in LLaMA, Llama-2, GPT-J, etc.)



Su, Jianlin, et al. "Roformer: Enhanced transformer with rotary position embedding." arXiv preprint arXiv:2104.09864 (2021).

Relative Positional Encoding: ?

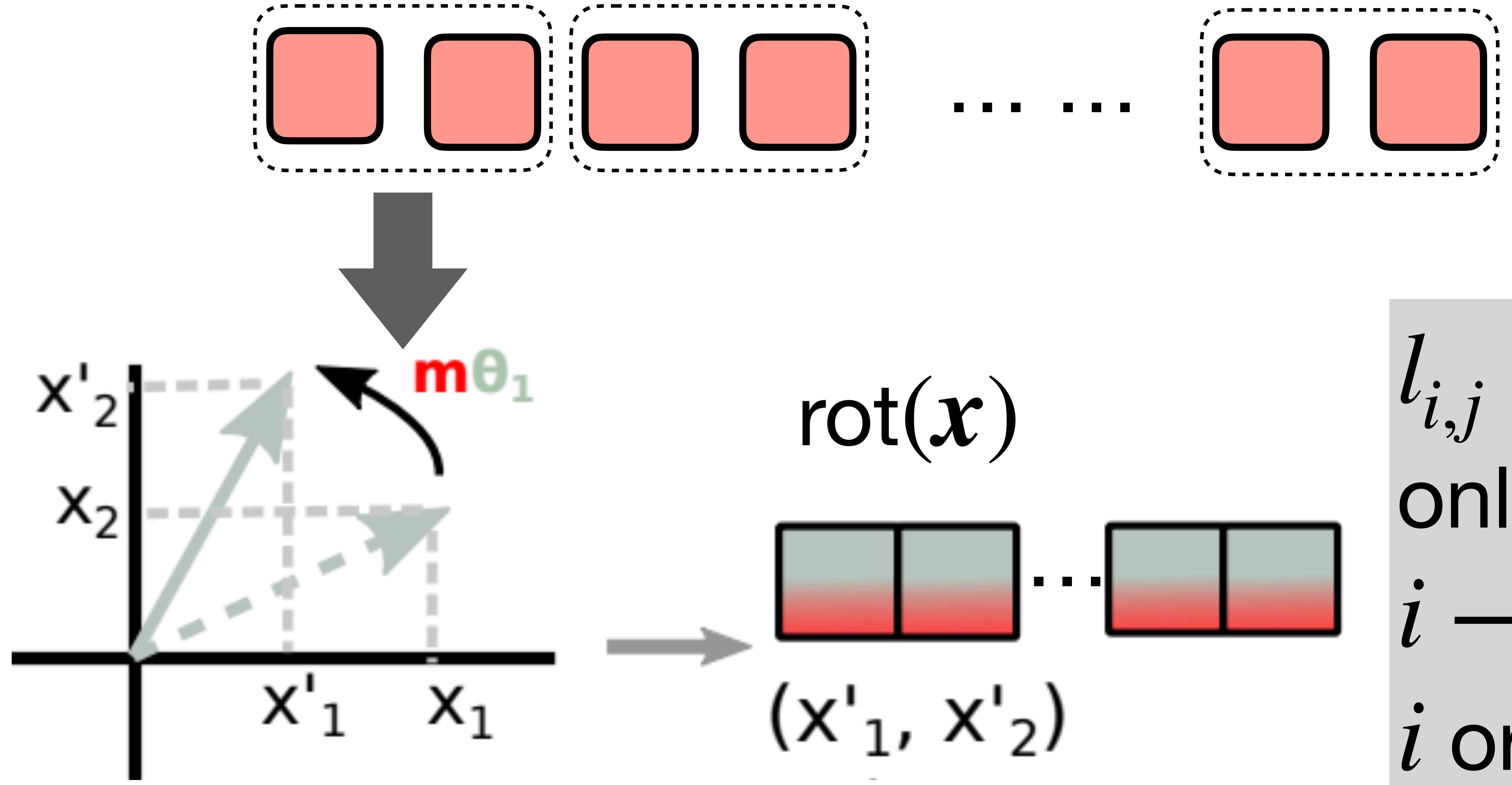
Relative positional encoding was proposed in the hope to alleviate this problem

Core idea: determining attention based on distance

$$\boldsymbol{x} = (x_1, \; x_2, \; x_3, \; x_4, \; \cdots \cdots \; x_{d-1}, x_d)$$

RoPE:

(Used in LLaMA, Llama-2, GPT-J, etc.)



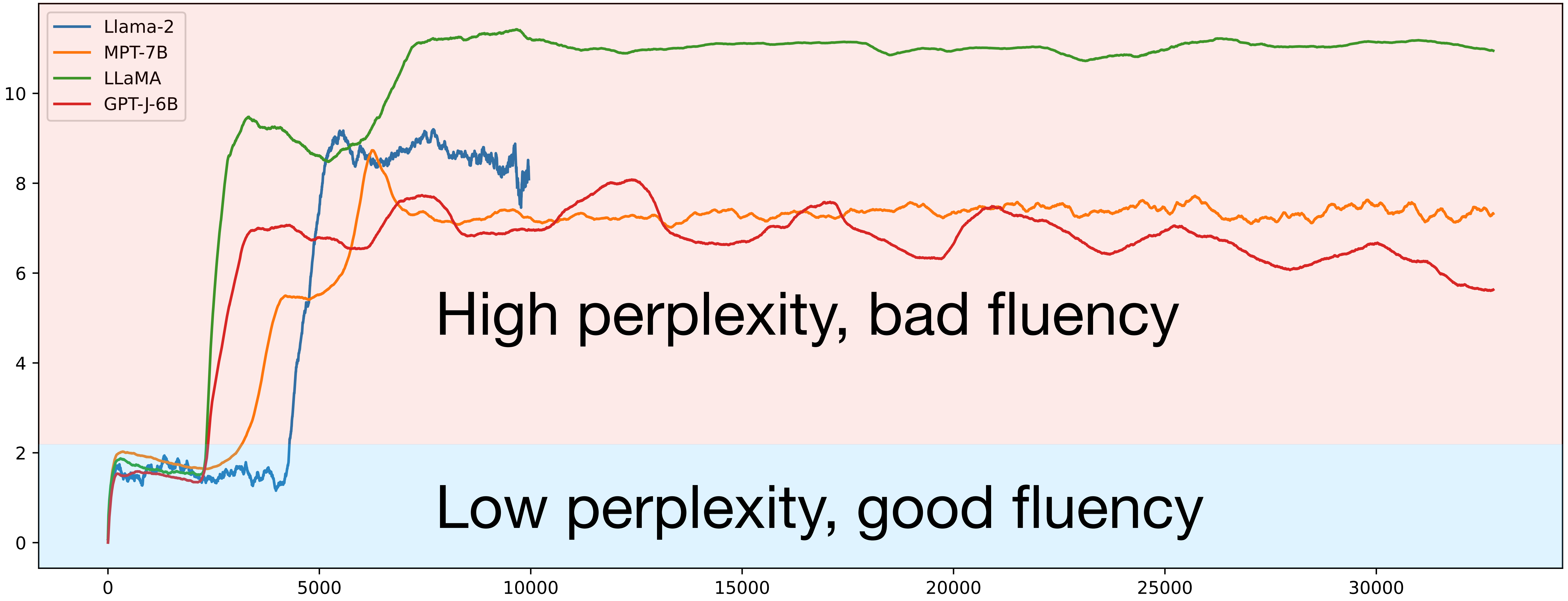
$l_{i,j} = \text{rot}(\boldsymbol{q}_i)^\top \text{rot}(\boldsymbol{k}_j)$ only depends on $i - j$, regardless of i or j.

Su, Jianlin, et al. "Roformer: Enhanced transformer with rotary position embedding." arXiv preprint arXiv:2104.09864 (2021).

Relative Positional Encoding: ?

However, current LLMs still struggle on unseen lengths.

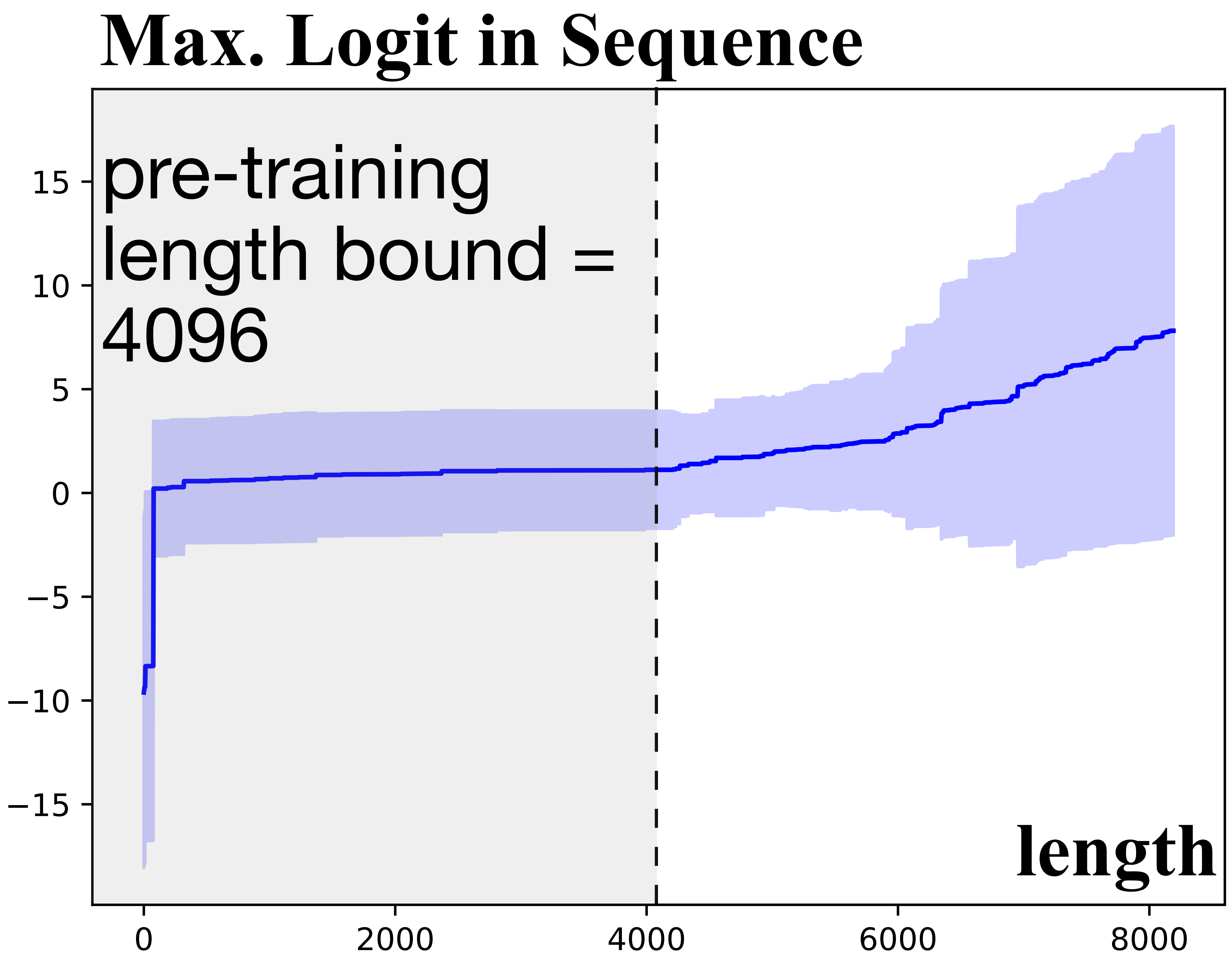
Negative Log-Likelihood (NLL, also =log(perplexity)**)** ↓



length

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.

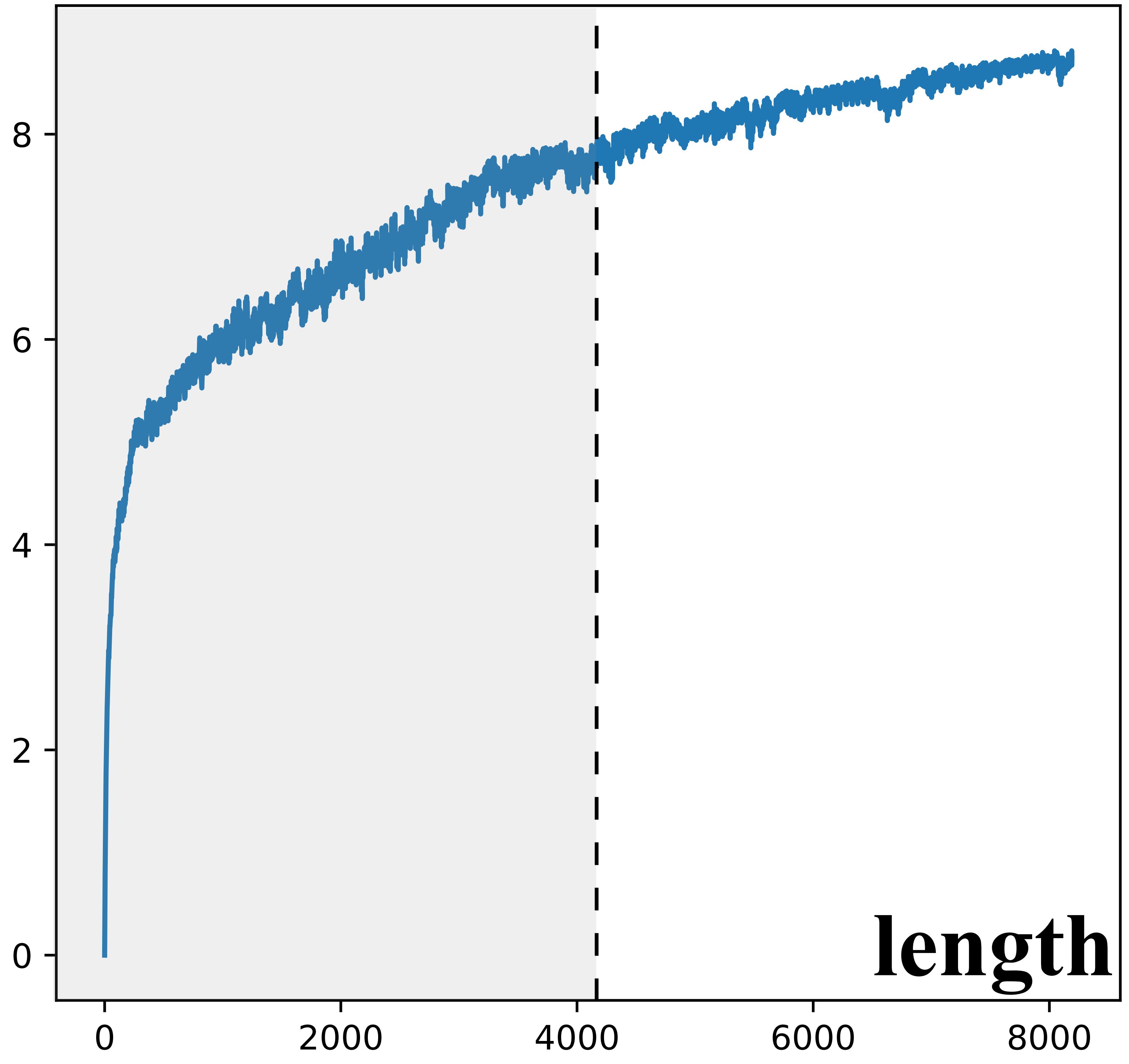
Han, Chi, et al. "LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

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.

Attention Entropy

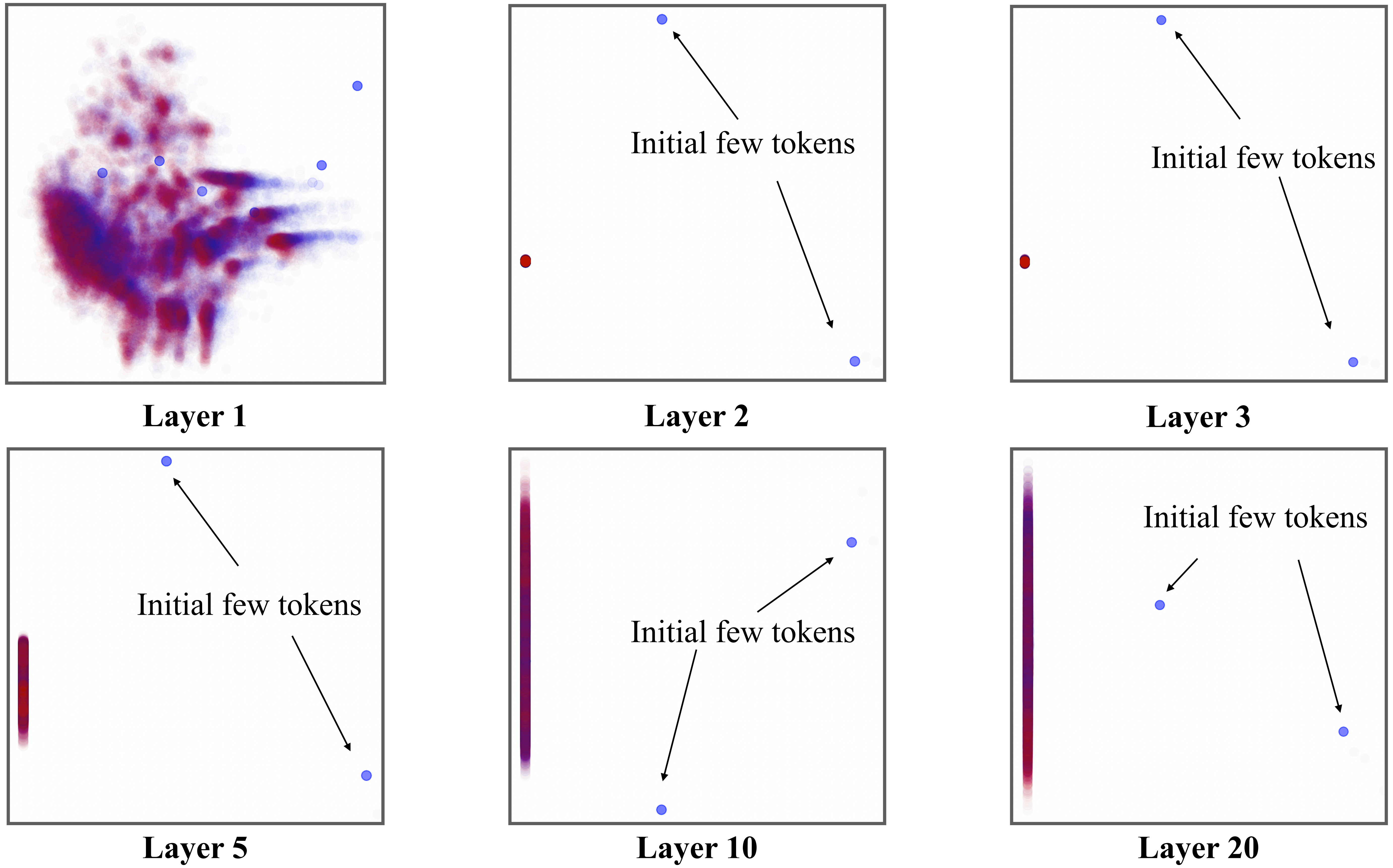


The entropy of attention distribution in Llama-2 continuously increases with length.

Han, Chi, et al. "LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

Factor 3: Implicitly Encoded Position

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



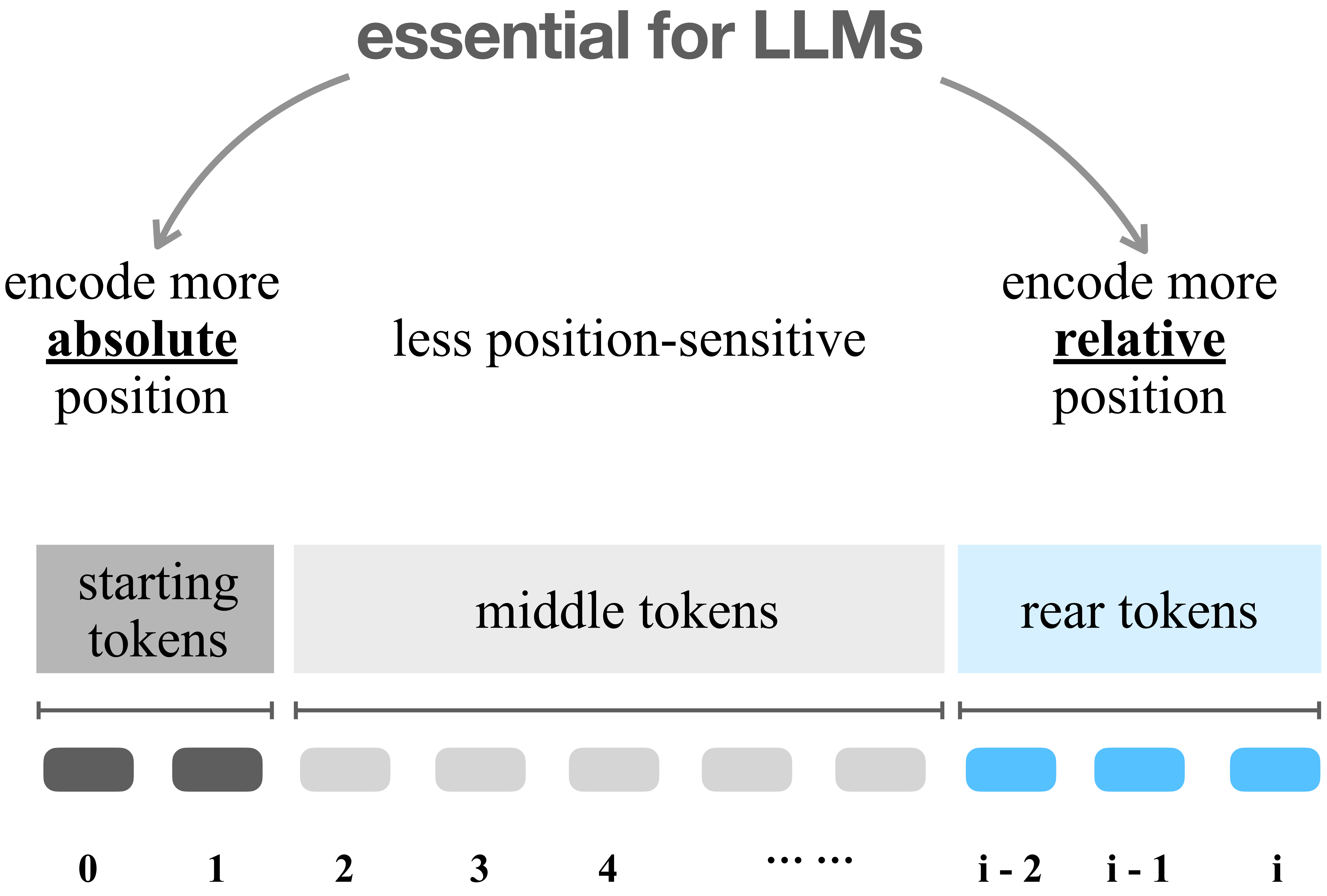
Theorem 3 (Informal): Even without absolute positional embeddings, attention can restore position information of tokens.

Han, Chi, et al. "LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

Kazemnejad, Amirhossein, et al. "The impact of positional encoding on length generalization in transformers." *Advances in Neural Information Processing Systems* 36 (2023): 24892-24928.

Topic 2: Attention - Question 1: Length

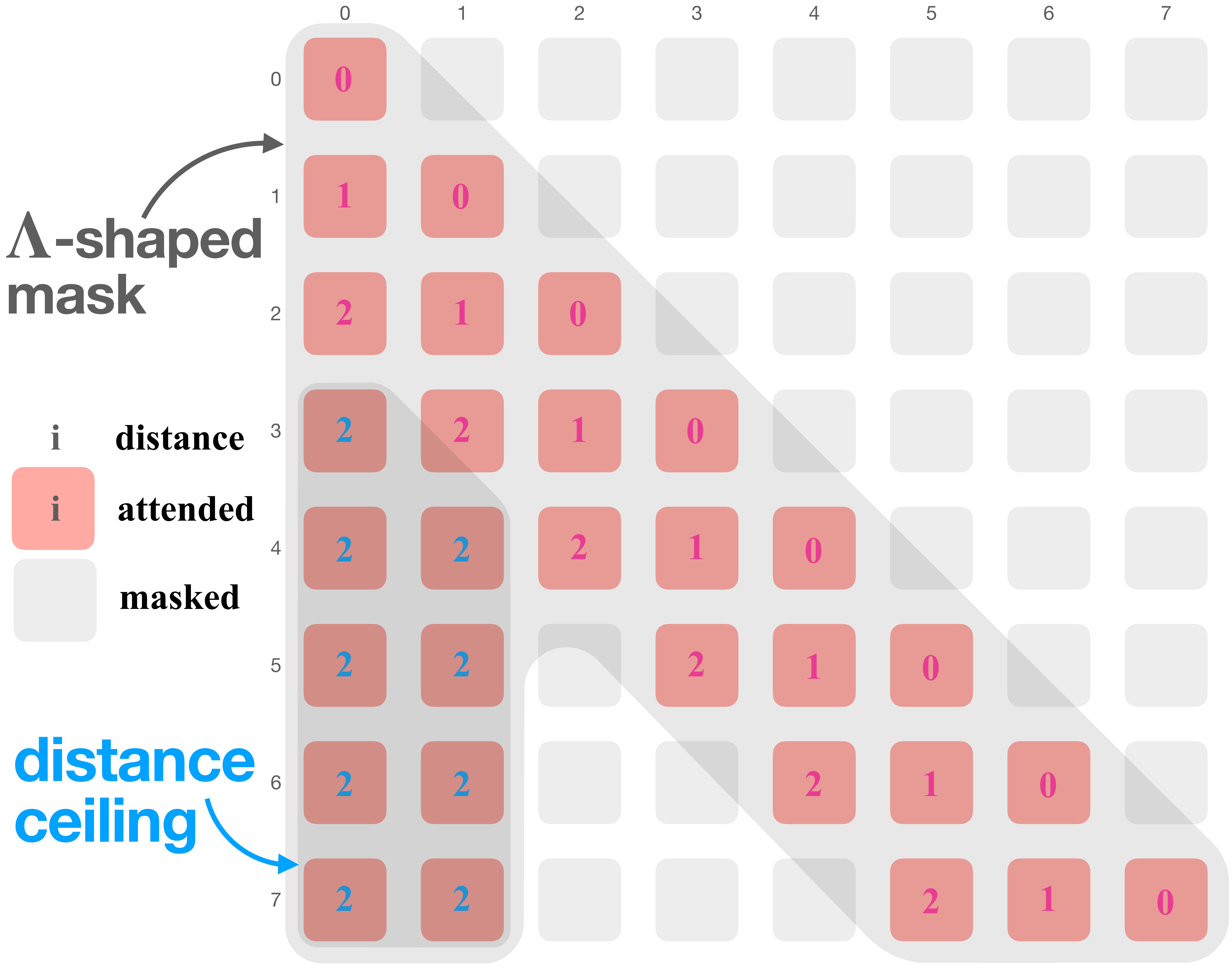
A Conceptual Model of Relative Position Encoding



Han, Chi, et al. "LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

Topic 2: Attention - Question 1: Length

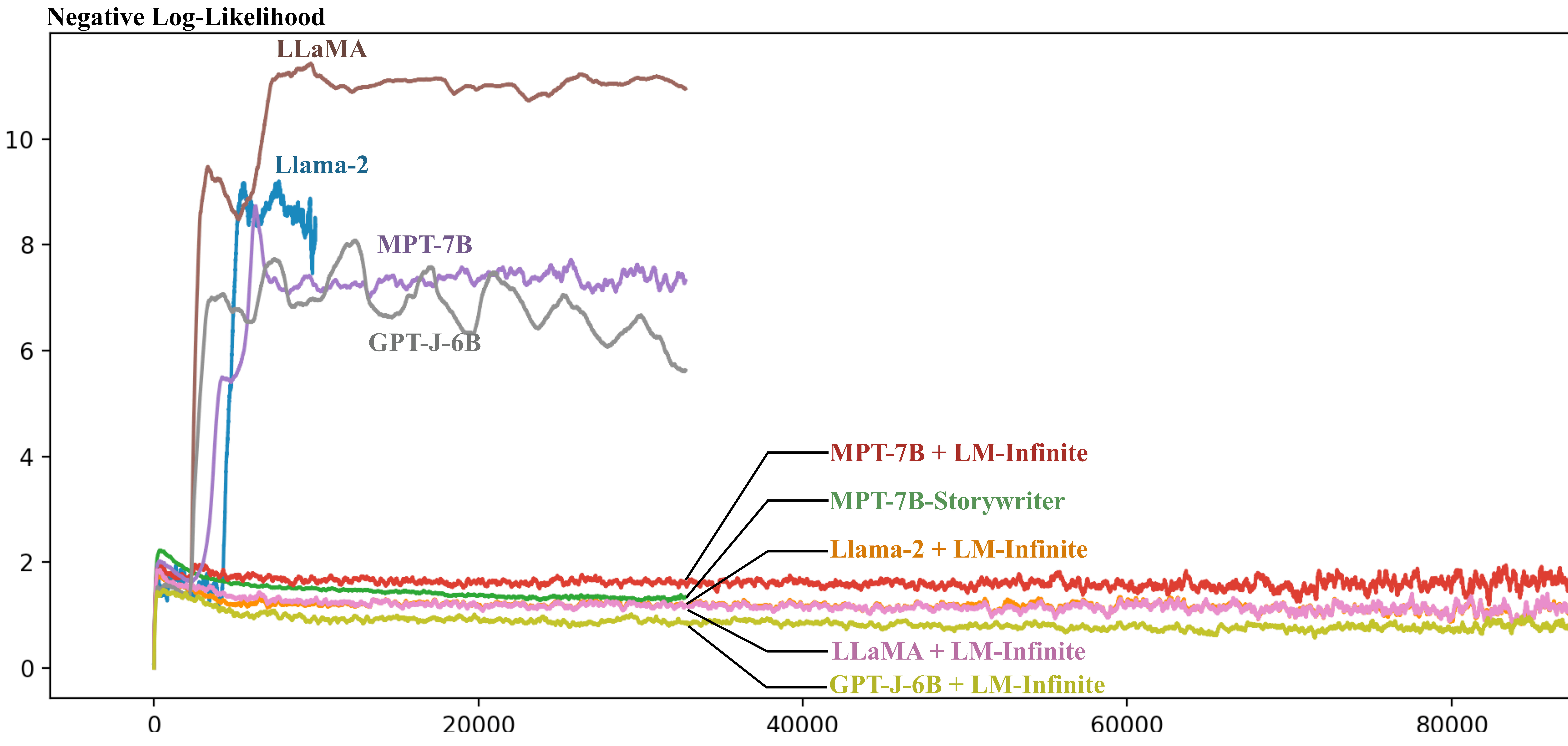
Solution: LM-Infinite



Han, Chi, et al. "LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024. (Outstanding Paper Award)

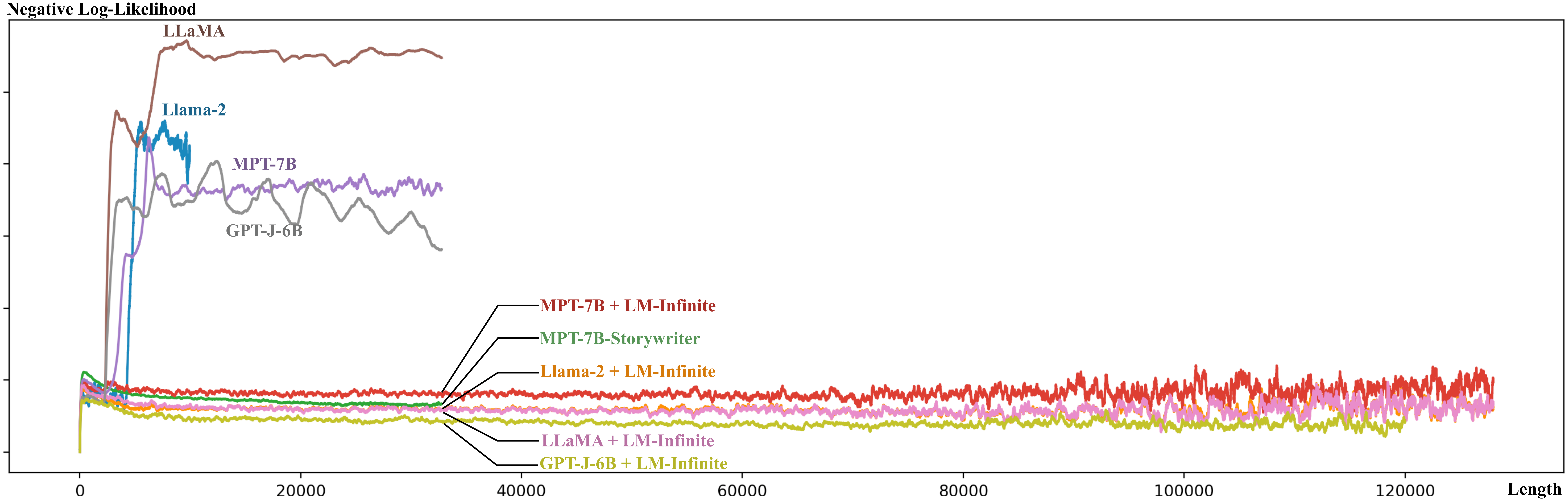
Topic 2: Attention - Question 1: Length

Length Generalization (to 200M length)



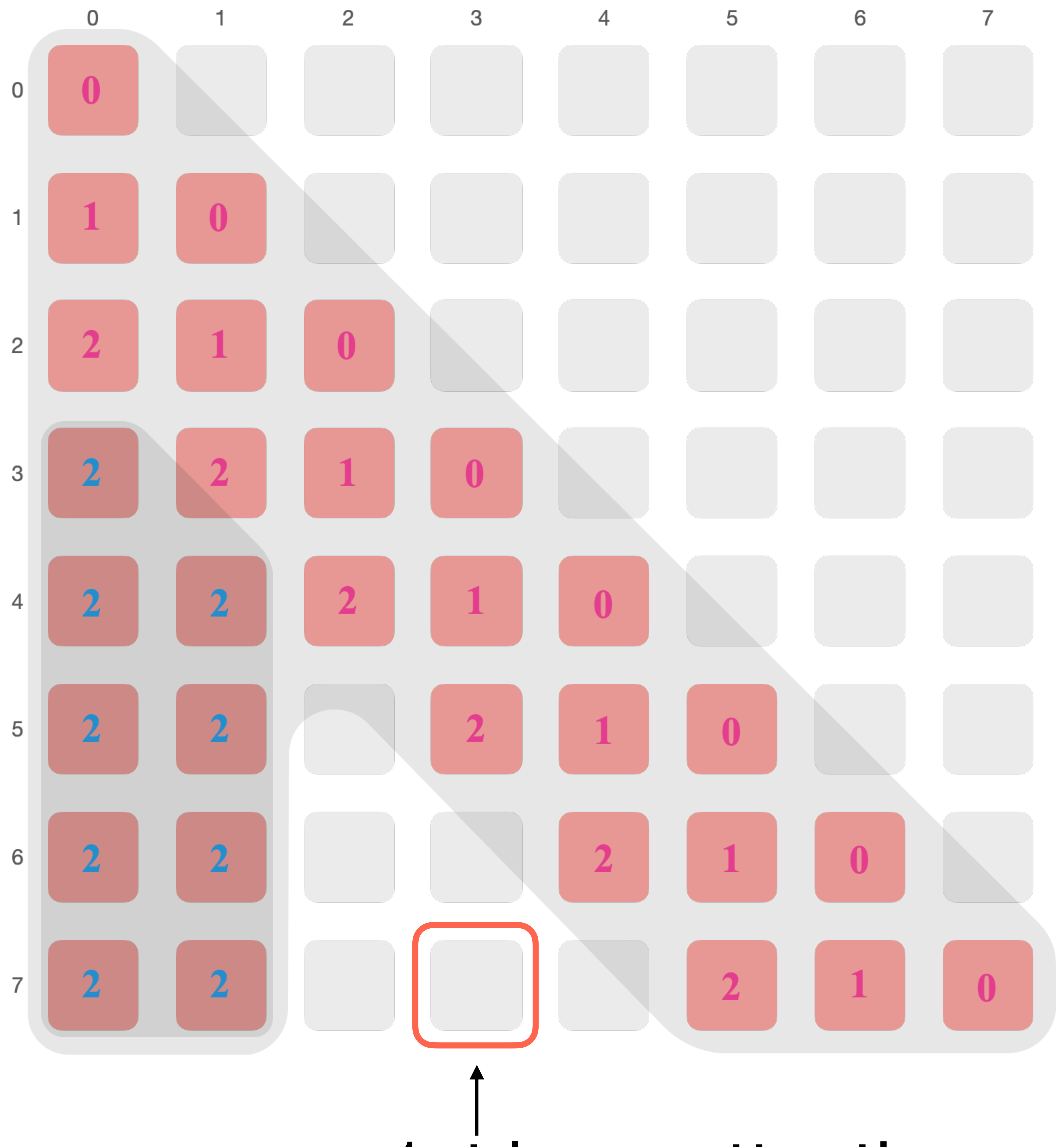
Topic 2: Attention - Question 1: Length

Length Generalization (to 200M length)



To Perceive Sensitive Information

Re-attending to top-k attention tokens



e.g. 1st large attention

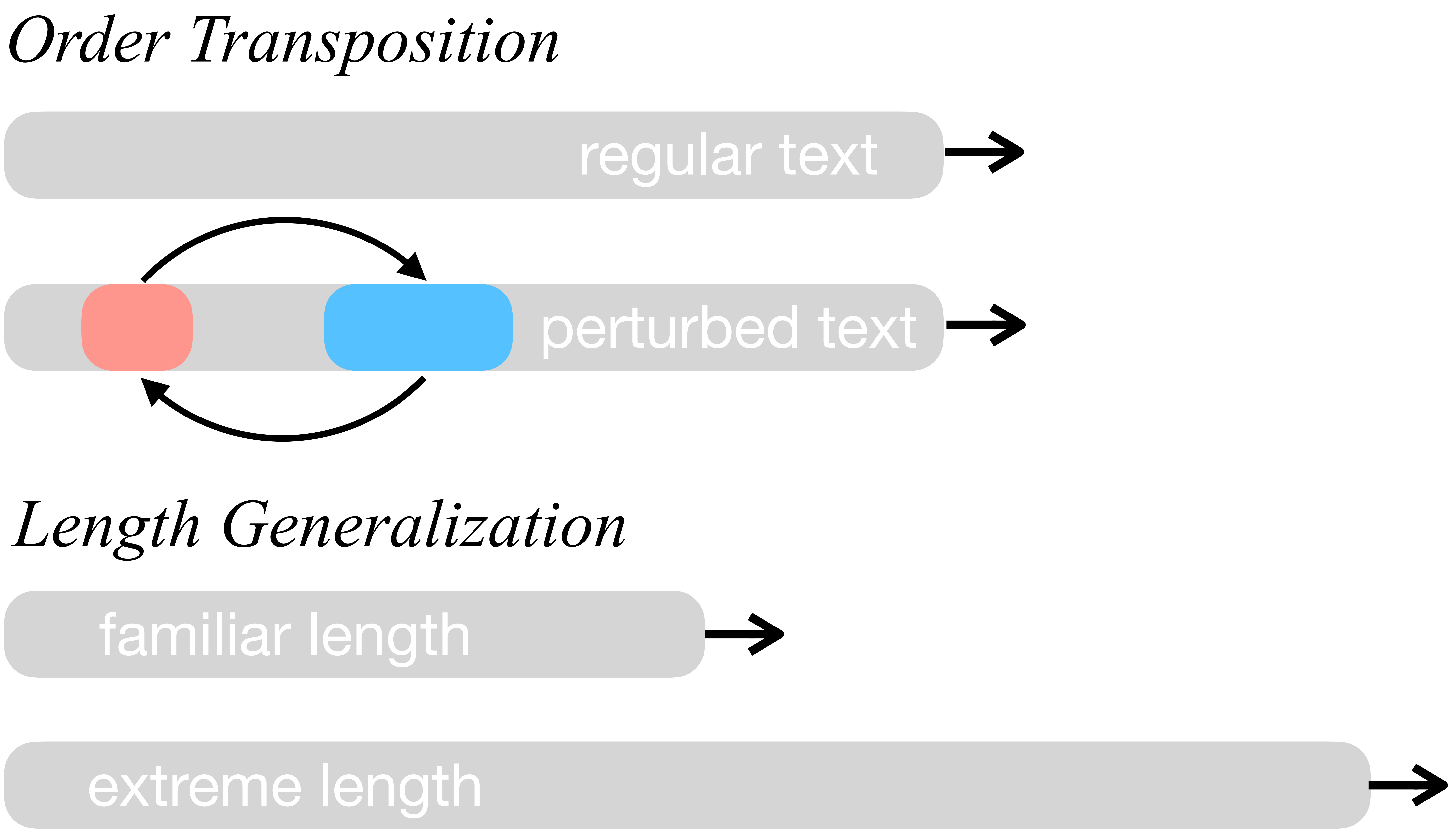
Why: to acquire key information that might be stored in the middle "ignored" region again.

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.

Positional Generalization Phenomenon

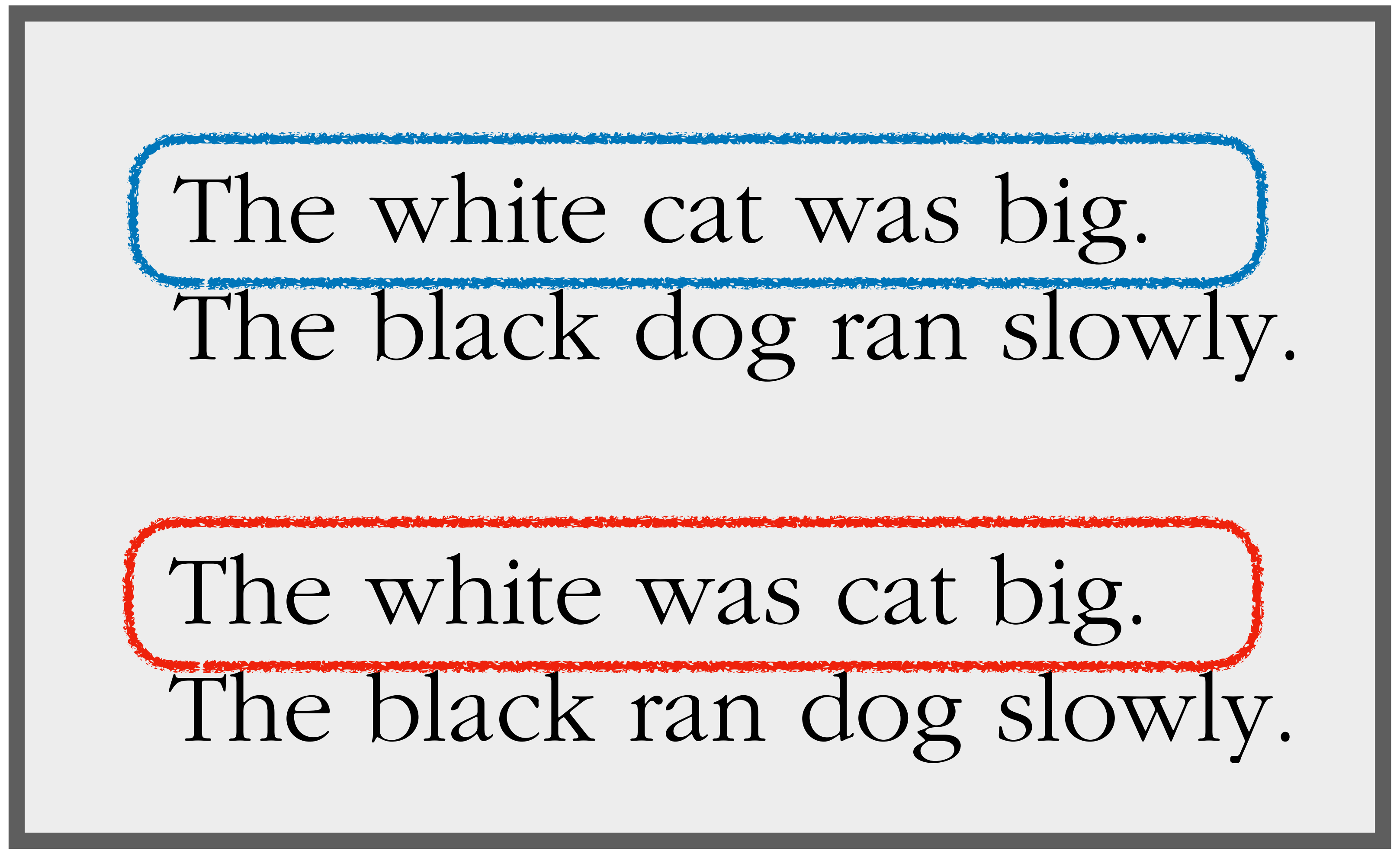
of both humans and language models

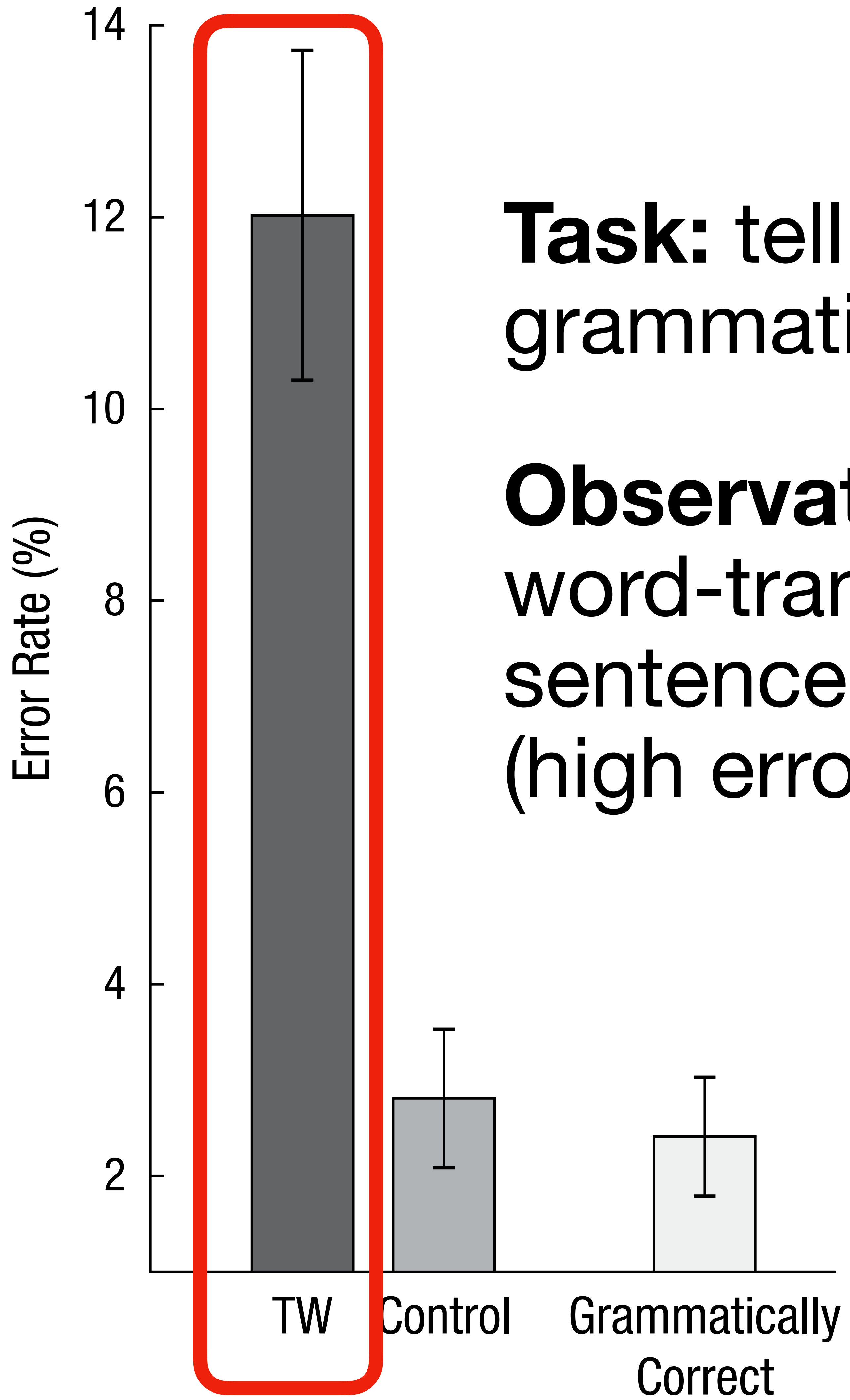


Han, Chi, et al. "Computation Mechanism Behind LLM Position Generalization" *arXiv preprint arXiv:2503.13305 (2025)*

Humans' Positional Generalization

Task: is the new sentence grammatically correct?





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)

Mirault, Jonathan, Joshua Snell, and Jonathan Grainger. "You that read wrong again! A transposed-word effect in grammaticality judgments." Psychological Science 29.12 (2018): 1922-1929.

Topic 2: Attention - Question 2: Position

LMs Can Understand Perturbed Language

Task: paraphrase

if two sentences are duplicate

Q_1 Does marijuana cause cancer?

Q_2 How can smoking marijuana give you lung cancer?

(a) Prediction: "duplicate" 0.96

Q_1 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

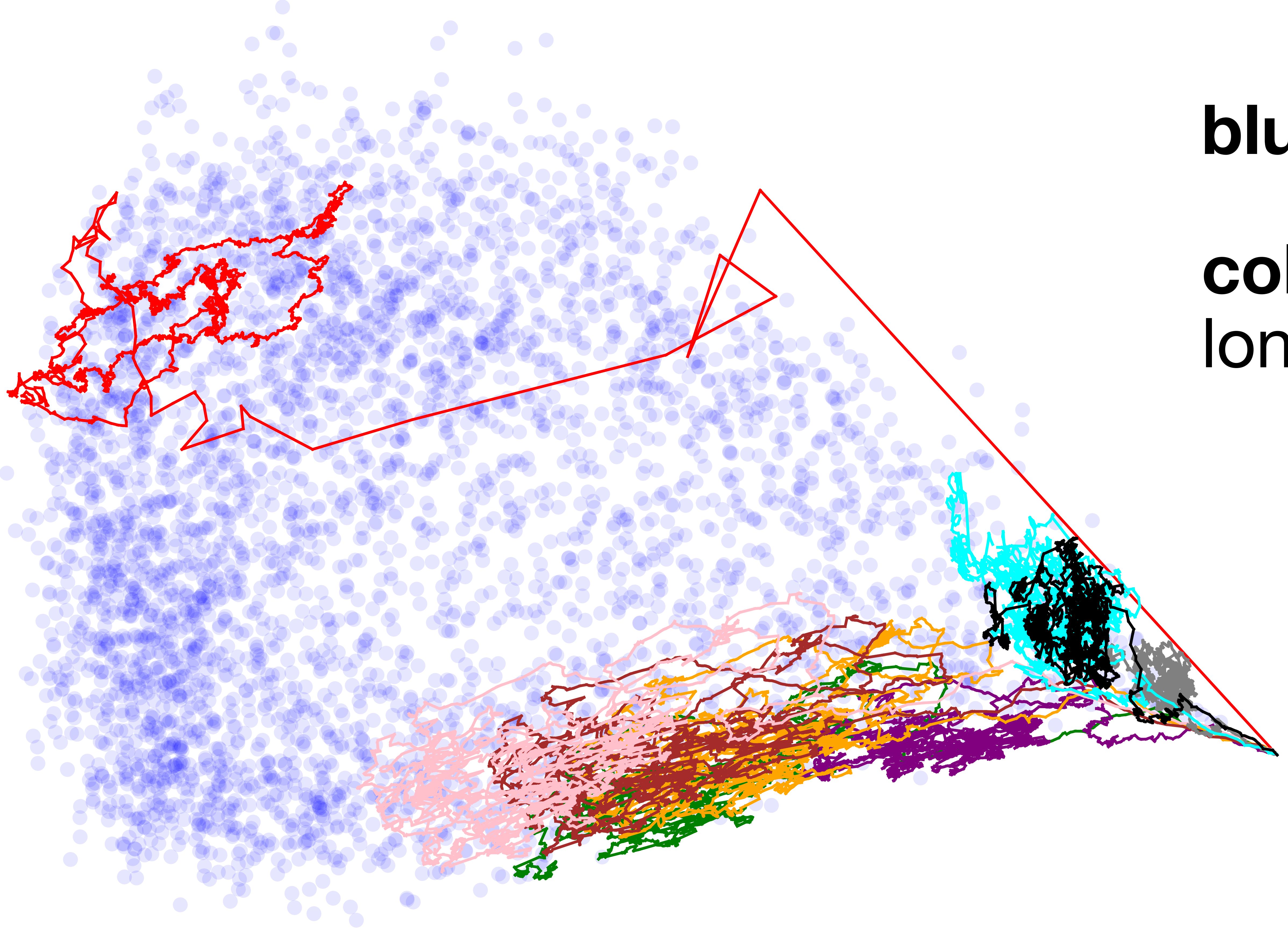
Task: entailment

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.	

Pham, Thang, et al. "Out of Order: How important is the sequential order of words in a sentence in Natural Language Understanding tasks?." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021.

In-Distribution Features under Length-Generalization

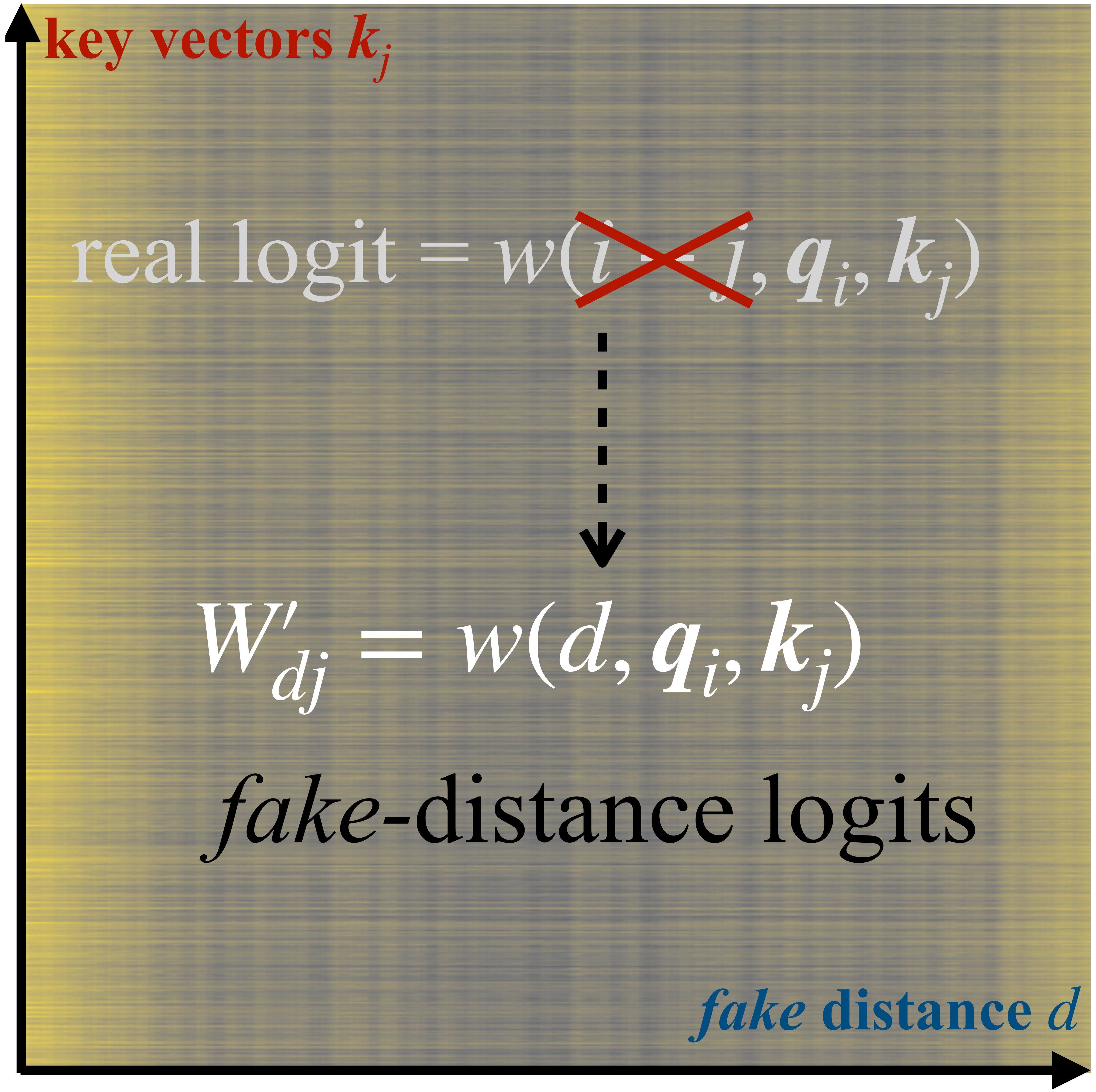


blue dots: normal features

colored lines: token features of in super-long context under length generalization

Han, Chi, et al. "Computation Mechanism Behind LLM Position Generalization" *arXiv preprint arXiv:2503.13305 (2025)*

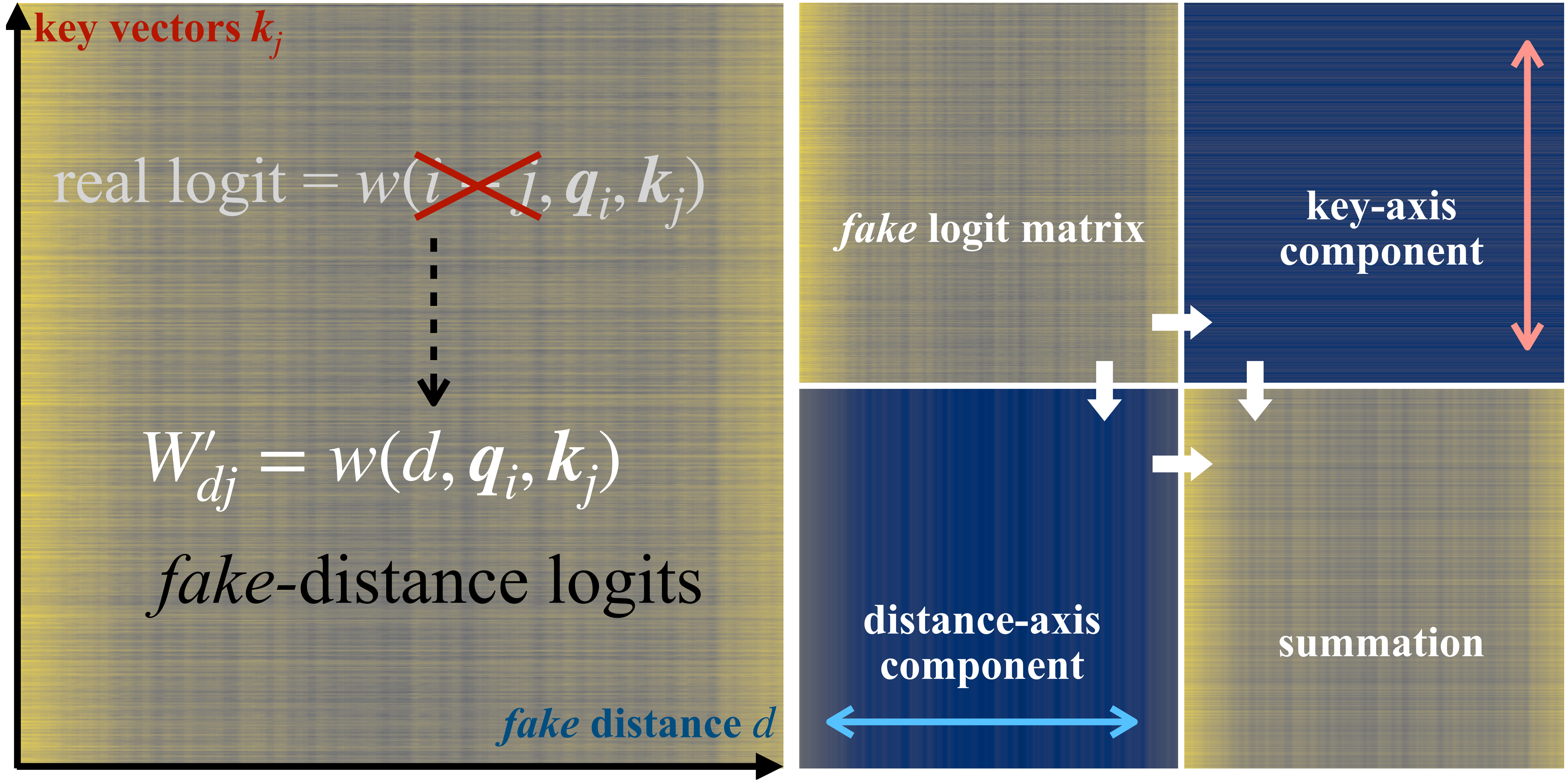
LM Attention Decomposes Position and Semantics



to isolate the effect of vectors $\boldsymbol{q}_i, \boldsymbol{k}_j$ and their distance $i - j$, let us use a "fake" distance d instead of $i - j$

Han, Chi, et al. "Computation Mechanism Behind LLM Position Generalization" *arXiv preprint arXiv:2503.13305 (2025)*

LM Attention Decomposes Position and Semantics

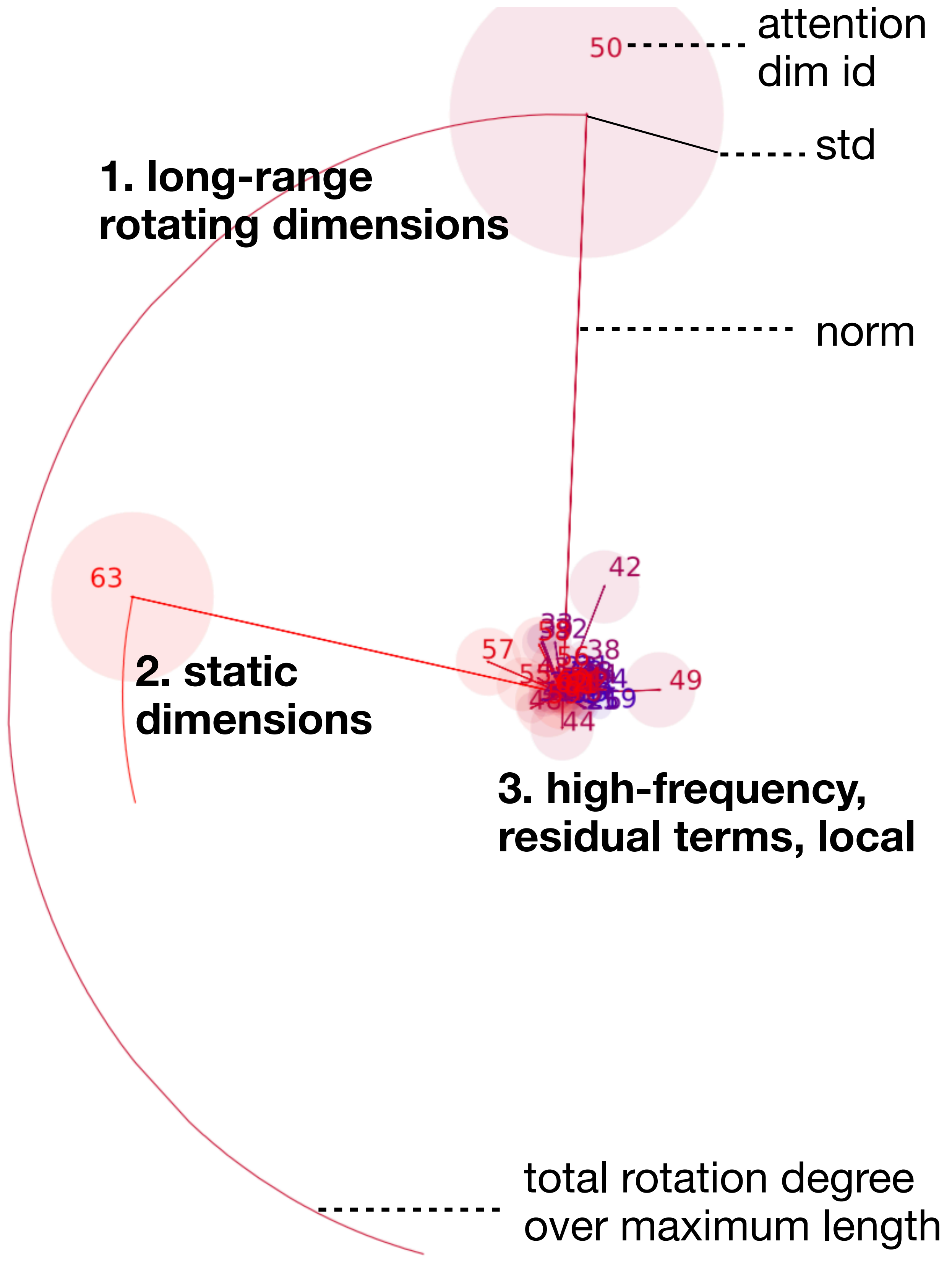


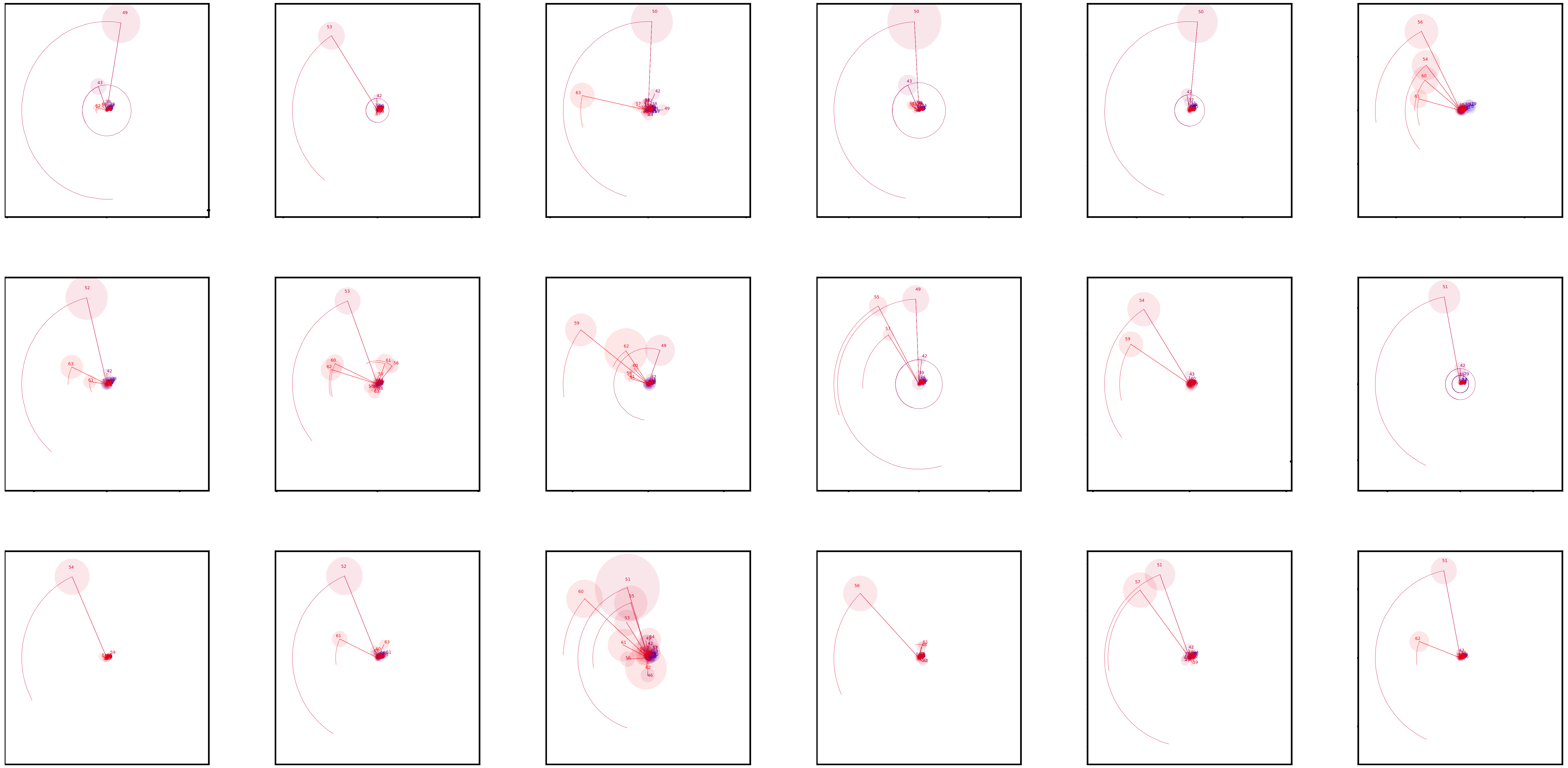
to isolate the effect of vectors q_i, k_j and their distance $i-j$, let us use a "fake" distance d instead of $i-j$

Han, Chi, et al. "Computation Mechanism Behind LLM Position Generalization" *arXiv preprint arXiv:2503.13305 (2025)*

Topic 2: Attention - Question 2: Position

An Intriguing Learned Feature Pattern





pattern in RoPE:
certain dimensions with slower "rotating speeds" have a dominant norm

Han, Chi, et al. "Computation Mechanism Behind LLM Position Generalization" *arXiv preprint arXiv:2503.13305 (2025)*

The Pattern Proves to Disentangle Attention

Theorem 1. *There exists functions $f(\boldsymbol{q}, i-j), g(\boldsymbol{q}, \boldsymbol{k})$ that so that the effect of $i-j$ and $\boldsymbol{k}$ can be asymptotically disentangled as:*

$$w(i-j, \boldsymbol{q}, \boldsymbol{k}) = f(\boldsymbol{q}, i-j) + g(\boldsymbol{q}, \boldsymbol{k}) + o(R) \quad (5)$$

, where

$$R = \max\left(Range(f), Range(g)\right)$$

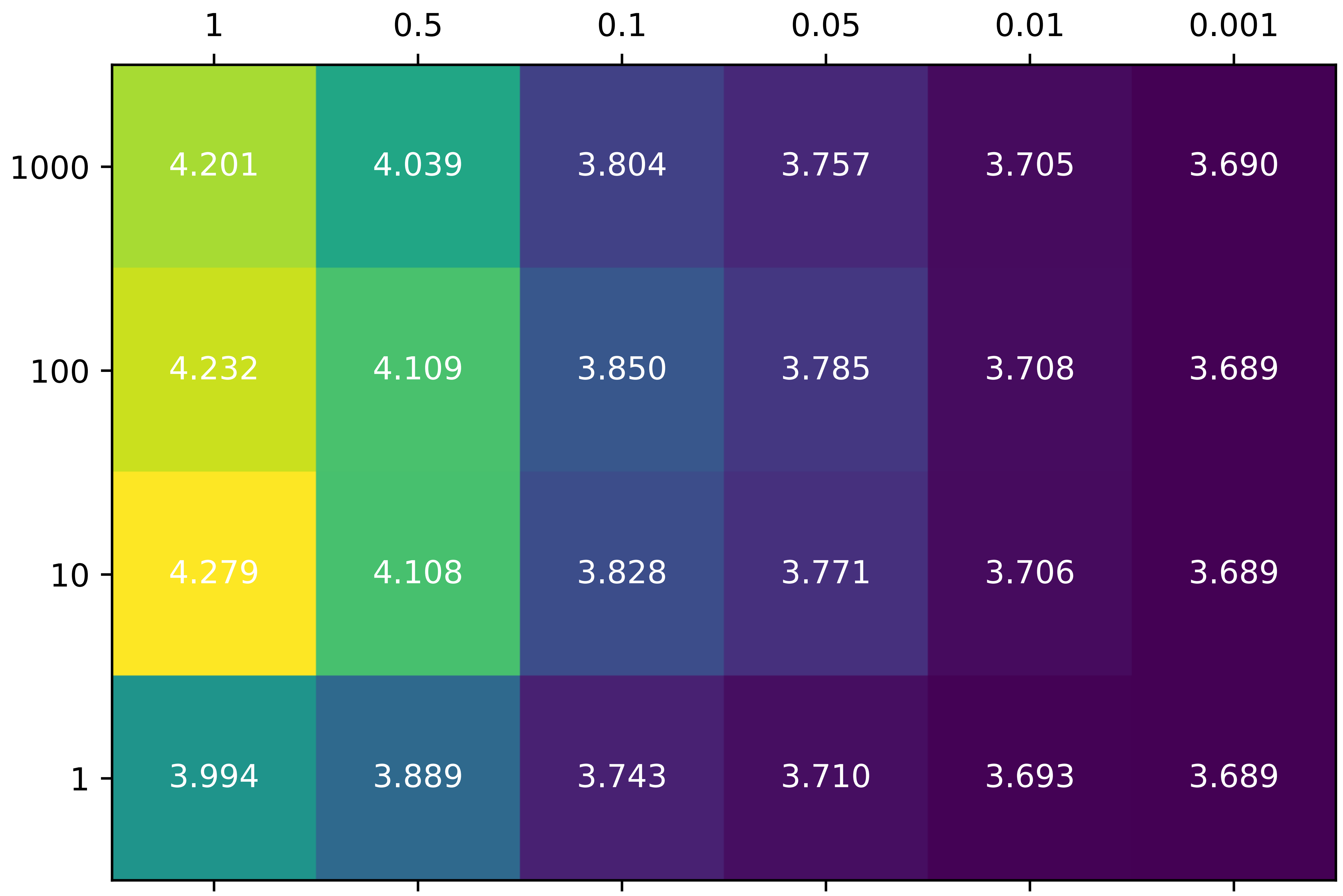
stands for the larger one of extreme range of f and g as $i, j, \boldsymbol{k}$ vary

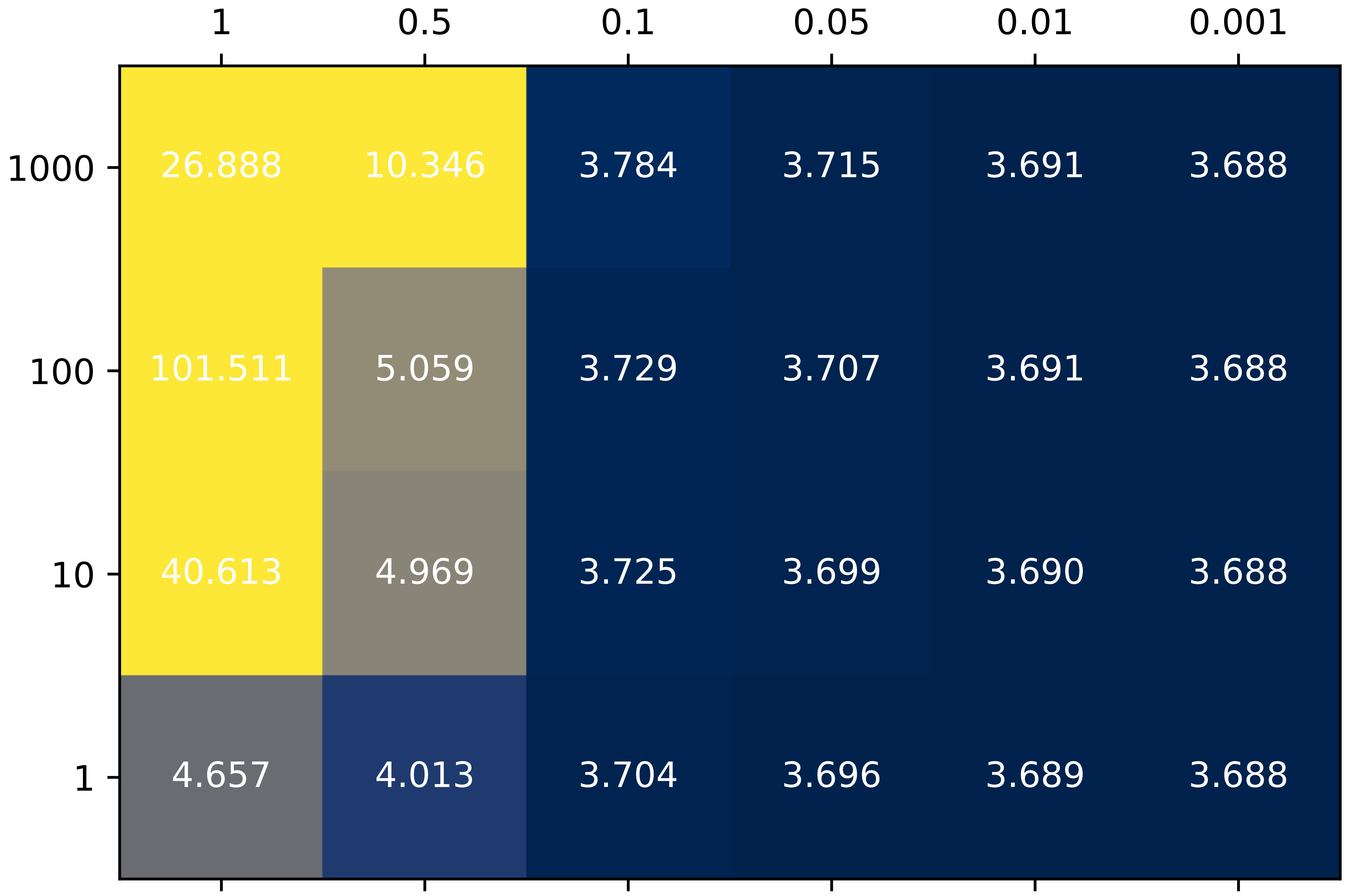
Message: LMs don't bond semantic feature $\boldsymbol{k}_j$ with their position relations $i-j$!

Han, Chi, et al. "Computation Mechanism Behind LLM Position Generalization" *arXiv preprint arXiv:2503.13305 (2025)*

Topic 2: Attention - Question 2: Position

LMs Are Stable on Disentangled Position





Perplexity if we shuffle γ ratio of *words* within max range D

Perplexity if we shuffle γ ratio of *features* within max range D

Han, Chi, et al. "Computation Mechanism Behind LLM Position Generalization" *arXiv preprint arXiv:2503.13305 (2025)*

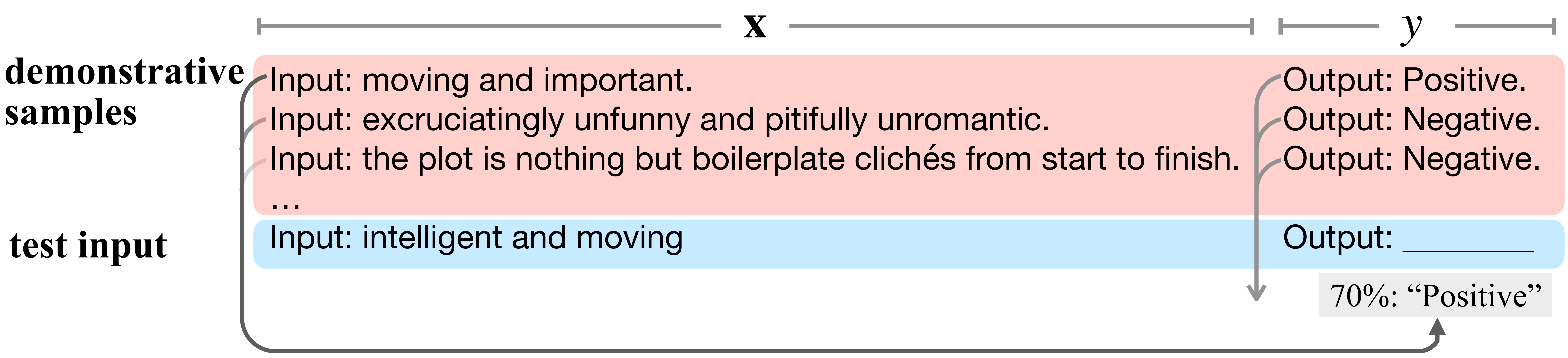
LMs Are Stable on Disentangled Position

Operation	Qasper Accuracy				
	0.5	0.1	0.05	0.01	0.001
Original			42.53		
Text Order	37.39	41.44	42.34	42.37	42.53
Feature Order	35.11	41.15	41.98	42.33	42.56

QA Task accuracy if we shuffle γ ratio of *words or features* within max range 5

Han, Chi, et al. "Computation Mechanism Behind LLM Position Generalization" *arXiv preprint arXiv:2503.13305 (2025)*

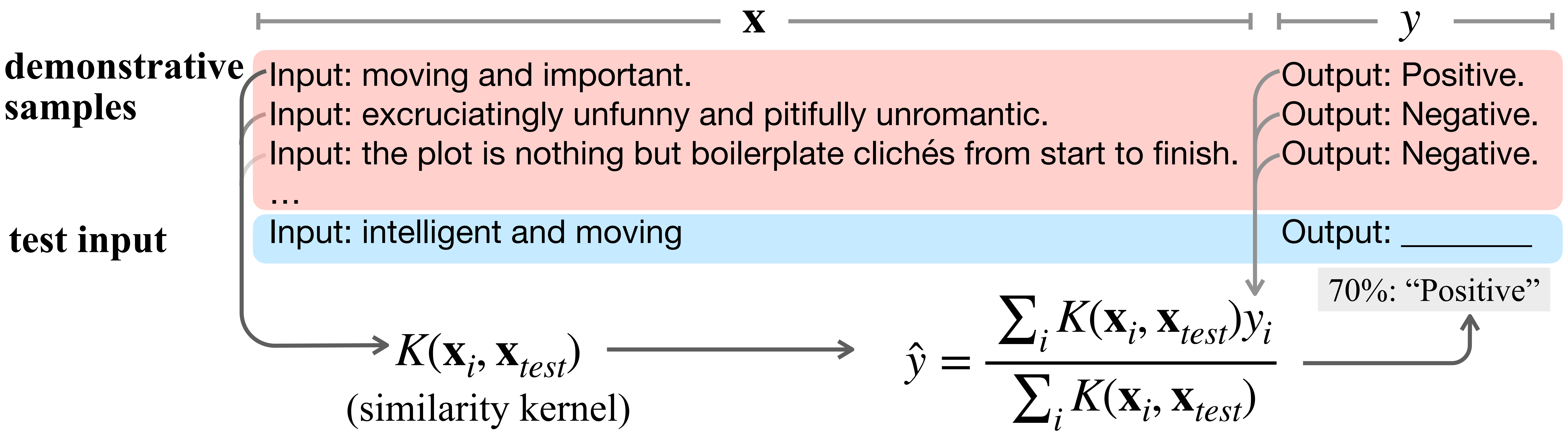
Attention Also Explains In-Context Learning



In-context learning: completing tasks based on demonstrations

Han, Chi, et al. "Explaining emergent in-context learning as kernel regression." *arXiv preprint arXiv:2305.12766* (2023).

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 $K(\boldsymbol{x}_i, \boldsymbol{x}_{text})$ (i.e., a kernel)

Han, Chi, et al. "Explaining emergent in-context learning as kernel regression." *arXiv preprint arXiv:2305.12766* (2023).

The Kernel Originates from Pre-Training

Kernel regression (hypothesized ICL algorithm)

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

Han, Chi, et al. "Explaining emergent in-context learning as kernel regression." *arXiv preprint arXiv:2305.12766* (2023).

Topic 2: Attention - Question 3: Contextual Knowledge

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)

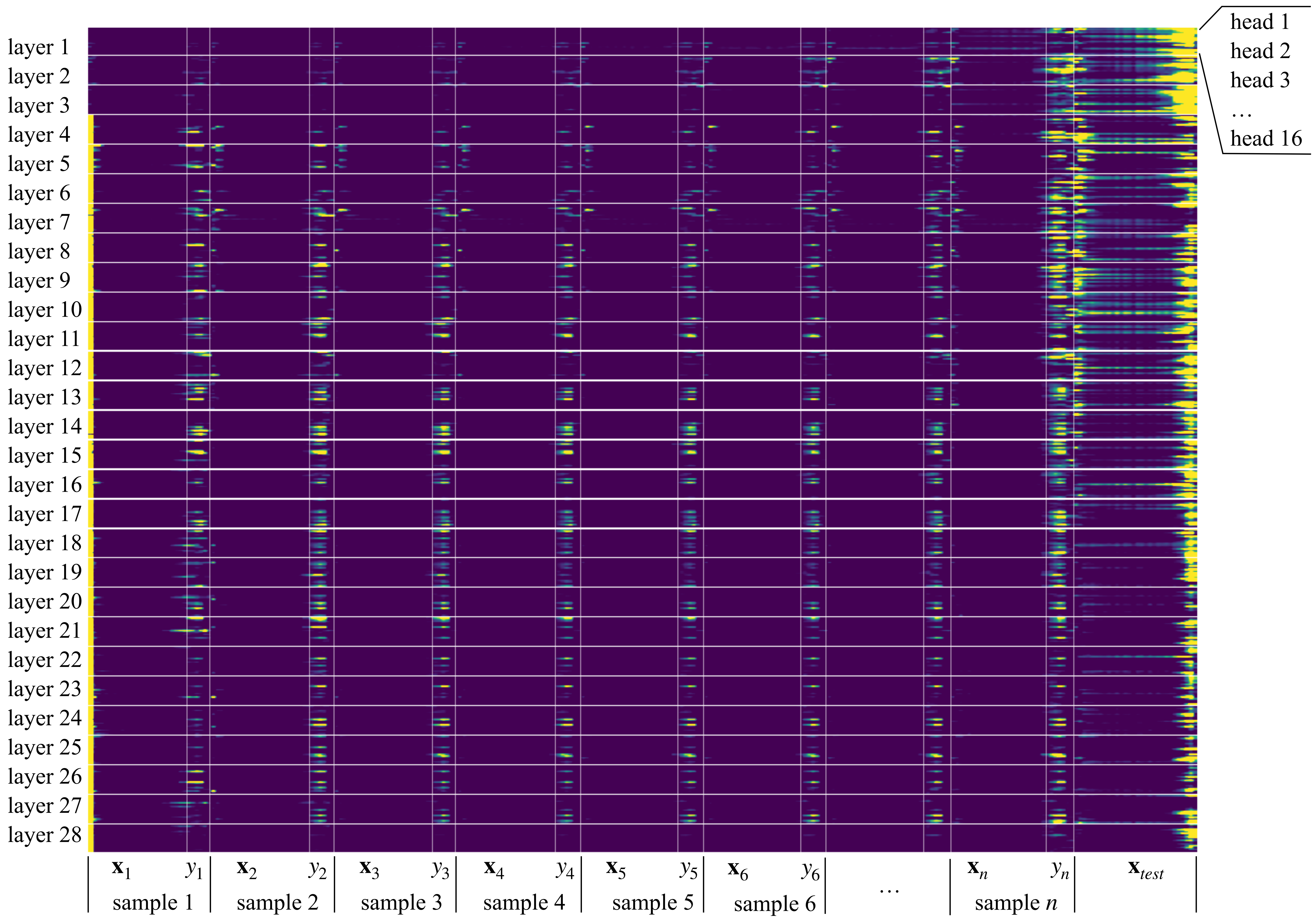
$$\mathcal{K}(\mathbf{x}, \mathbf{x}') = \mathrm{vec}(T_{\mathbf{x}})^{\top} \Sigma^{-1}_{p_{\textit{pre-train}}}, \mathrm{vec}(T_{\mathbf{x}'})$$

A representation of sample input x for predicting the next token

A matrix about the pre-training objective

Han, Chi, et al. "Explaining emergent in-context learning as kernel regression." *arXiv preprint arXiv:2305.12766* (2023).

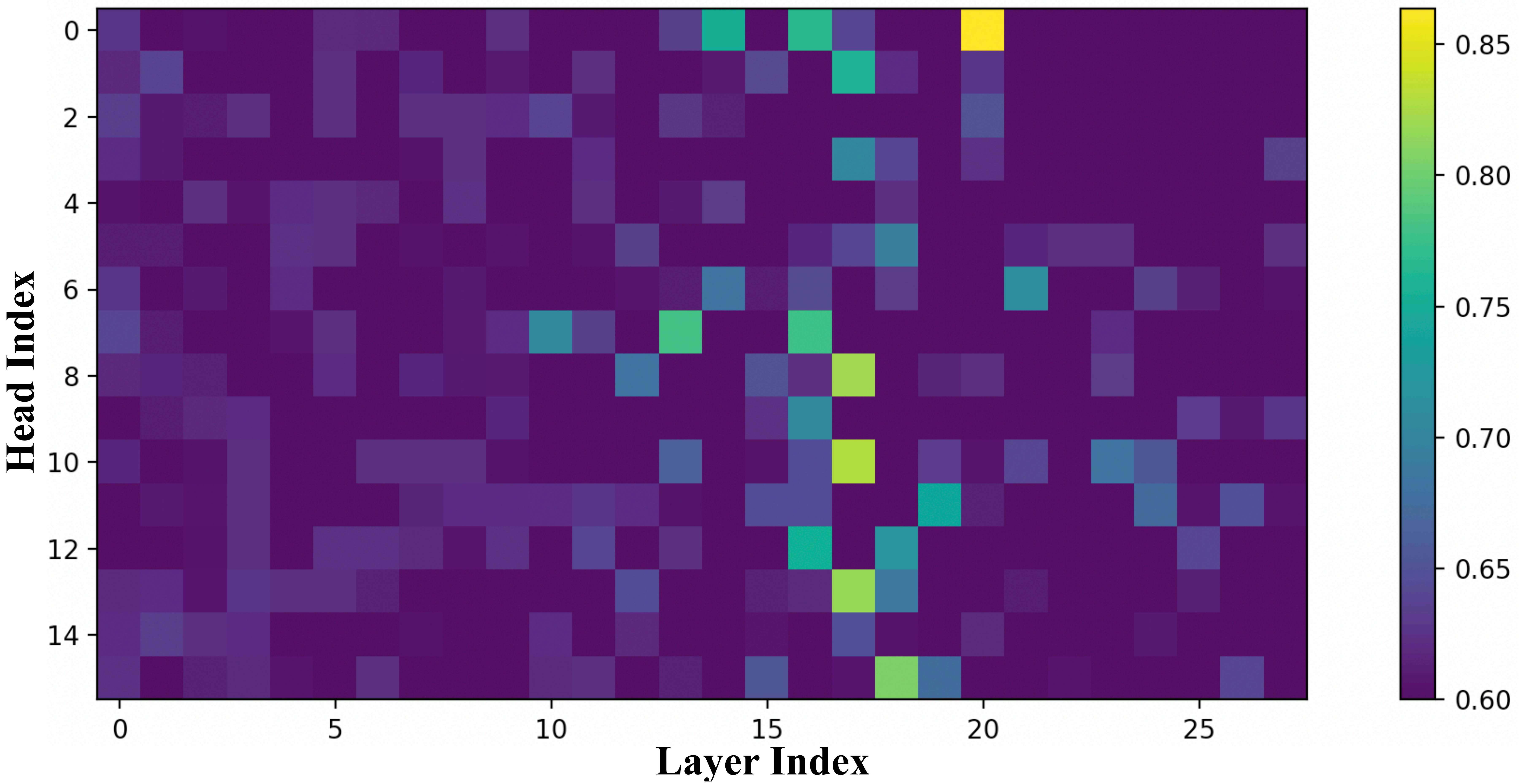
The Attention Applies to y_i As Kernel Regression



Han, Chi, et al. "Explaining emergent in-context learning as kernel regression." *arXiv preprint arXiv:2305.12766* (2023).

Topic 2: Attention - Question 3: Contextual Knowledge

The Explanation Aligns With the Model Output



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

Han, Chi, et al. "Explaining emergent in-context learning as kernel regression." *arXiv preprint arXiv:2305.12766* (2023).

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

Han, Chi, et al. "Explaining emergent in-context learning as kernel regression." *arXiv preprint arXiv:2305.12766* (2023).

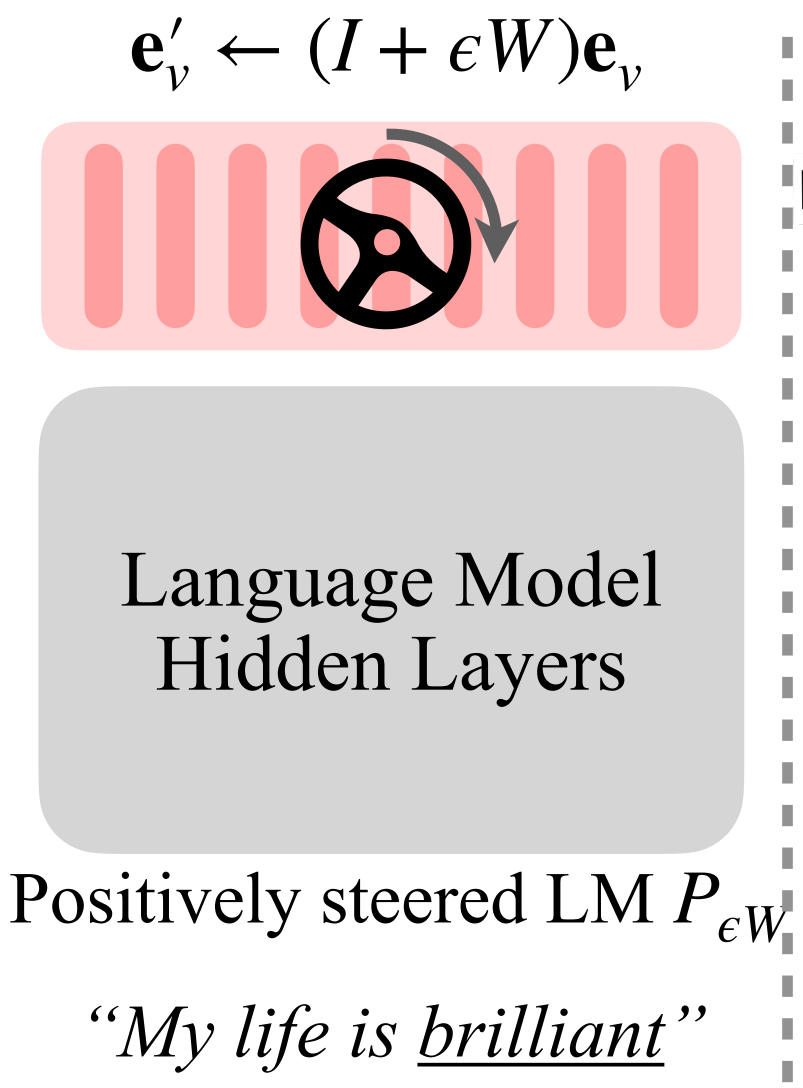
Room for Future Research

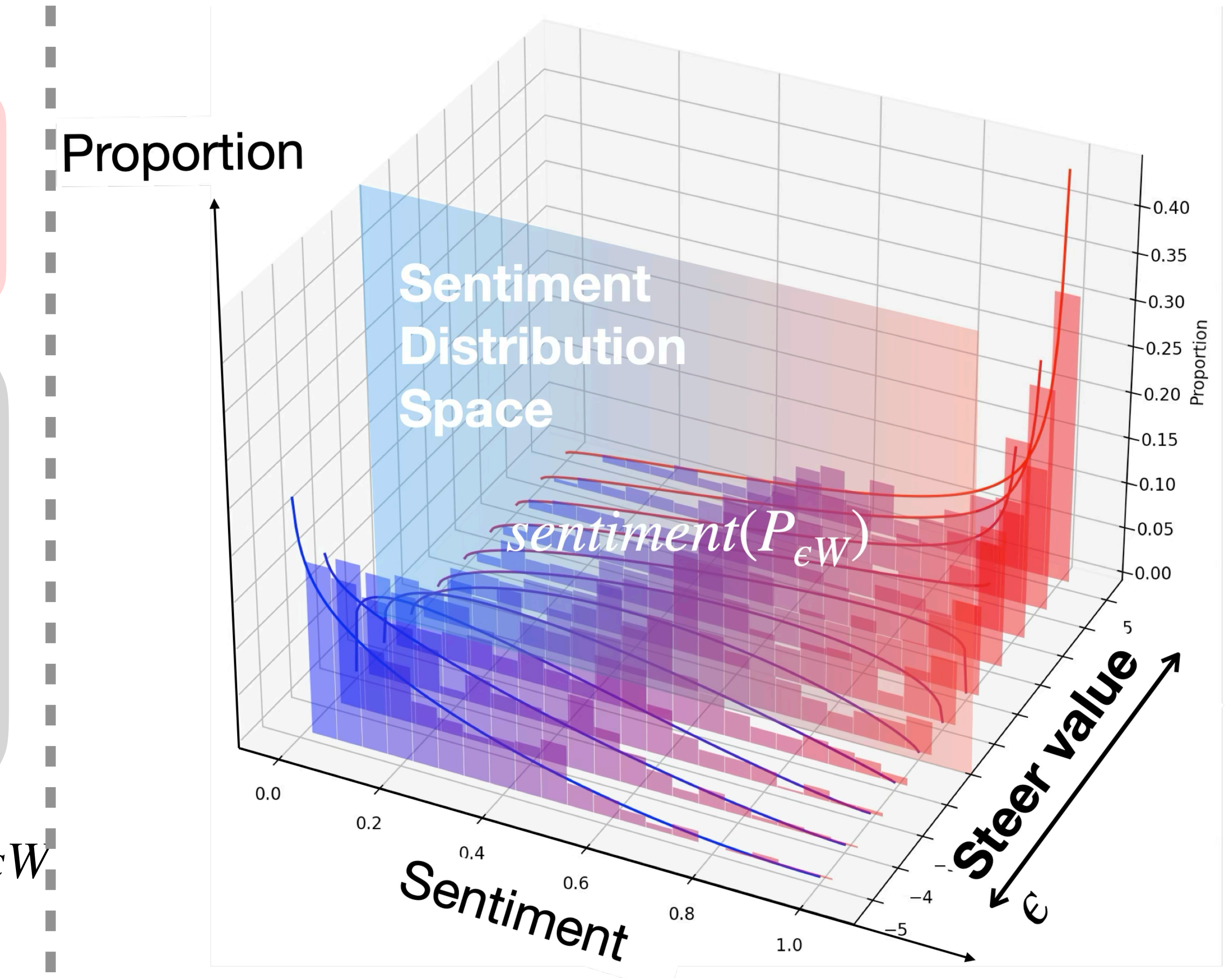
- Attention module’s role in syntax and word order processing
- More precise categorization of attention’s role in demonstration learning
- Explaining and addressing and lost-in-the-middle and position bias problem
- Extension to other model architectures

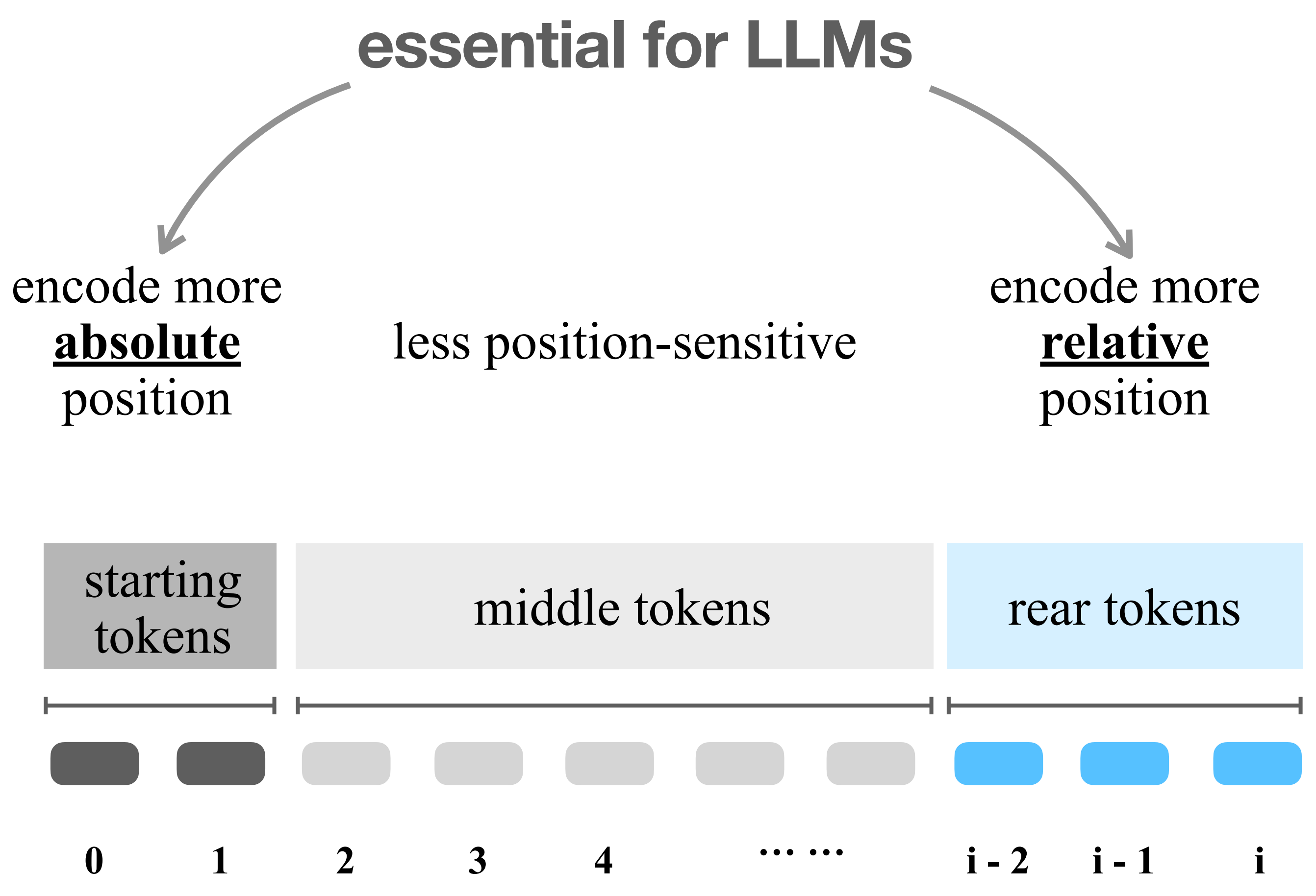
Summary

Towards a Physiology of Language Models: Elucidating and Utilizing Hidden Language Representation

- Topic 1: What Is the Function of Word Embeddings
- Topic 2: Attention, Position and Context
    - Q1: How LMs Deal with Context Length
    - Q2: How LMs Process Position Information
    - Q3: How LMs Comprehend Contextual Knowledge



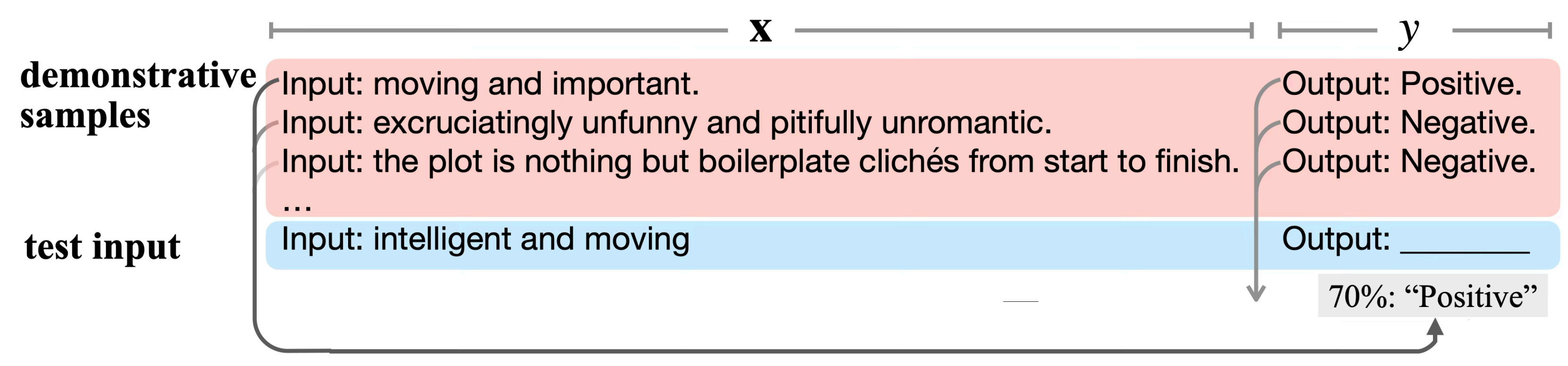




$$w(i-j, \boldsymbol{q}, \boldsymbol{k}) = f(\boldsymbol{q}, i-j) + g(\boldsymbol{q}, \boldsymbol{k}) + o(R) \quad (5)$$

, *where*

$$R = \max\left(Range(f), Range(g)\right)$$



A Retrospect of Science of LMs