Towards
A Science of LMs

CS546 Guest Lecture

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

Why Do We Need A New Science?

Doesn't ML already provide a scientific basis for LM?

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

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…

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

PAC theory,
optimization, …

Gradient Descent,
Neural Tangent Kernel,
…

A Sciences of LMs

https://www.twinkl.com/parenting-wiki/particle, https://www.google.com/url?sa=i&url=https://simple.wikipedia.org/wiki/Shock_wave, https://www.google.com/url?sa=i&url=https://bitwiseacademy.com/what-is-fluid-mechanics/&psig=AOvVaw1hA8LI-JEcyxC67WonlgY8A&ust=1729744191924000

Why Do We Need A Science at All?

Empirical Rules Can Be Misleading

- *“The car allergic to vanilla ice cream”*

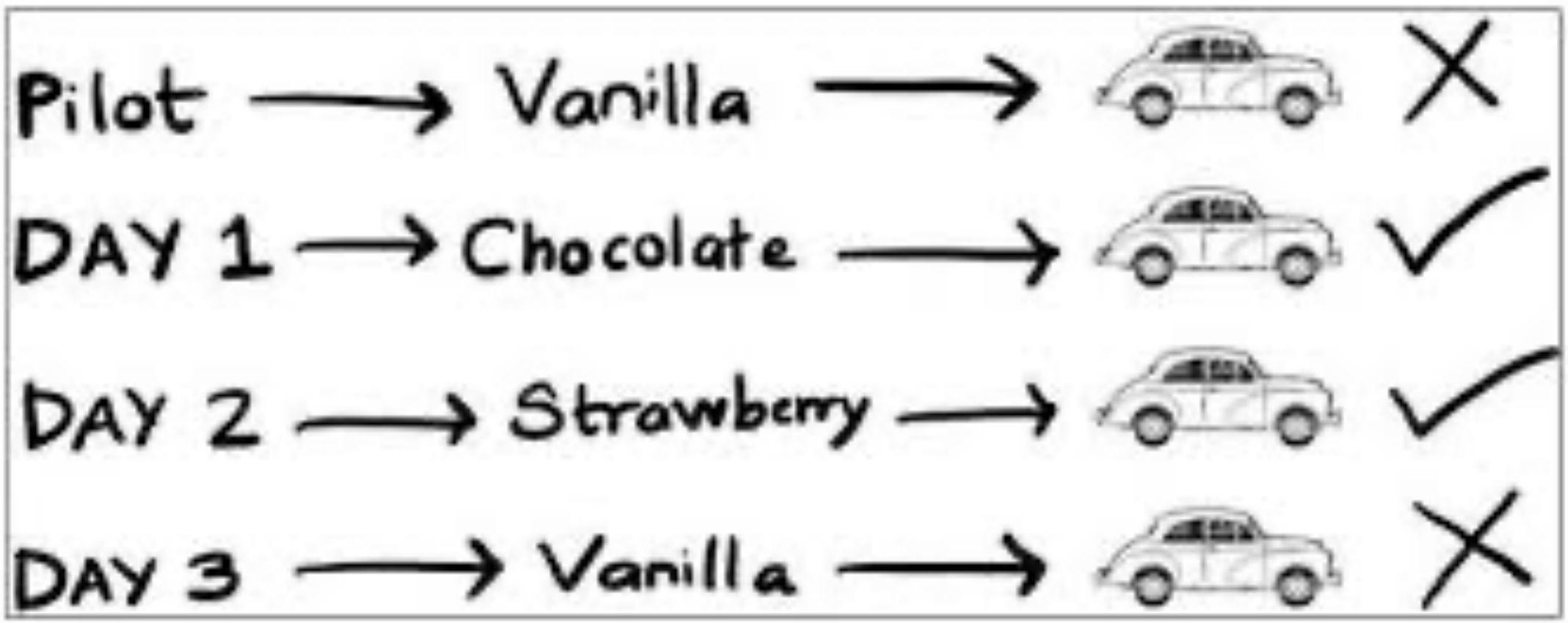
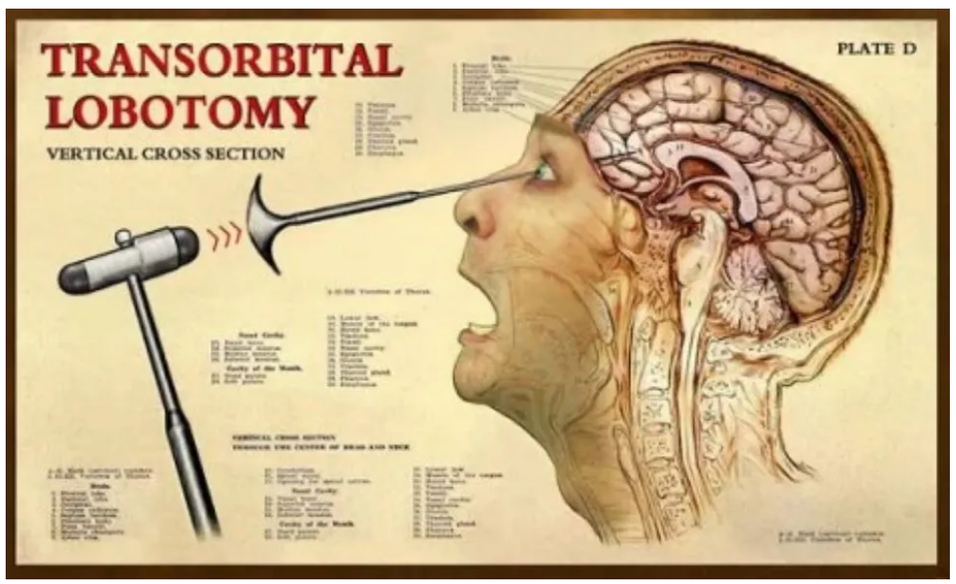


image: https://www.google.com/url?sa=i&url=https://www.facebook.com/photo.php?fbid=611022205721805&id=441524076004953&set=a.444237809066913&locale=hi_IN&psig=AOvVaw0iIPqPpWfkyvzXzvCOTspc&ust=1729715035064000&source=images&cd=vfe&opi=89978449&ved=0CBcQjhxqFwoTCPCqu-PoookDFQAAAAAdAAAAABAE

Mistakes From Lack of Science

Science	Mistakes
physics & thermodynamics	attempts in perpetual motion machines
neuroscience	transorbital lobotomies
psychology	exorcisms or bloodletting
chemistry	using lead in water pipes and medicine

https://www.stellariplaw.com/post/2019/02/20/are-perpetual-motion-machines-patentable

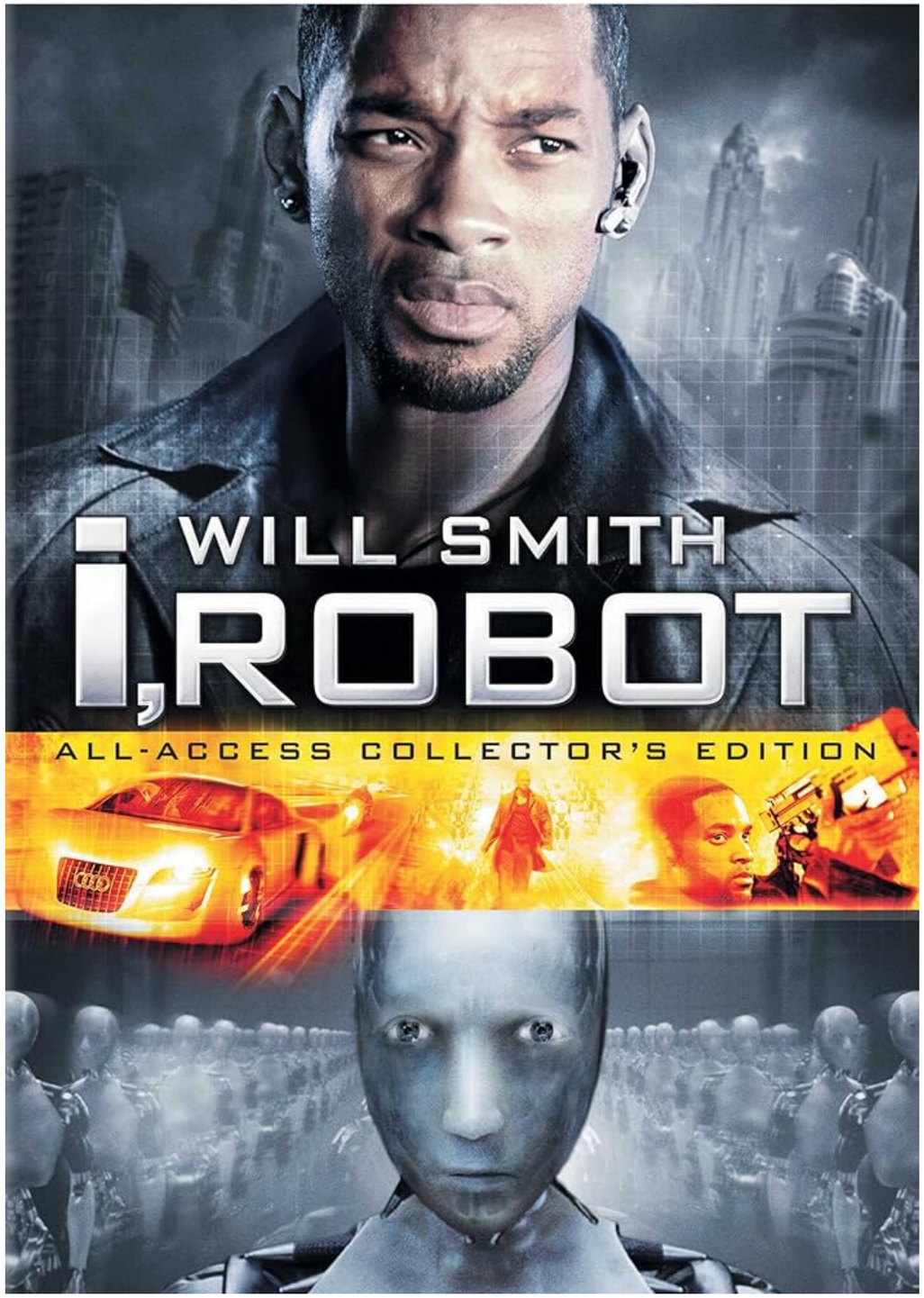


https://www.google.com/url?sa=i&url=https://sashaayad.com/parallels-between-lobotomy-and-childhood-gender-transition&psig=AOvVaw0r2fjwop6XQ-nrvppl12sp&ust=1729715180024000&source=images&cd=vfe&opi=89

https://www.google.com/url?sa=i&url=https://thecatalystnews.com/2020/04/09/the-benefits-of-bloodletting/&psig=AOvVaw2389ahnlIA9gl66NPLQYFN&ust=1729715254971000&source=images&cd=vfe&opi=89978449&ved=0CBcQjhxqFwoTCLjhxdTpookDFQAAAAAdAAAAABAE

What If Similar Accidents Happen to LMs?

- Waste of research efforts
- Trustworthy
- Safety
- Public confidence
- …



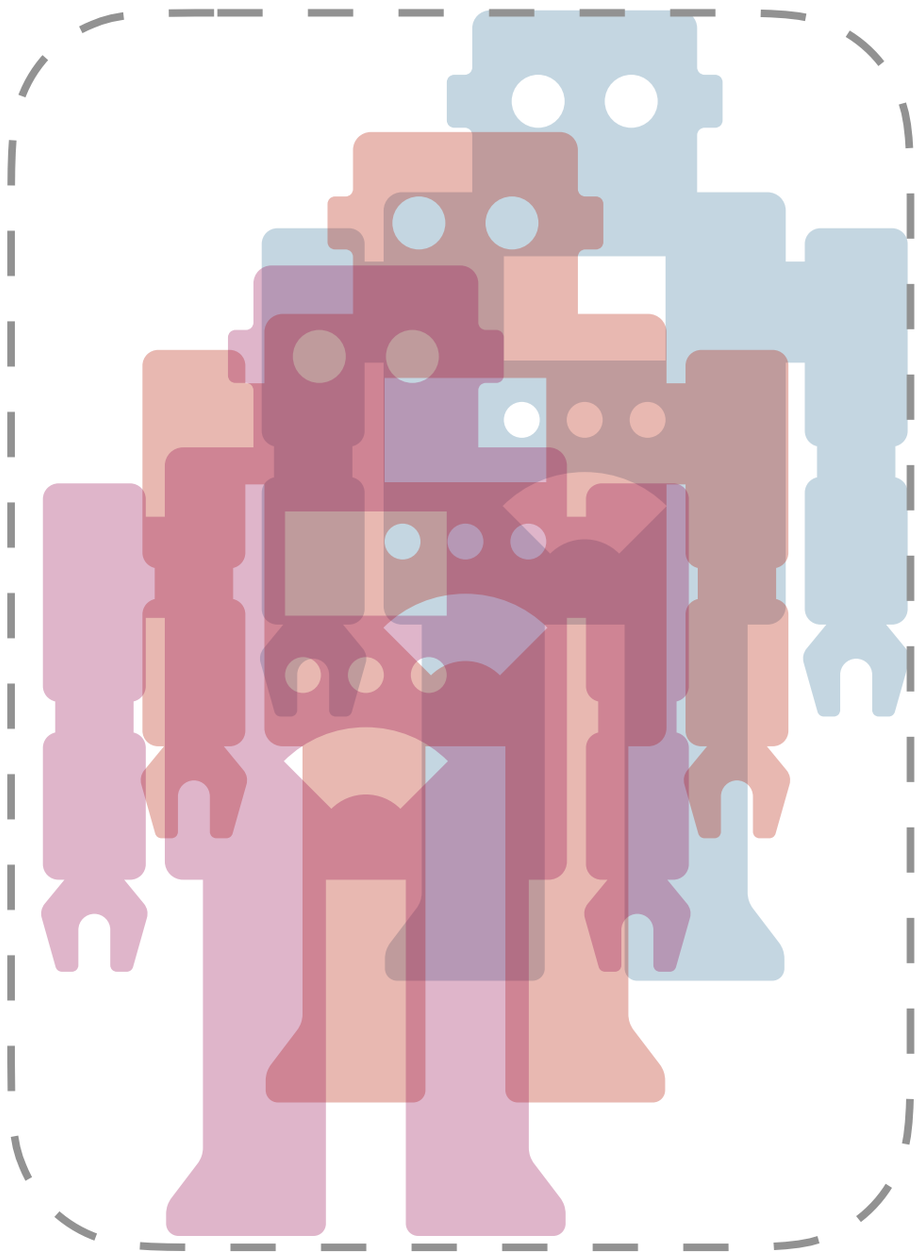
https://www.amazon.com/I-Robot-Two-Disc-All-Access-Collectors/dp/B0007PALSE

Levels of Sciences of LMs

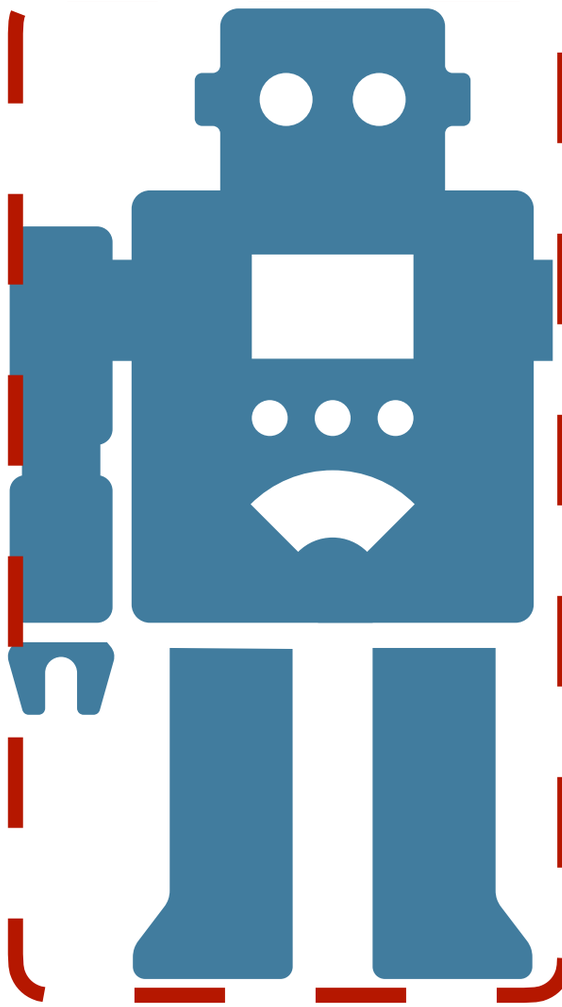
high-level

low-level

Physics of LMs
(group level, rule induction)

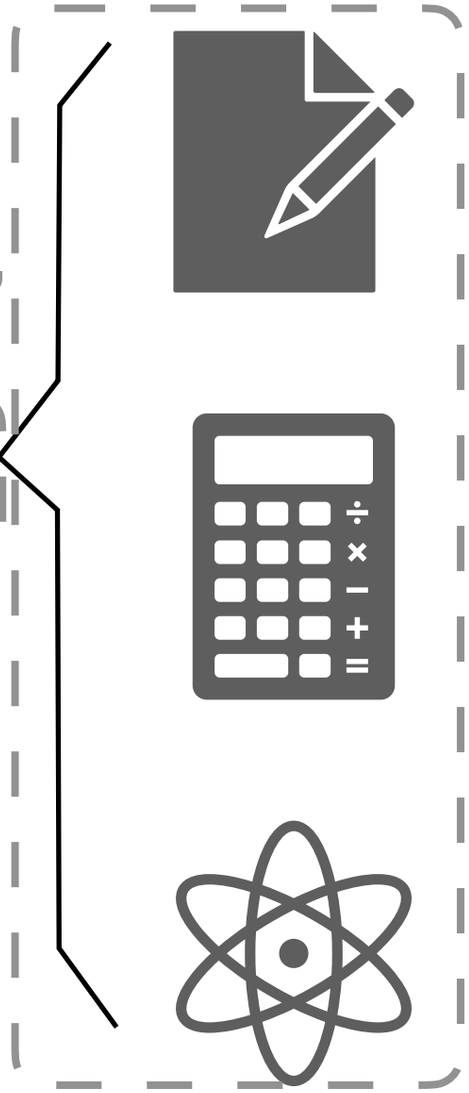


Physiology of LMs
(components-level, mechanism)



Ethology
(Instance level, behaviors)

Performance:
(Task level, scores)



Part 1:

Physics of LMs

- Scaling Laws
- Language, Reasoning & Knowledge Acquisition (Allen Zhu’s work)

General Principle

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

https://en.wikipedia.org/wiki/File:Galileolarge.png

1.1 Scaling Laws

Scaling Laws

Is Model Performance Predictable?

In physics:

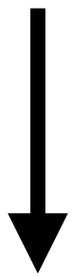
Intuition:
larger force + smaller weight
→ moving faster

Newton's Law:

$$F = ma$$

In LMs:

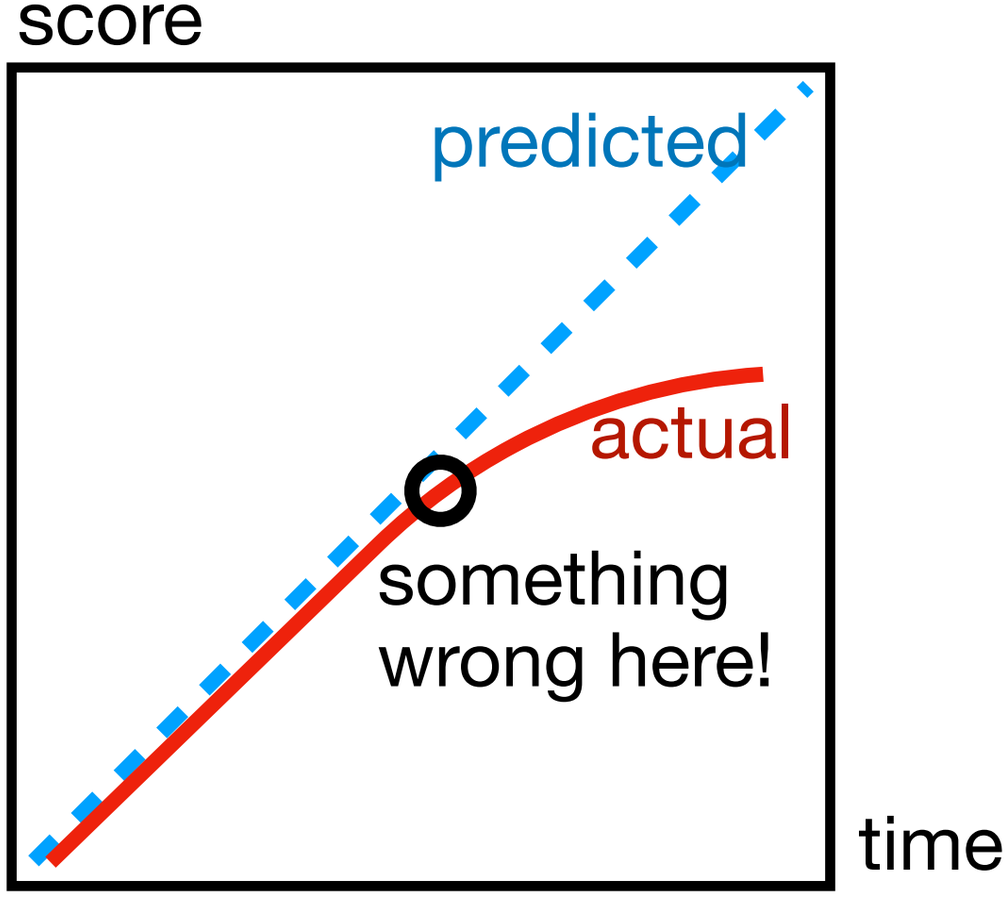
Intuition:
larger model + more data →
higher score

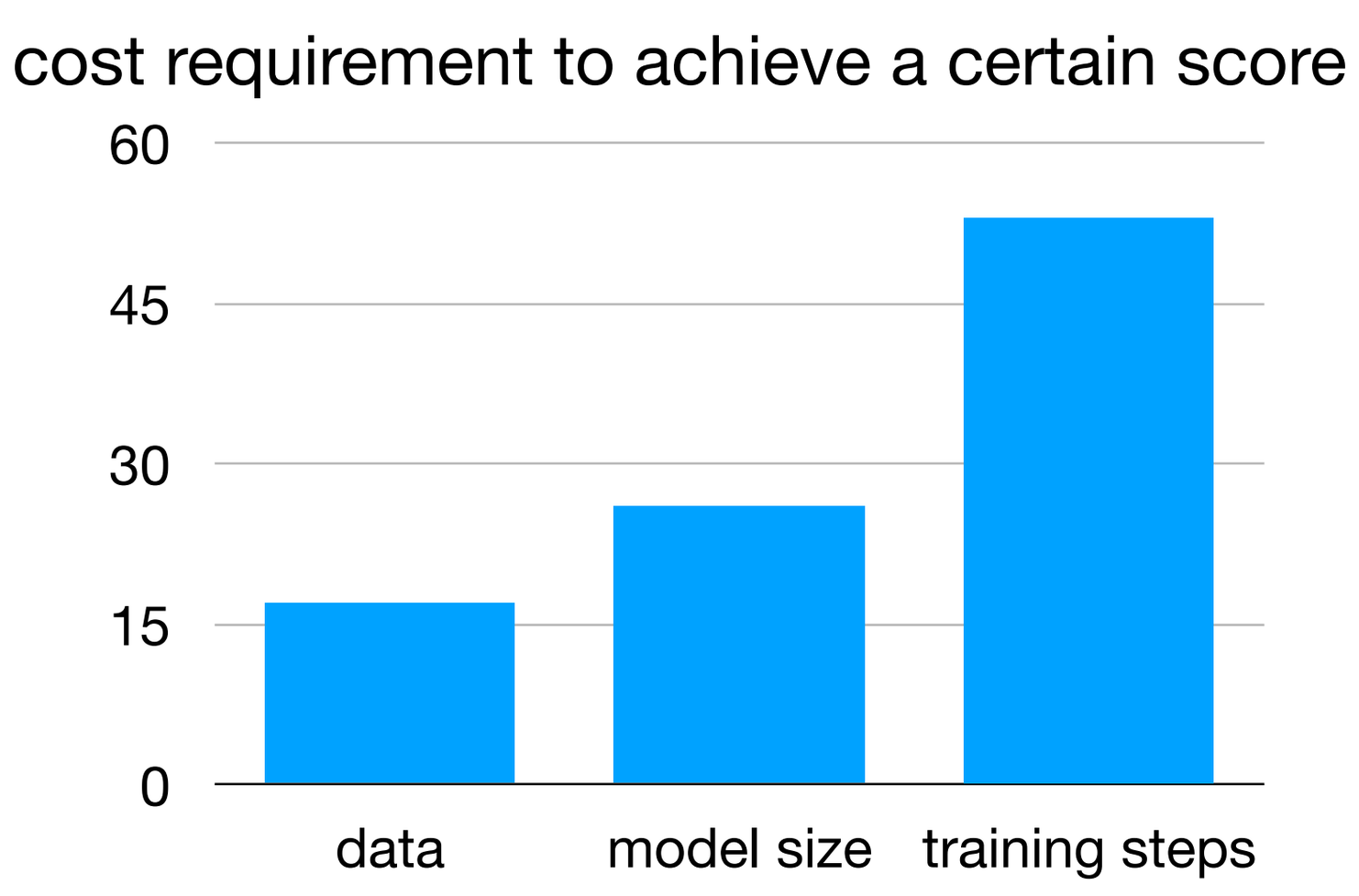


Any law to predict scores before training?

Why Do We Need Scaling Laws?

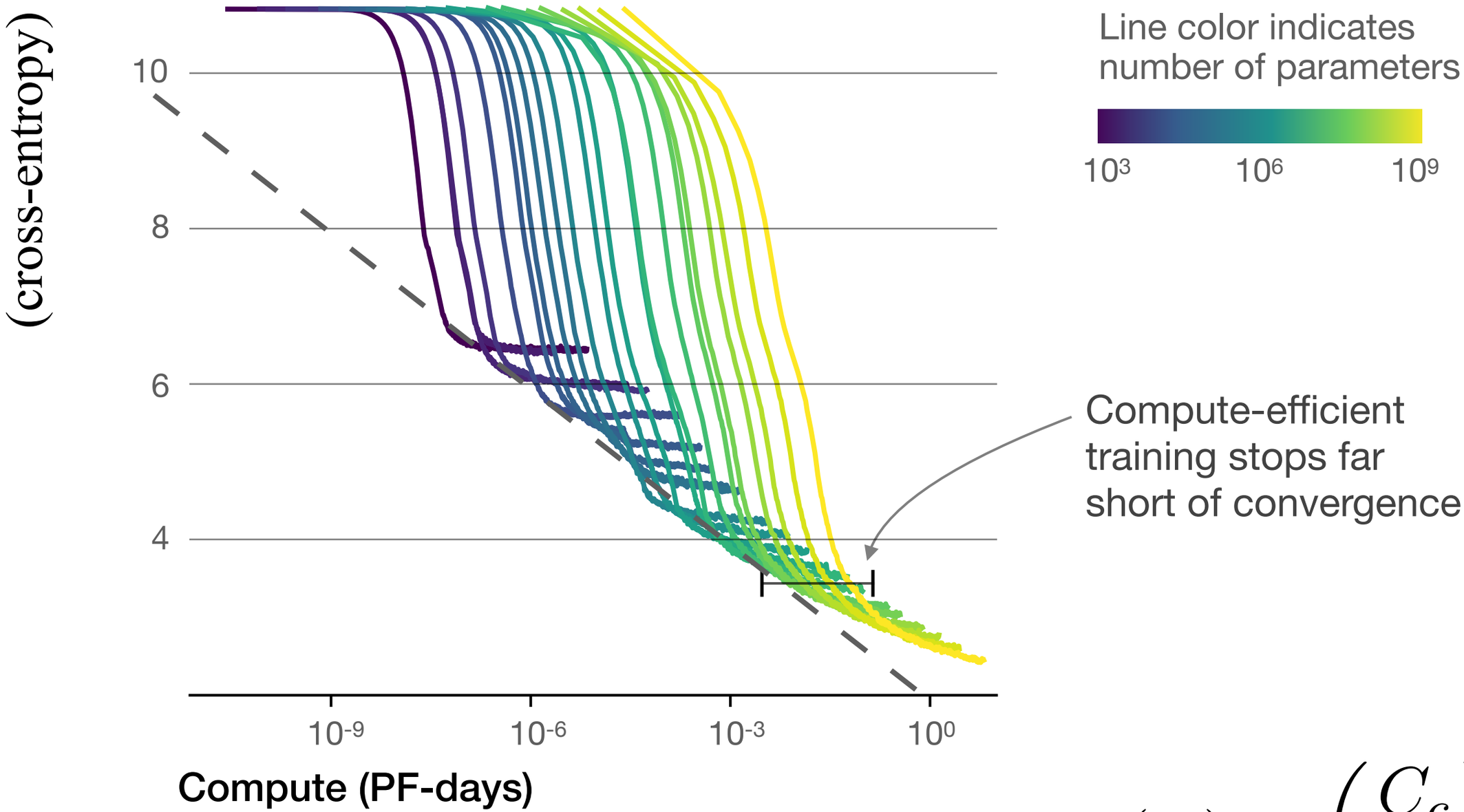
1. Curiosity
2. Early debugging
3. Better allocation of the resources





Law on Training Compute

Fitting points: early-stop points

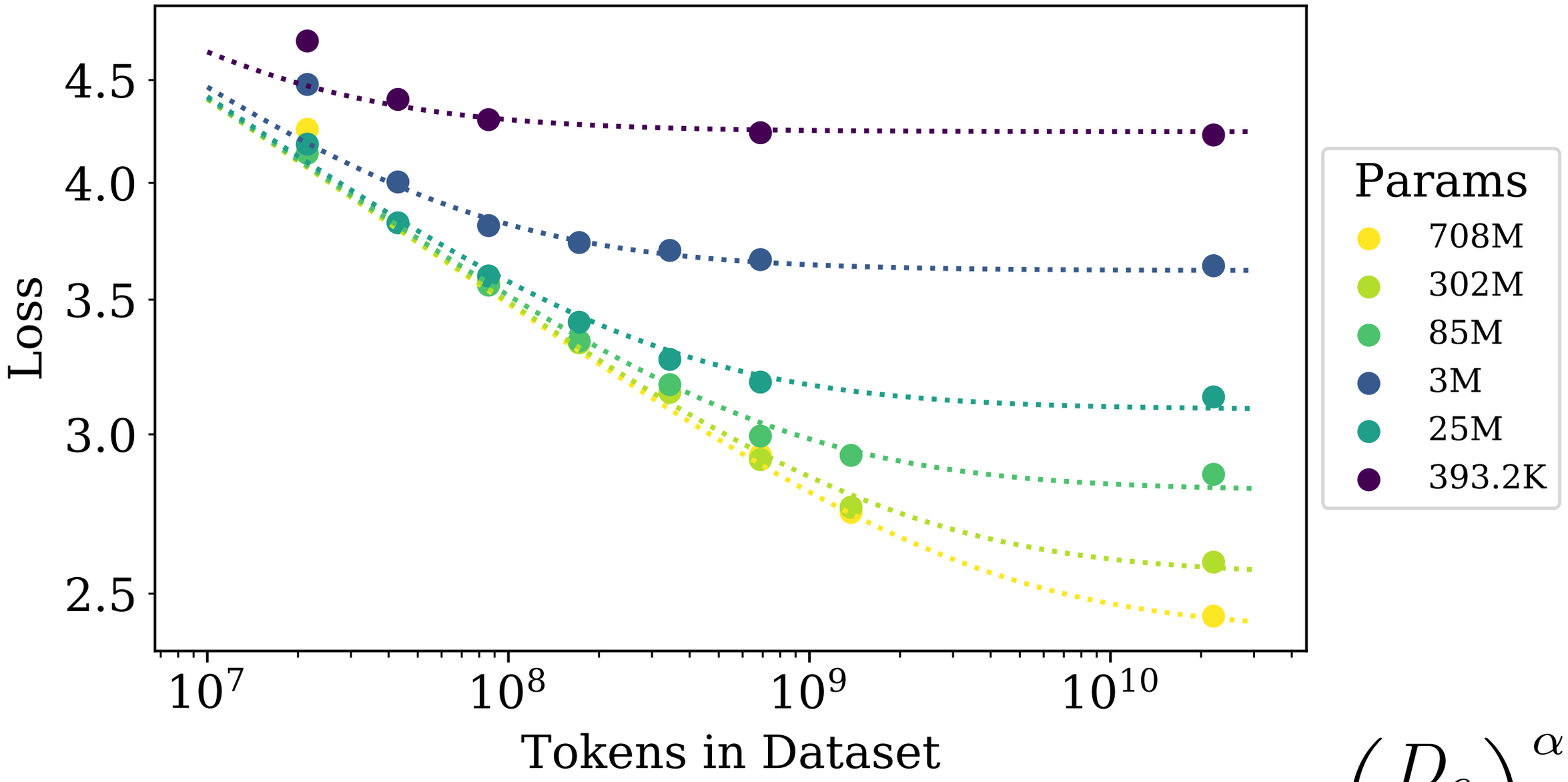


a PF-Day = 8.63e19 floating point operations

$$L(C) \approx \left(\frac{C_c}{C}\right)^{\alpha_C}$$

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

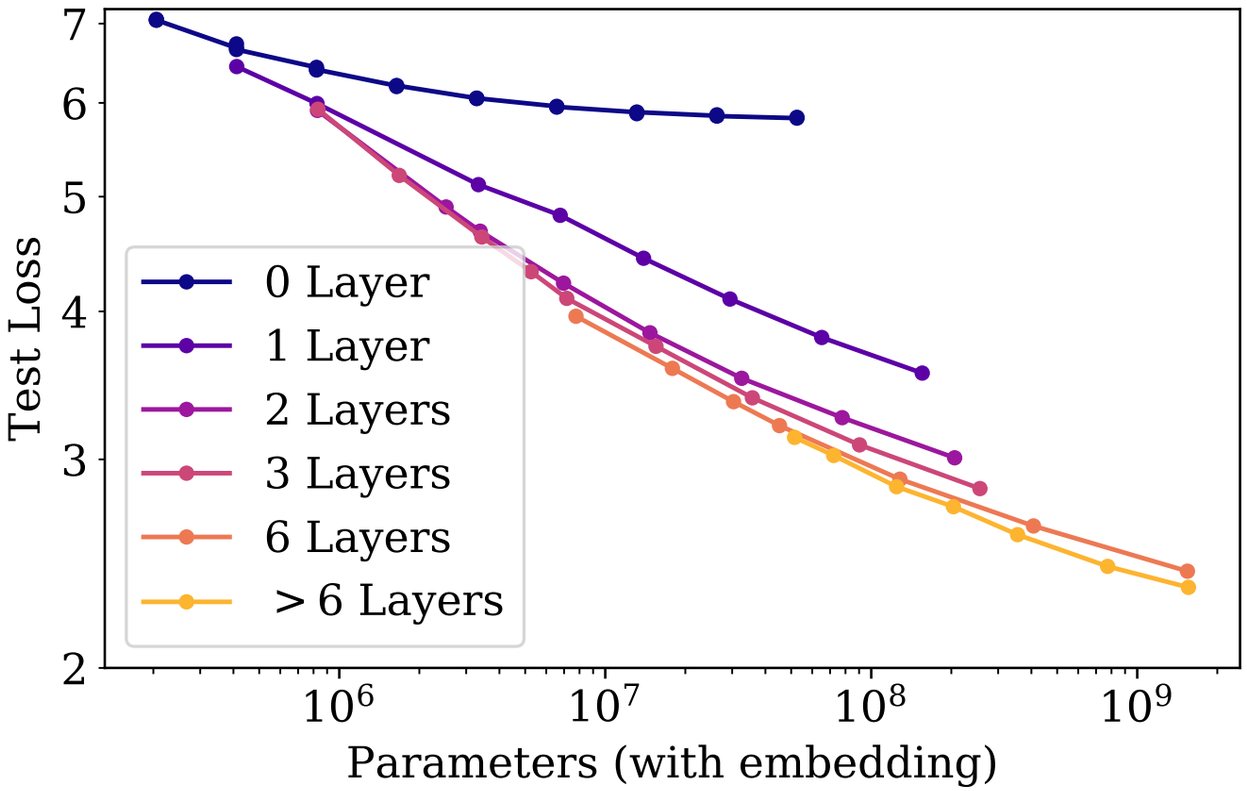
Law on Dataset Size



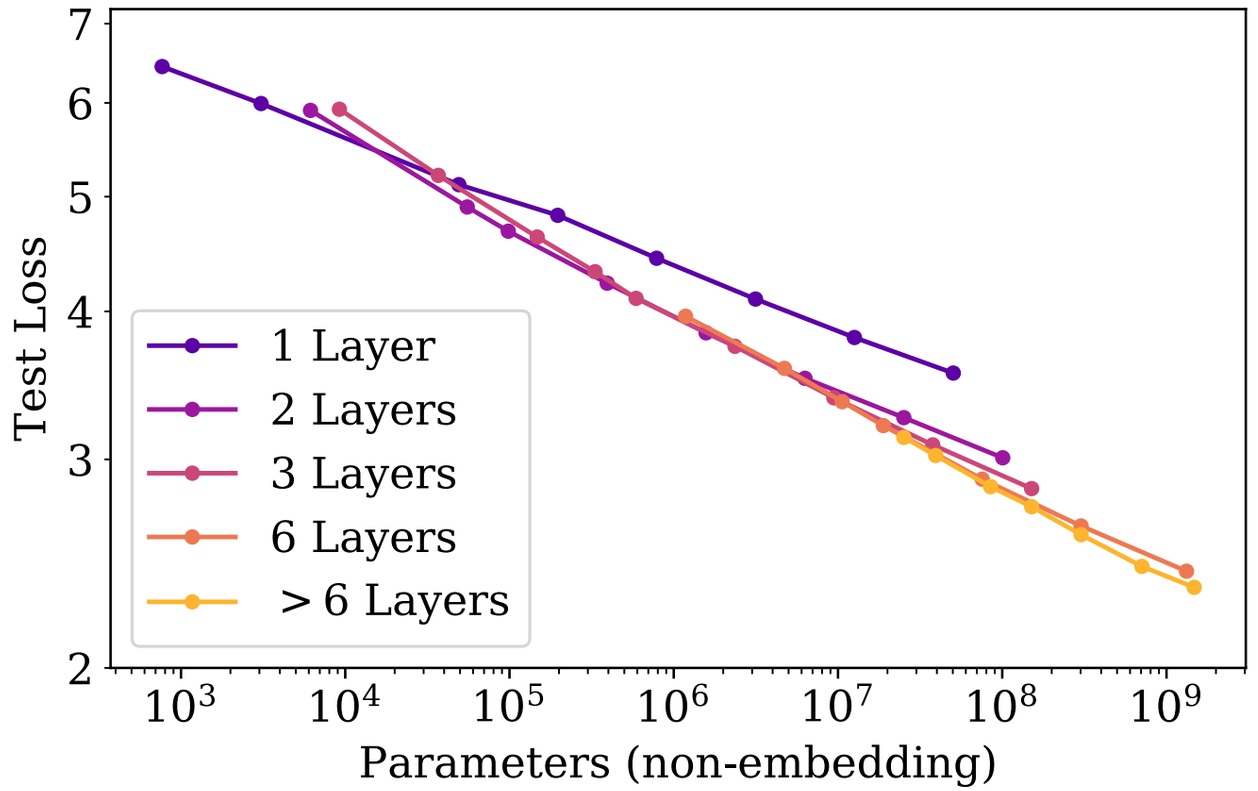
$$L(D) \approx \left(\frac{D_c}{D}\right)^{\alpha_D}$$

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

Law on Model Size



w/ embeddings: harder to fit



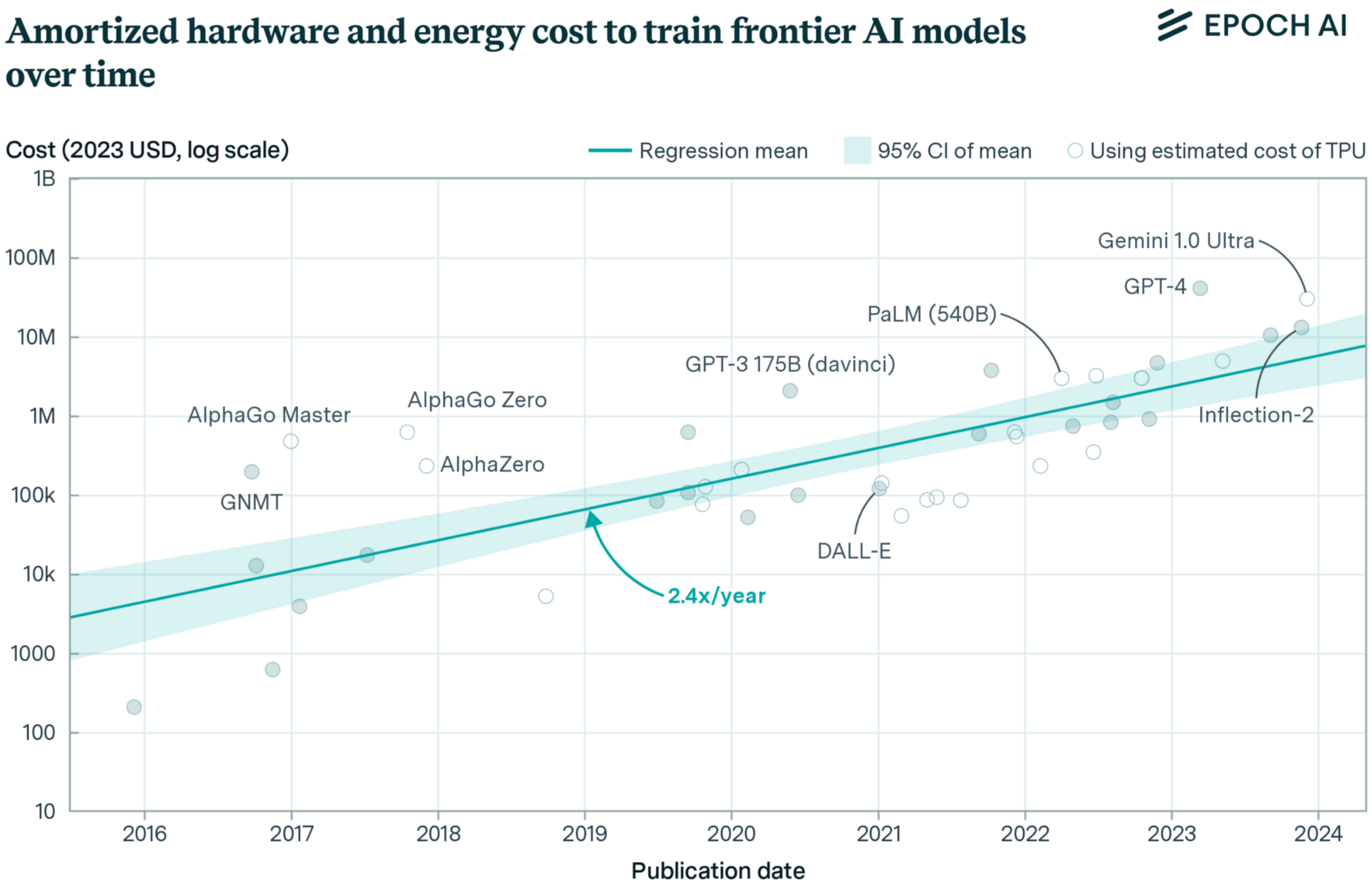
w/o embeddings: neater trend

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

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

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

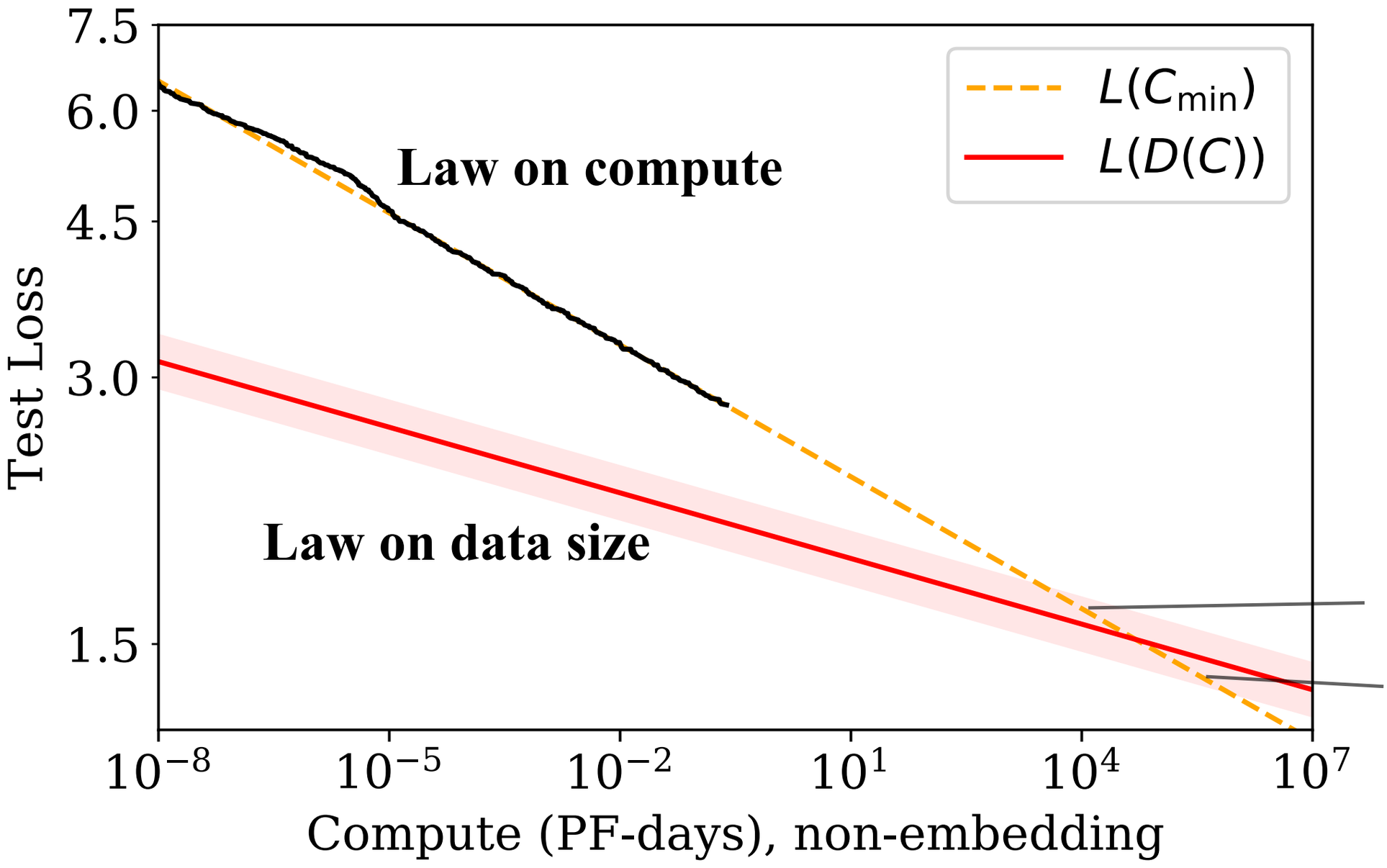
Temporal Laws



not necessarily a “scaling law” but tells us useful trends

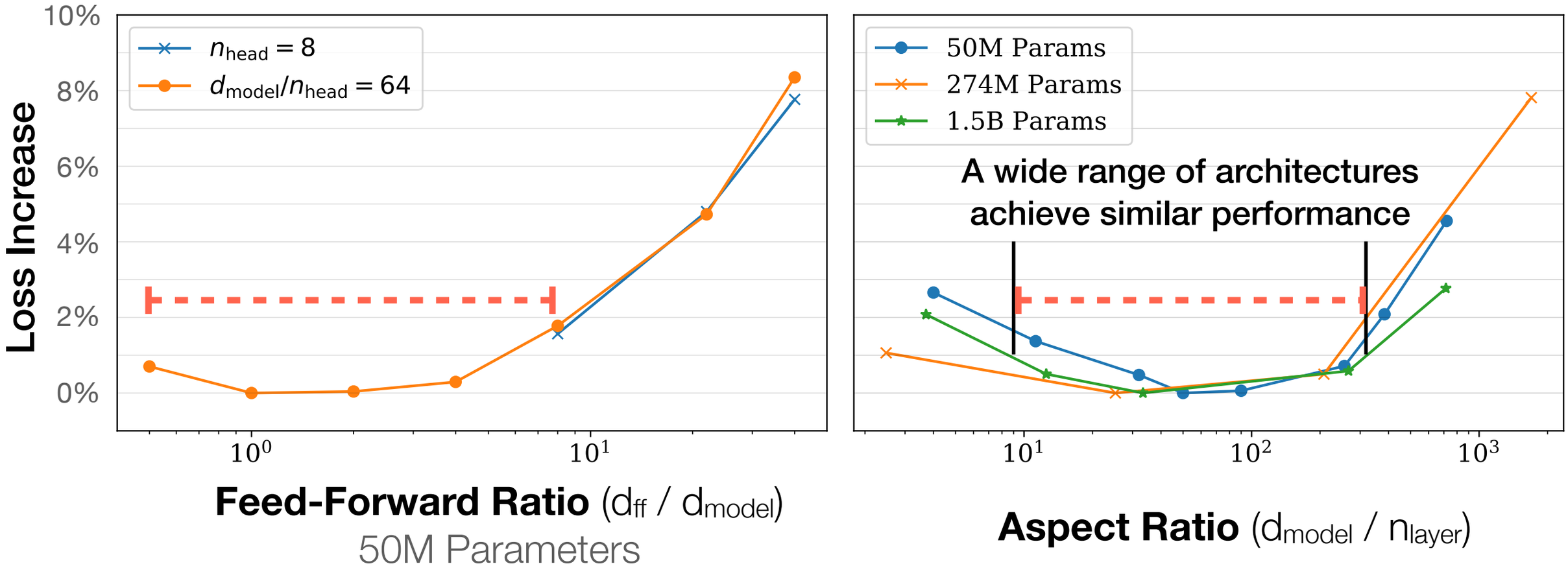
https://epochai.org/

Not Scaling Forever



Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

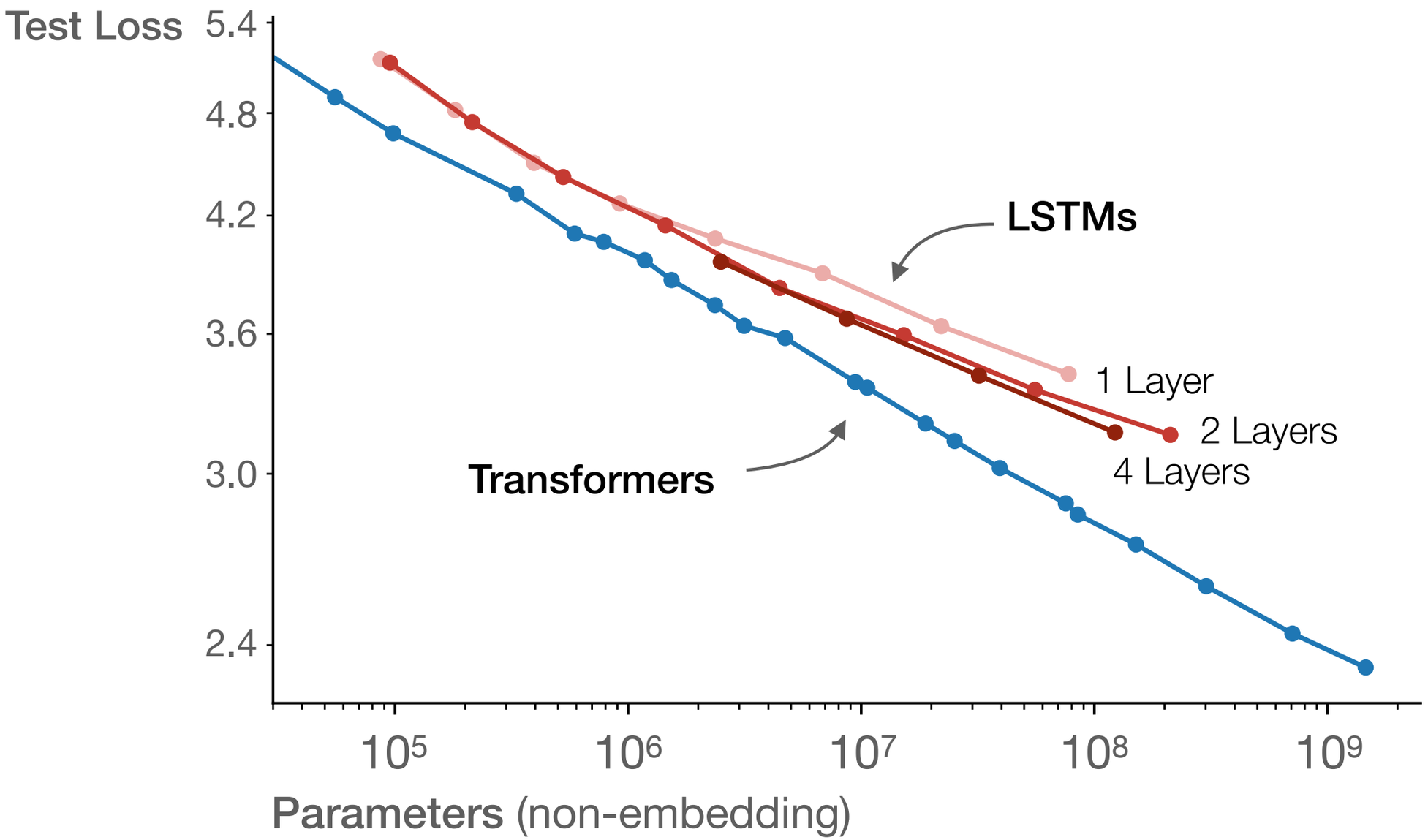
Model Shape Matters Less



Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

But Architecture Matters

Transformers asymptotically outperform LSTMs
due to improved use of long contexts



Open question: what decides this “model factor”?

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

Can Laws Be Unified?

The pursuit of a unified theory in physics has been a long-standing dream.

~~gravity~~

electromagnetic force

strong force

weak force

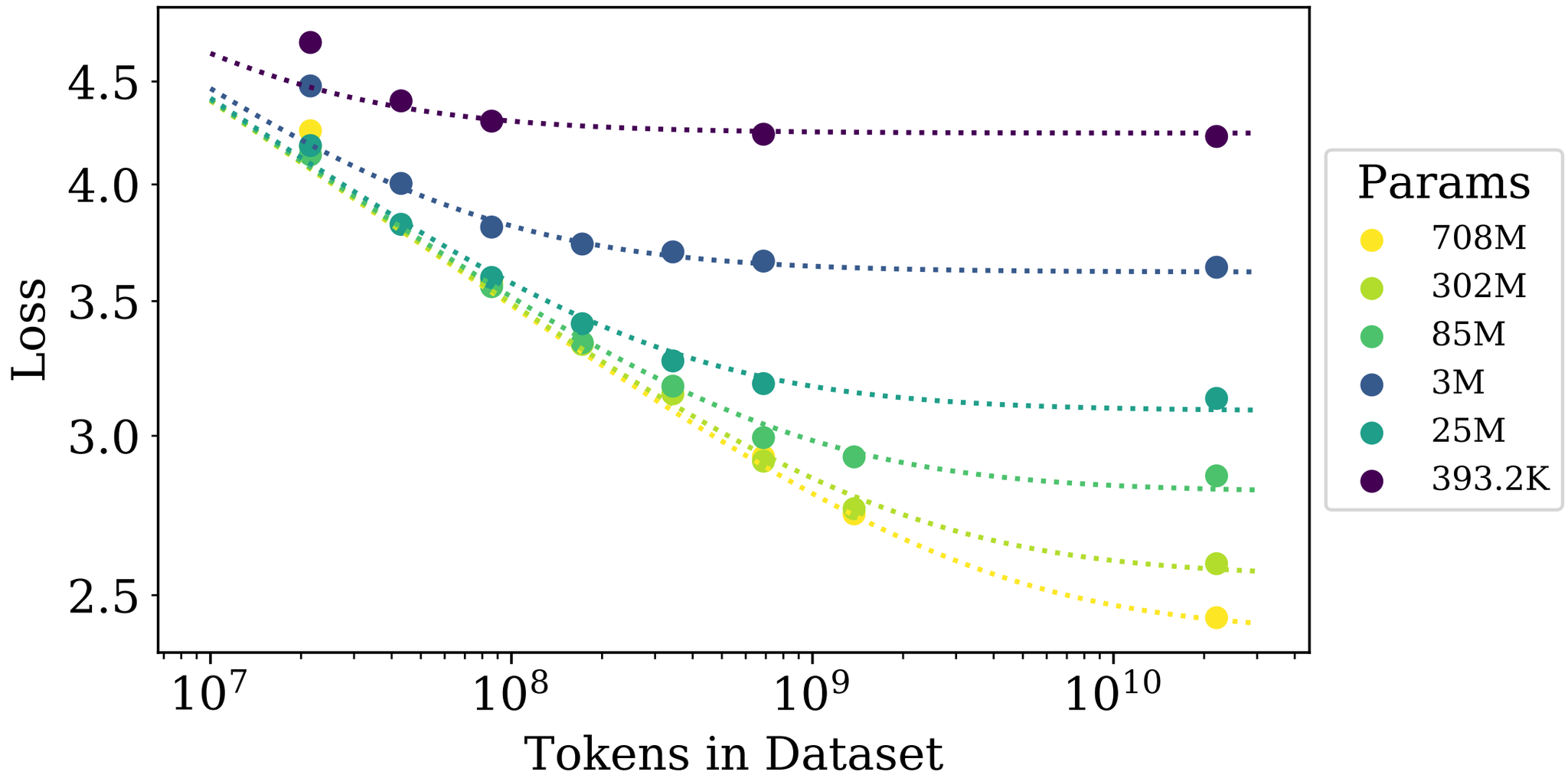
$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$$
$$+i\bar{\psi}D\psi + h.c.$$
$$+\bar{\psi}_i y_{ij}\psi_j\phi + h.c.$$
$$+|D_\mu\phi|^2 - V(\phi)$$

The “simplified” equation of (attempted) unification of forces

Can Laws Be Unified?

Unifying model size and data size:

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



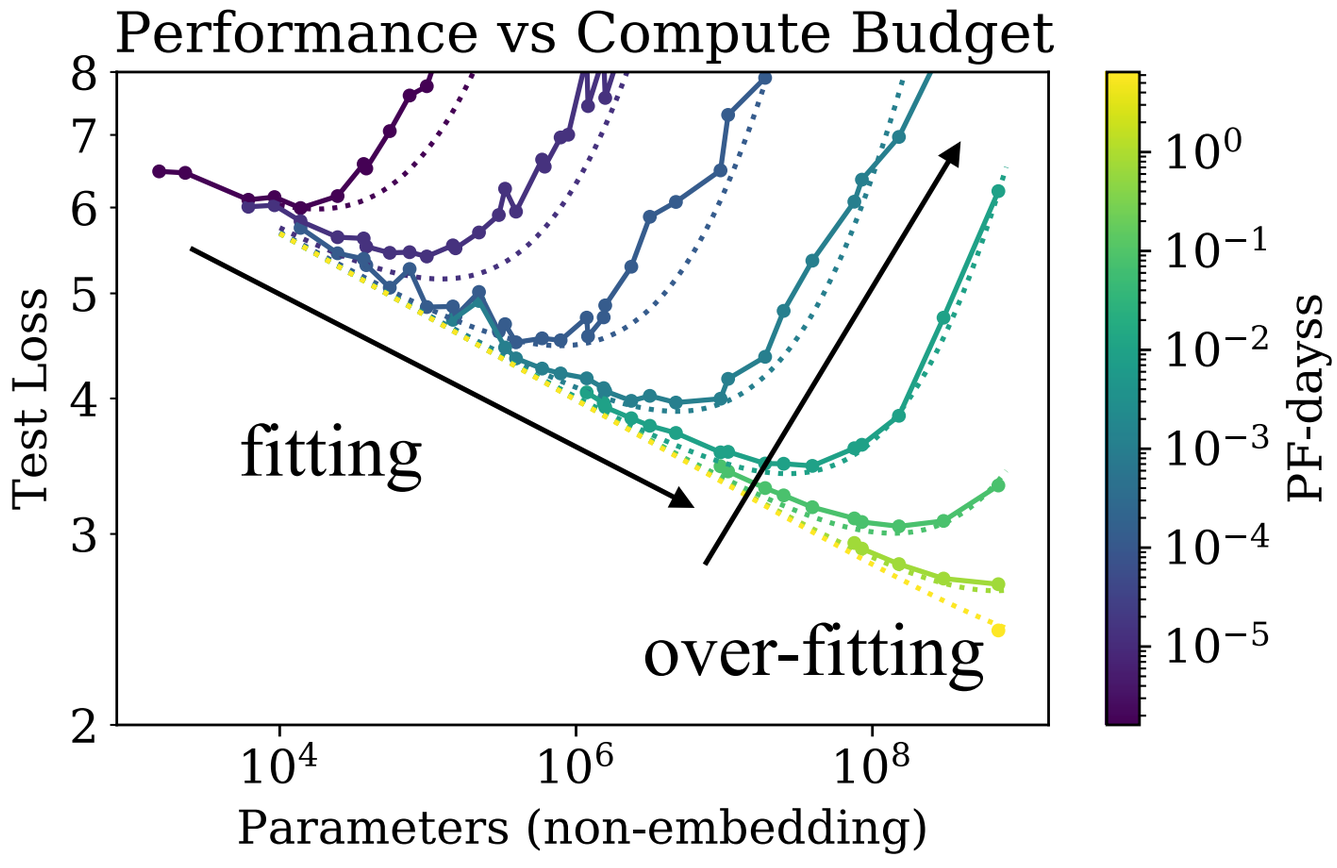
Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

The Size-Compute Law Expanded

Unifying model size and training compute:

$$L(N,S) = \left(\frac{N_c}{N}\right)^{\alpha_N} + \left(\frac{S_c}{S_{\min}(S)}\right)^{\alpha_S} = \left(\frac{N_c}{N}\right)^{\alpha_N} + \left(\frac{6NBS_c(1+B_{crit}(L)/B)}{C}\right)^{\alpha_S}$$

not so elegant…

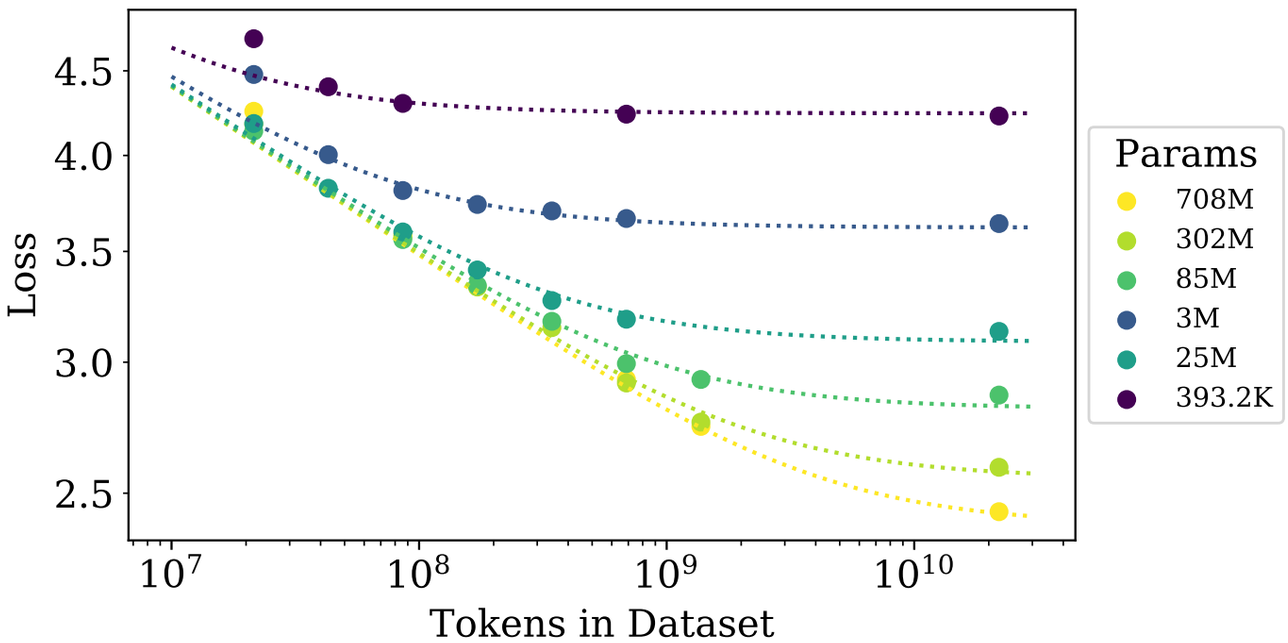


Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

Can Laws Be Unified?

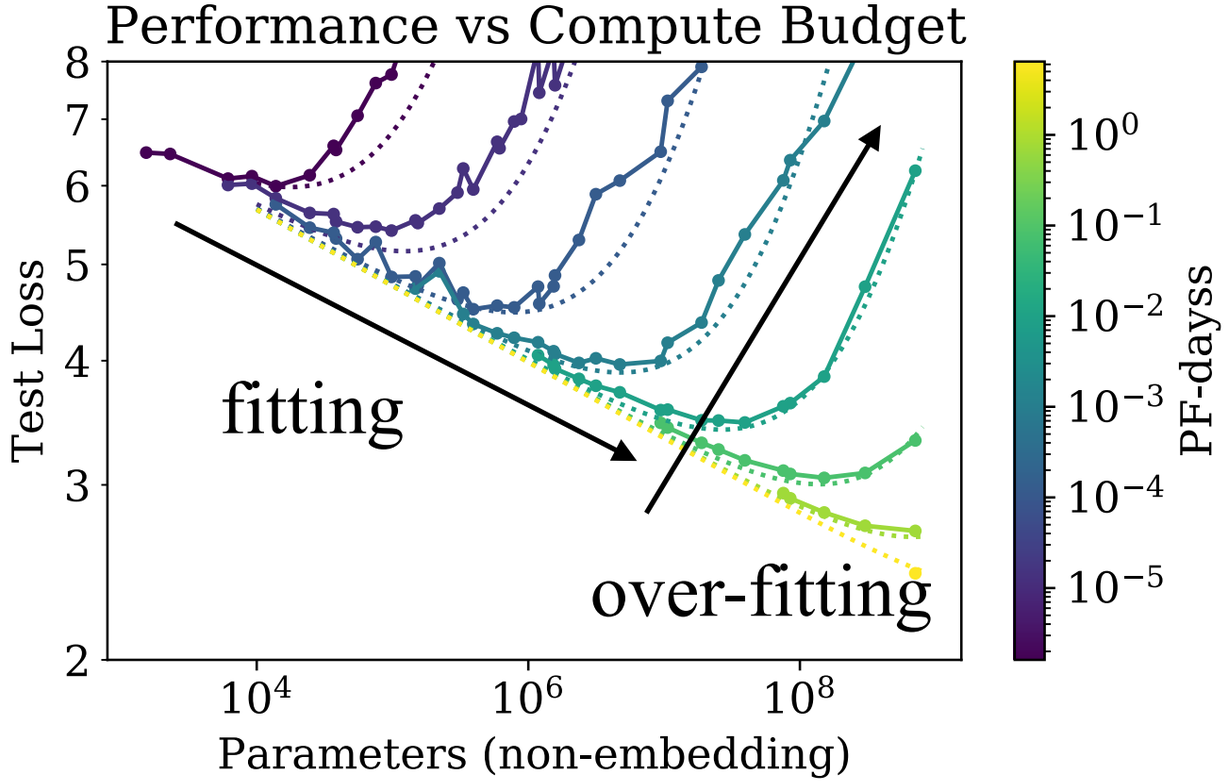
Unifying model size and data size:

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



Unifying model size and training compute:

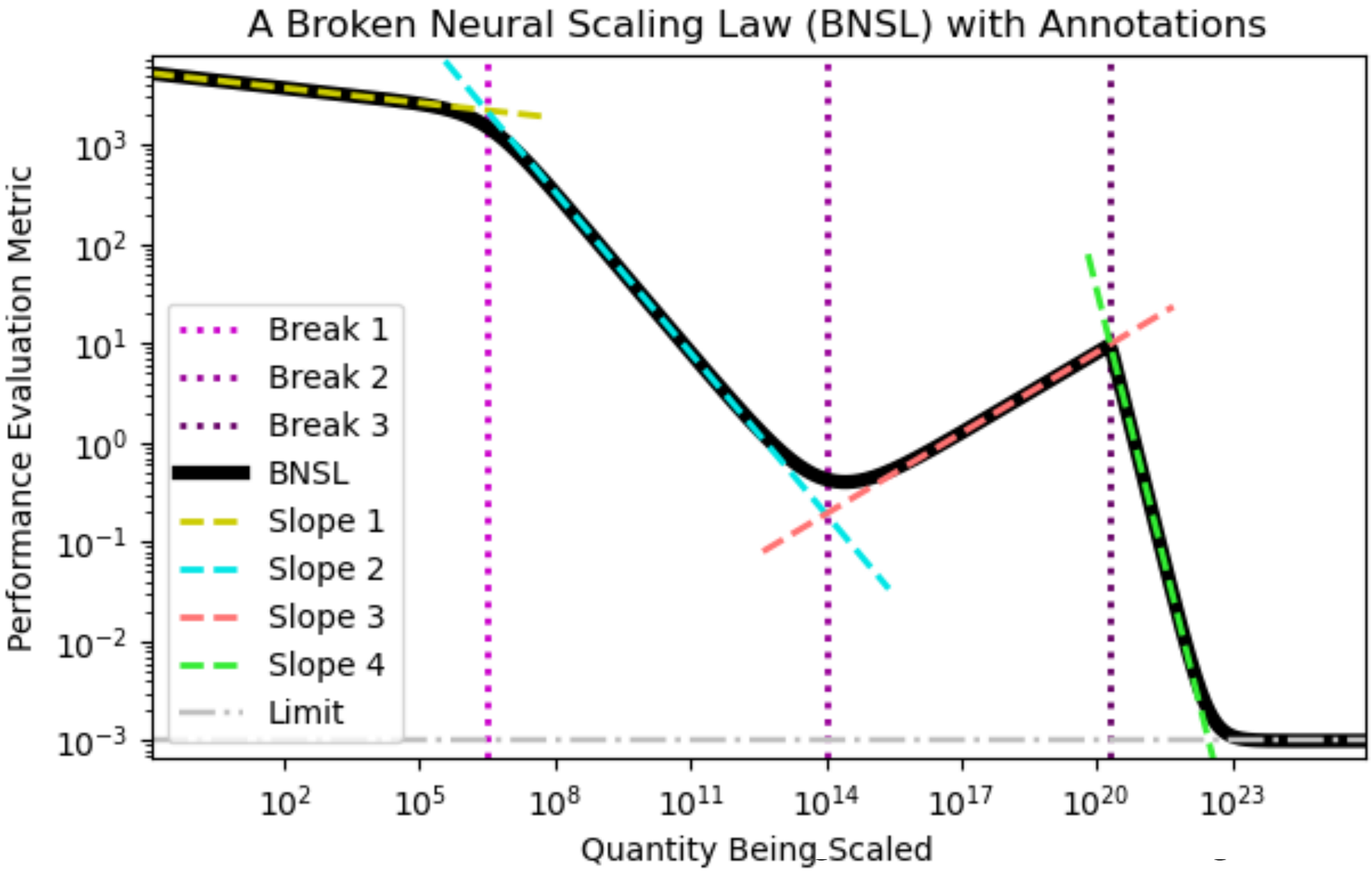
$$L(N, S) = \left(\frac{N_c}{N} \right)^{\alpha_N} + \left(\frac{S_c}{S_{\min}(S)} \right)^{\alpha_S}$$



Can they be further unified? The authors didn't say.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

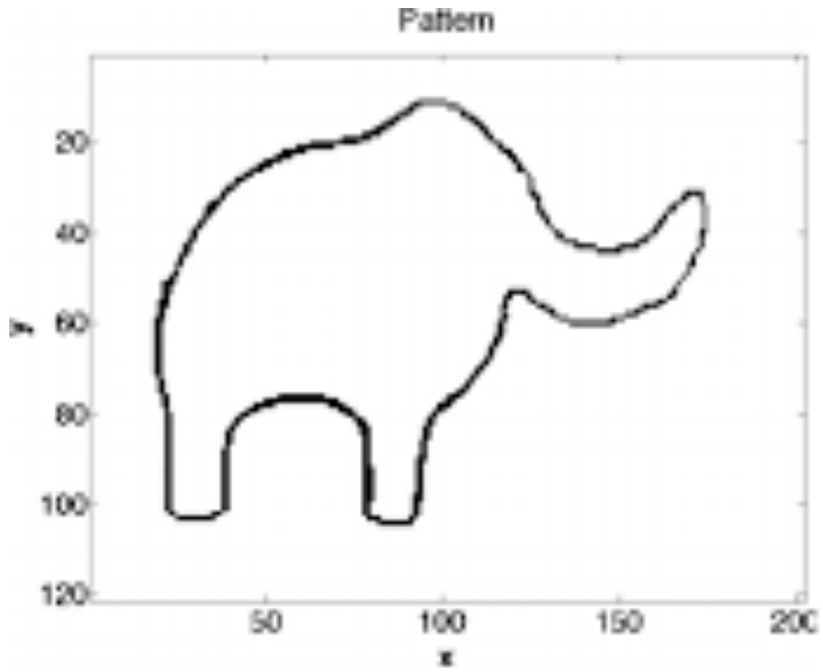
Some Law Become Complex



$$y = a + \left(bx^{-c_0}\right) \prod_{i=1}^{n} \left(1 + \left(\frac{x}{d_i}\right)^{1/f_i}\right)^{-c_i * f_i}$$

Caballero, Ethan, Kshitij Gupta, Irina Rish, and David Krueger. "Broken Neural Scaling Laws." In *The Eleventh International Conference on Learning Representations*.

Occam's Razor and Von Neumann's Elephant



https://en.wikipedia.org/wiki/Von_Neumann's_elephant

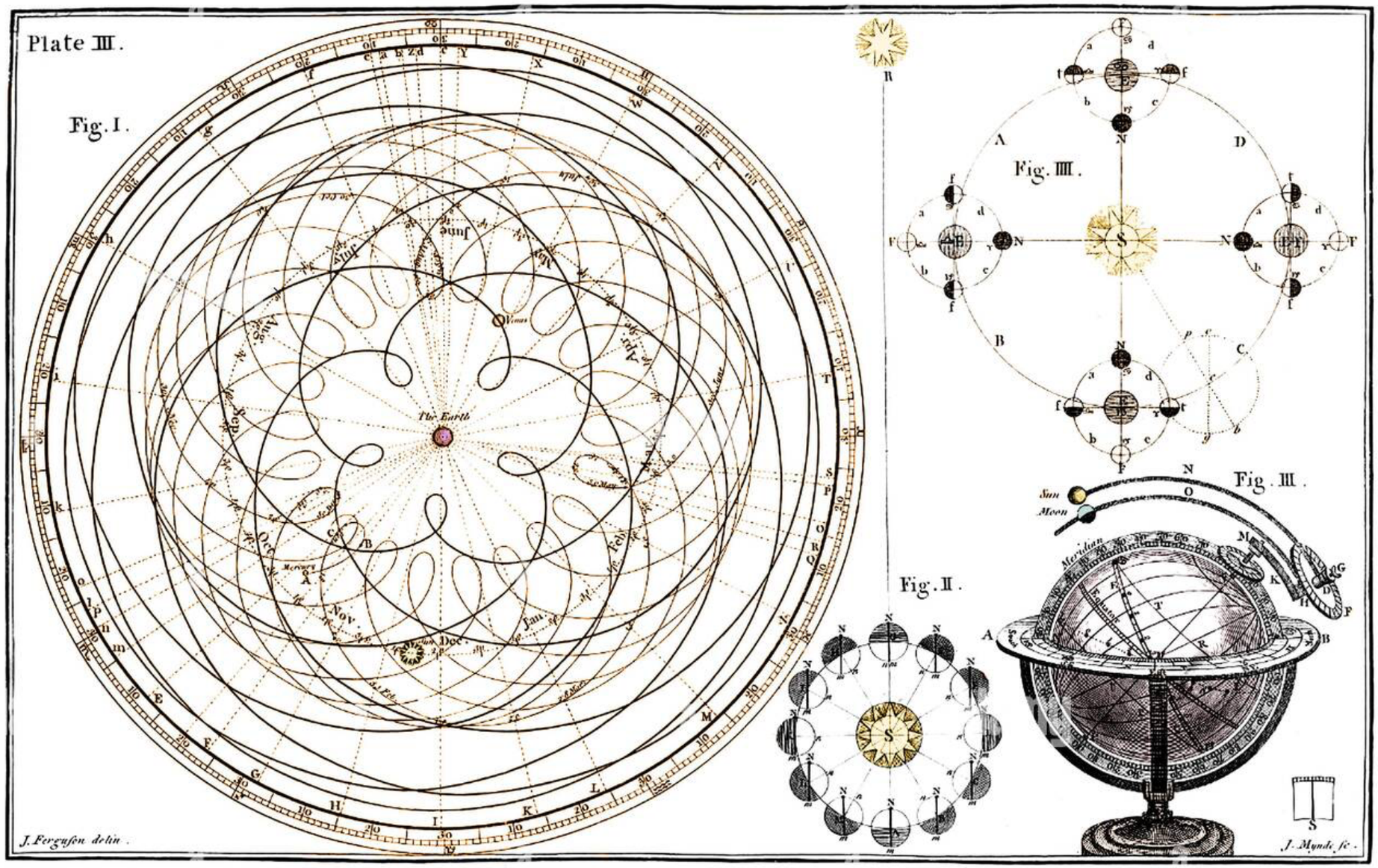
Occam's razor: "Simpler, better"

"With four parameters, I can fit an elephant, and with five, I can make him wiggle his trunk."

——Von Neumann

Note: selection of math form is also a parameter!

When Humans Overthink, They Overfit



Ptolemaic system trying to prove geocentric theory

Resource Allocation by Laws

model size

$$L(N) = (N_c/N)^{\alpha_N}$$

$$N \propto C^{\alpha_C^{\min}/\alpha_N},$$

training compute

$$L(C_{\min}) = (C_c^{\min}/C_{\min})^{\alpha_C^{\min}}$$

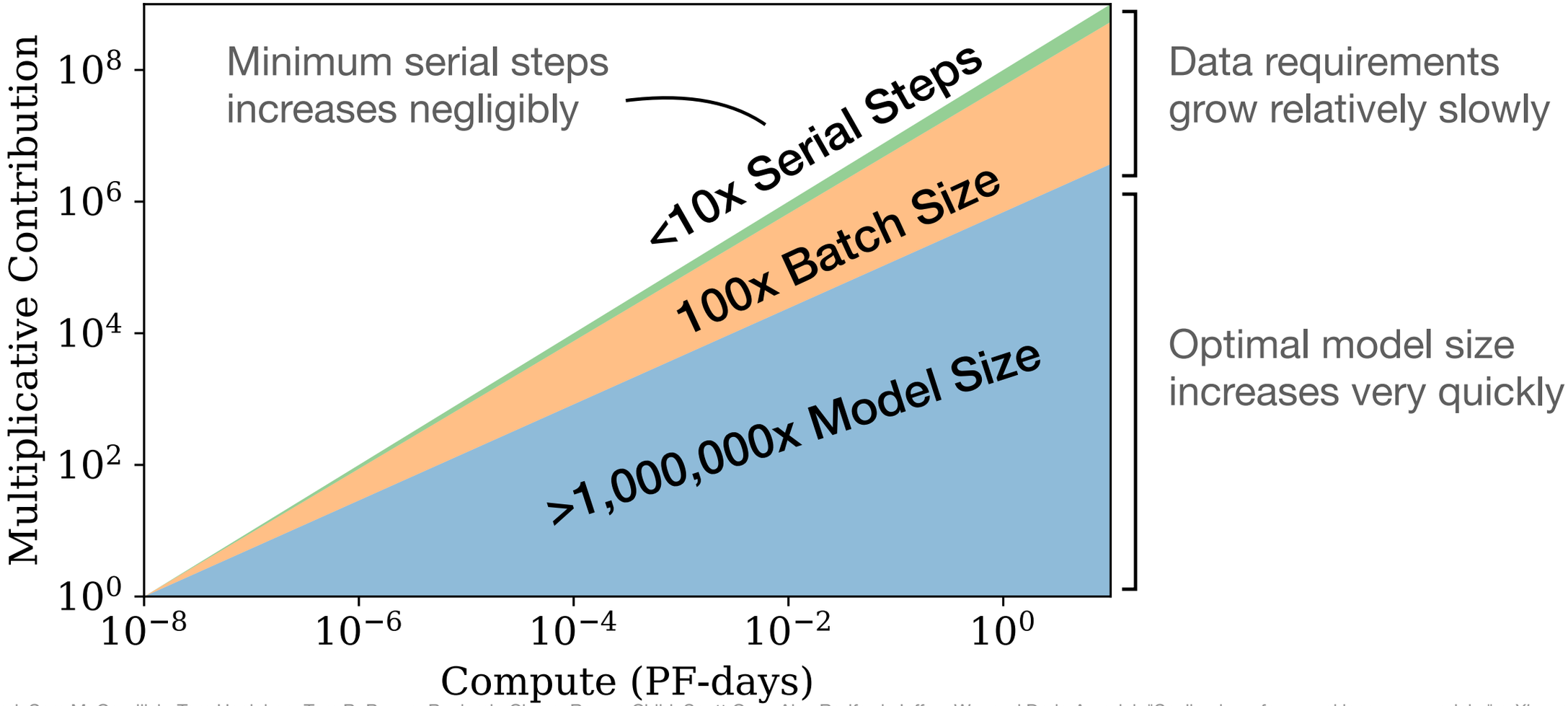
$$B \propto C^{\alpha_C^{\min}/\alpha_B},$$

$$S \propto C^{\alpha_C^{\min}/\alpha_S},$$

data size

$$L(D) = (D_c/D)^{\alpha_D}$$

$$D = B \cdot S$$



Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint*

Chinchilla's Law for Model and Data

Chinchilla

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

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

Kaplan (the one we just saw)

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

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

Approach	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50 (0.488, 0.502)	0.50 (0.501, 0.512)
2. IsoFLOP profiles	0.49 (0.462, 0.534)	0.51 (0.483, 0.529)
3. Parametric modelling of the loss	0.46 (0.454, 0.455)	0.54 (0.542, 0.543)
Kaplan et al. (2020)	0.73	0.27

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas et al. "Training Compute-Optimal Large Language Models."

Open Questions

1. **Does a unified law exist for all factors (i.e., how they interact)?**

2. **What causes these laws?**

3. **What determines the constants α_*?**

4. **Scaling laws for other factors, like data quality, composition, and diversity?**

5. **Scaling laws for other metrics?**
   (Much concurrent work)

1.2 Linguistics, Reasoning and Knowledge Acquisition

Allen Zhu's "Physics of LMs" series

https://physics.allen-zhu.com/,

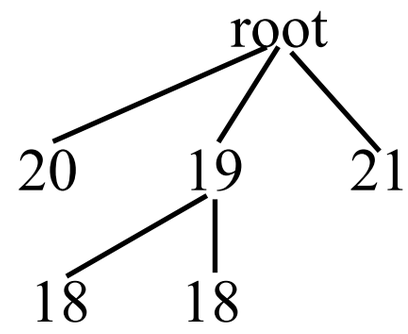
```
@misc{AllenZhu-icml2024-tutorial,
    author = {{Allen-Zhu}, Zeyuan},
    title = {{ICML 2024 Tutorial: Physics of Language Models}},
    year = {2024},
    month = {July},
    note = {Project page: \url{https://physics.allen-zhu.com/}}
}
```

Part 1: Language Acquisition

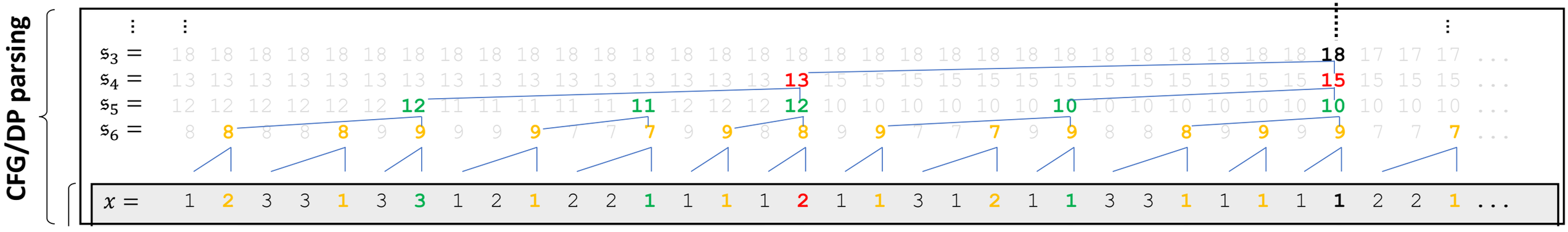
Investigation Tool: context-free grammar (CFG)

CFG Production rules: Each time randomly replace a symbol with a string

root |->20 21	19|->18 16 18	16|->15 15	13|->11 12	10|->8 9 9	7|->2 2 1
root |->20 19 21	19|->17 18	16|->13 15 13	13|->12 11 12	10|->9 7 9	7|->3 2 2
root |->21 19 19	19|->18 18	16|->14 13	13|->10 12 11	10|->7 9 9	7|->3 1 2
root |->20 20	20|->16 16	16|->14 14	14|->10 12	11|->8 8	7|->3 2
	20|->16 17	17|->15 14 13	14|->12 10 12	11|->9 7	8|->3 1 1
	20|->17 16 18	17|->14 15	14|->12 11	11|->9 7 7	8|->1 2
	21|->18 17	17|->15 14	14|->10 12 12	12|->7 9 7	8|->3 3 1
	21|->17 16	18|->14 15 13	15|->10 11 11	12|->9 8	9|->1 2 1
	21|->16 17 18	18|->15 13 13	15|->11 11 10	12|->8 8 9	9|->3 3
	21|->16 18	18|->13 15	15|->10 10		9|->1 1
			15|->12 12 11		

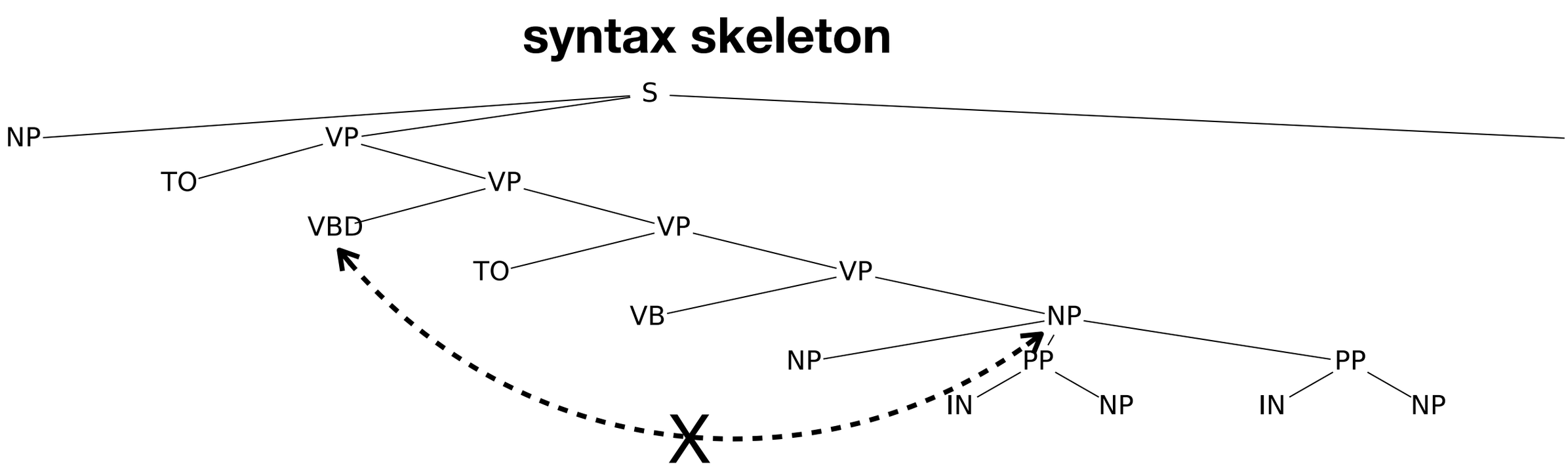


Example:



Allen-Zhu, Zeyuan, and Yuanzhi Li. "Physics of language models: Part 1, context-free grammar." *arXiv preprint arXiv:2305.13673* (2023).

CFG ≈ Syntax Structure



Difference: CFG has little semantics. There is little association between tokens far away.

(forcing semantics into CFG requiring expensive rules)

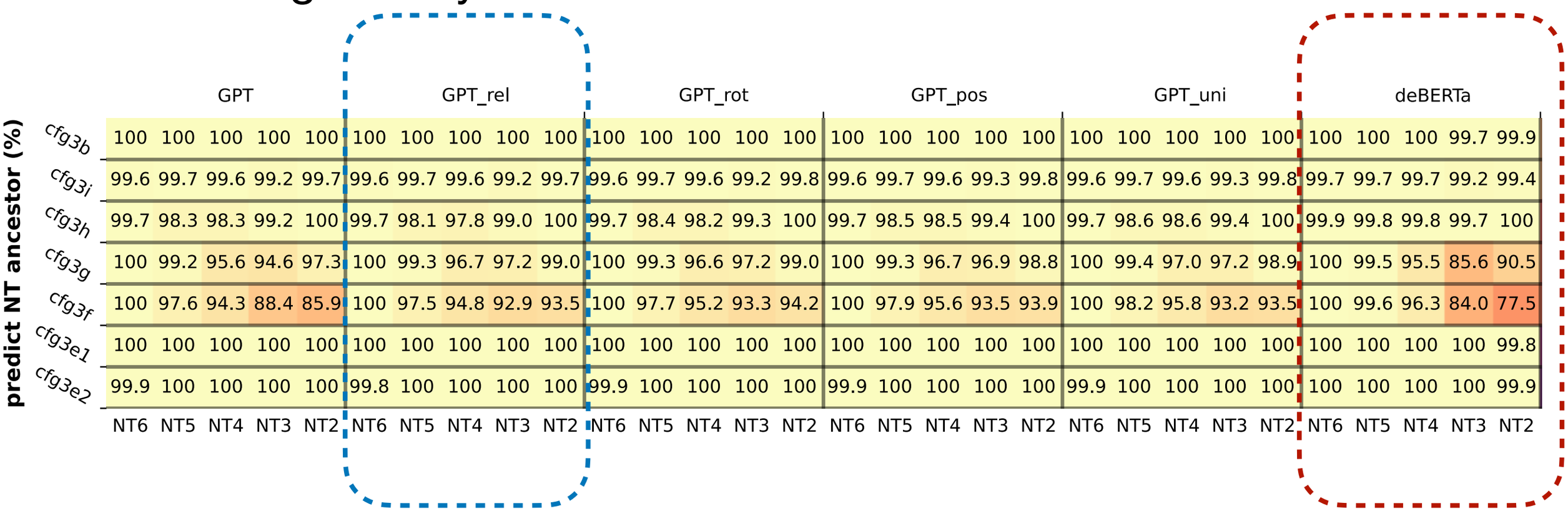
Allen-Zhu, Zeyuan, and Yuanzhi Li. "Physics of language models: Part 1, context-free grammar." *arXiv preprint arXiv:2305.13673* (2023).

Generative LM Learns CFG Better

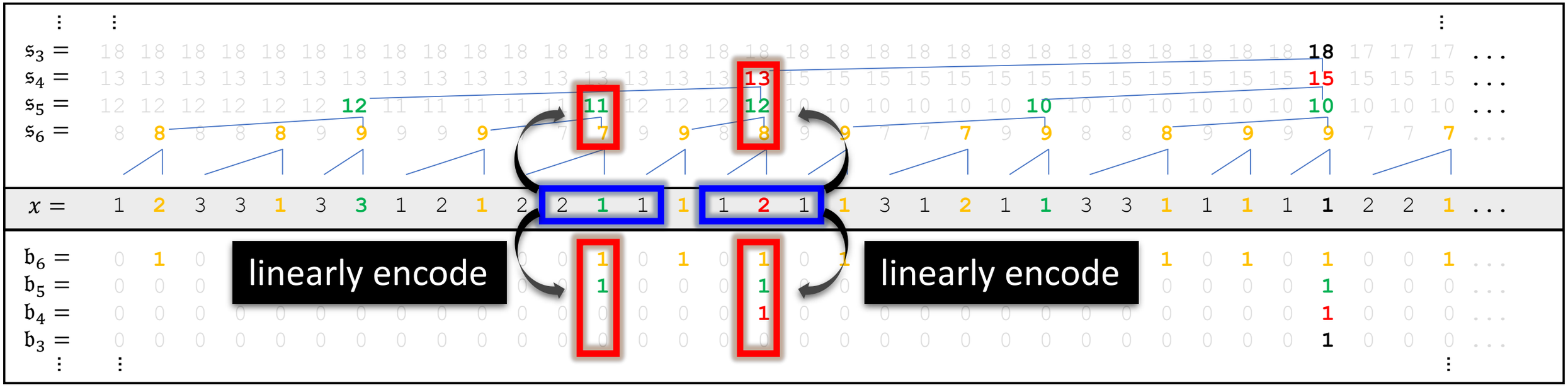
Task: predicting the NT ancestor (the symbol generating)

relative position
generally better

encoder-based

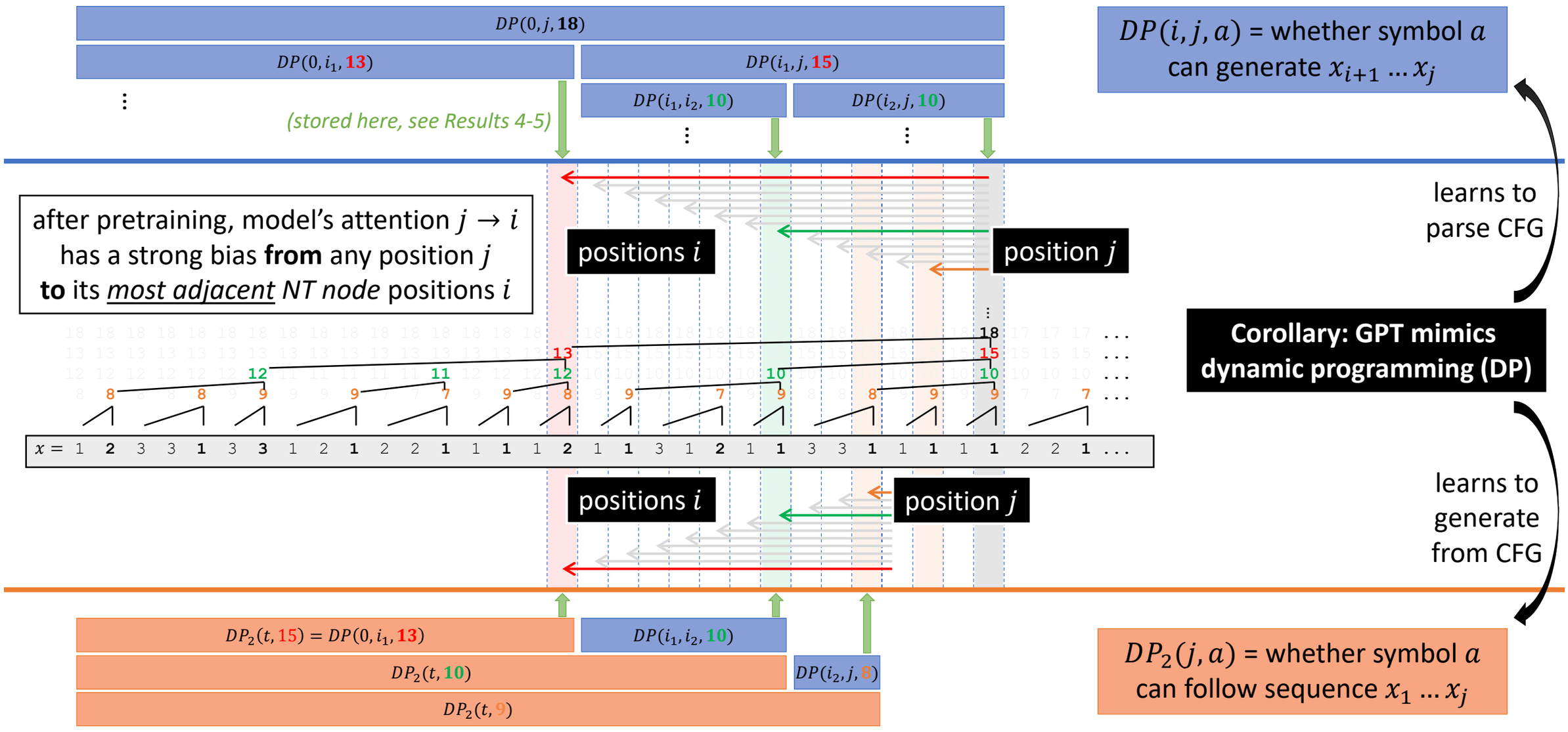


Tokens Encoding Local Syntax

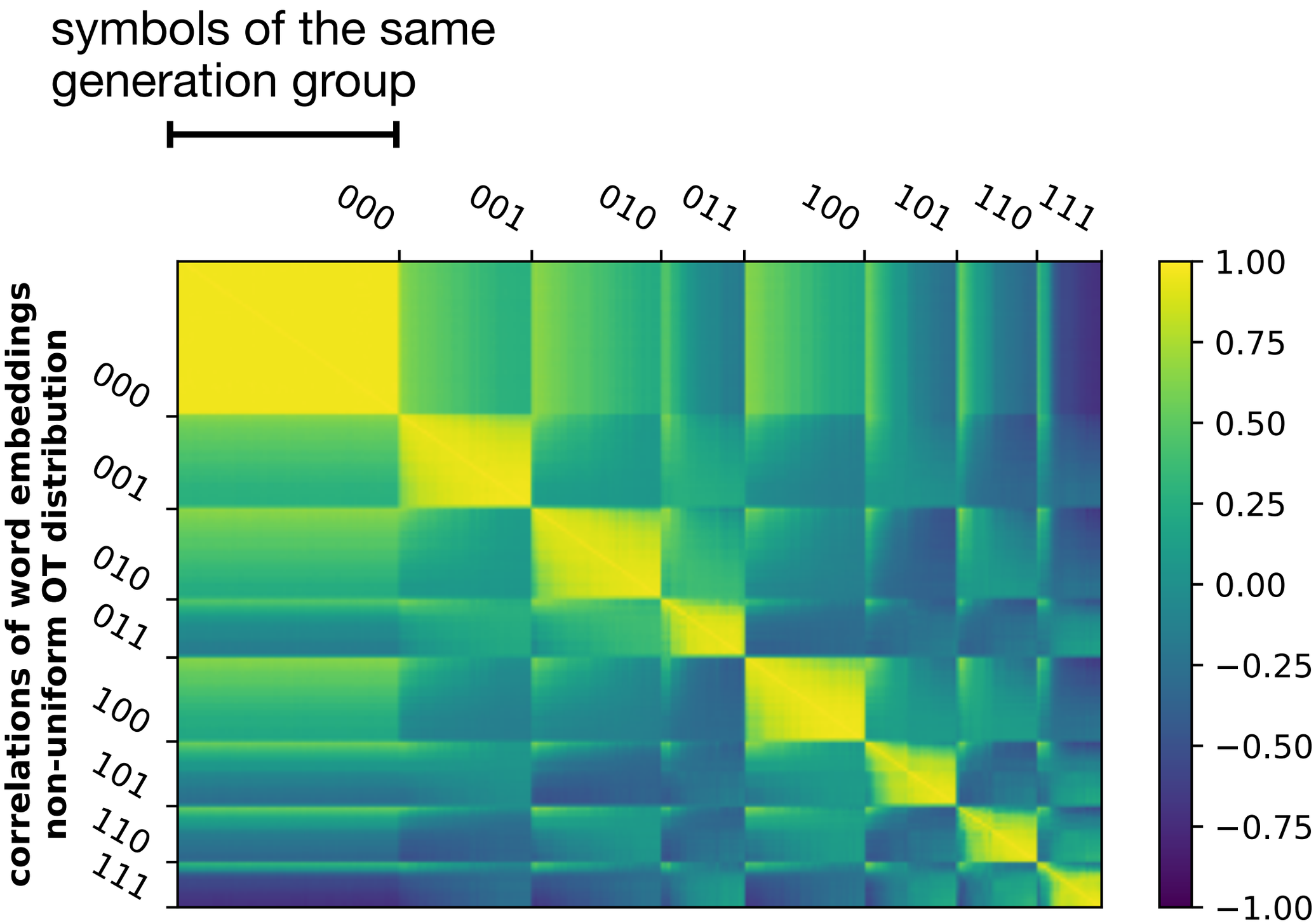


LM Encodes Dynamic Programming Features

Dynamic programming (DP) is the standard algorithm for parsing CFG that humans use.

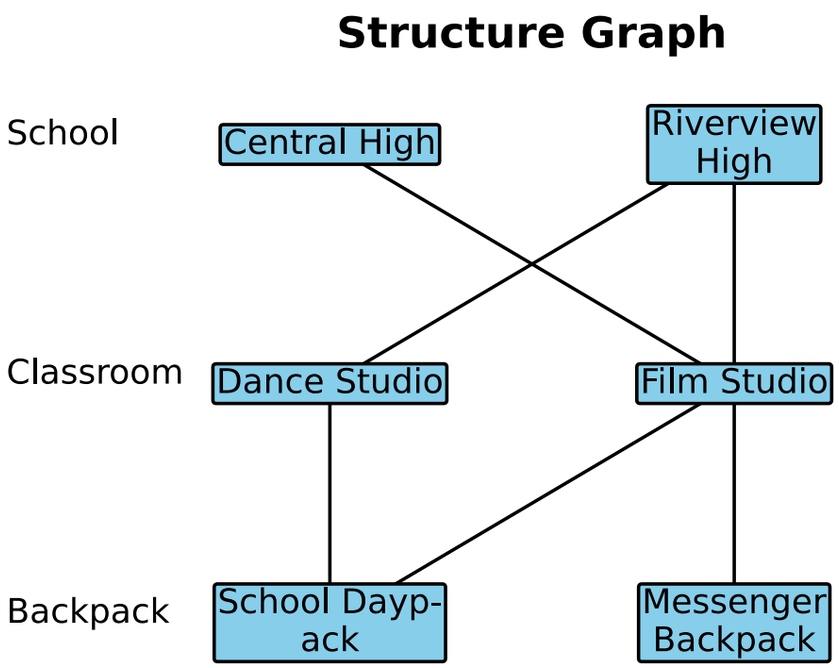


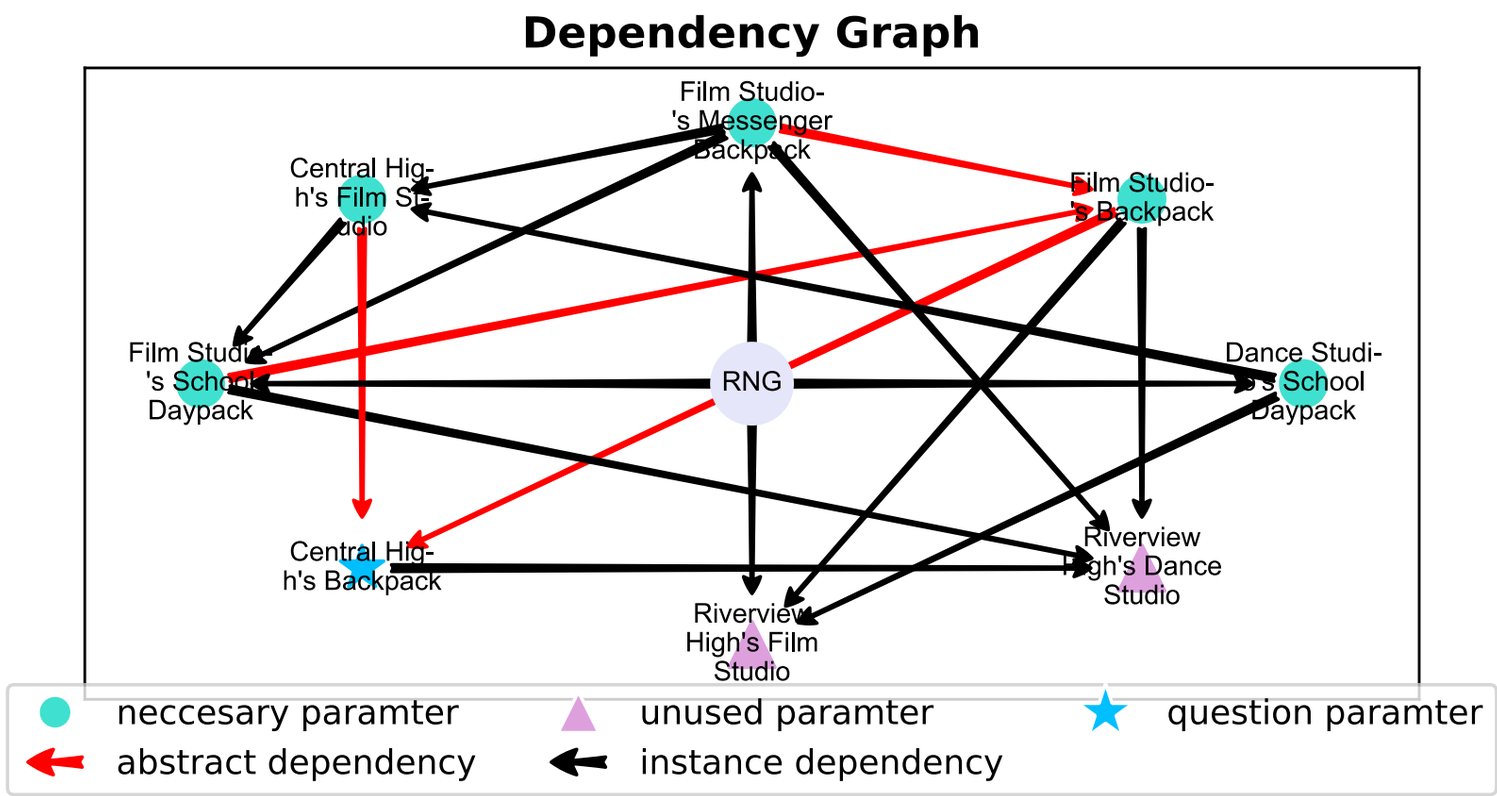
Word Embeddings Reflects CFG Syntax



Part 2: Reasoning Acquisition

Synthetic math reasoning construction

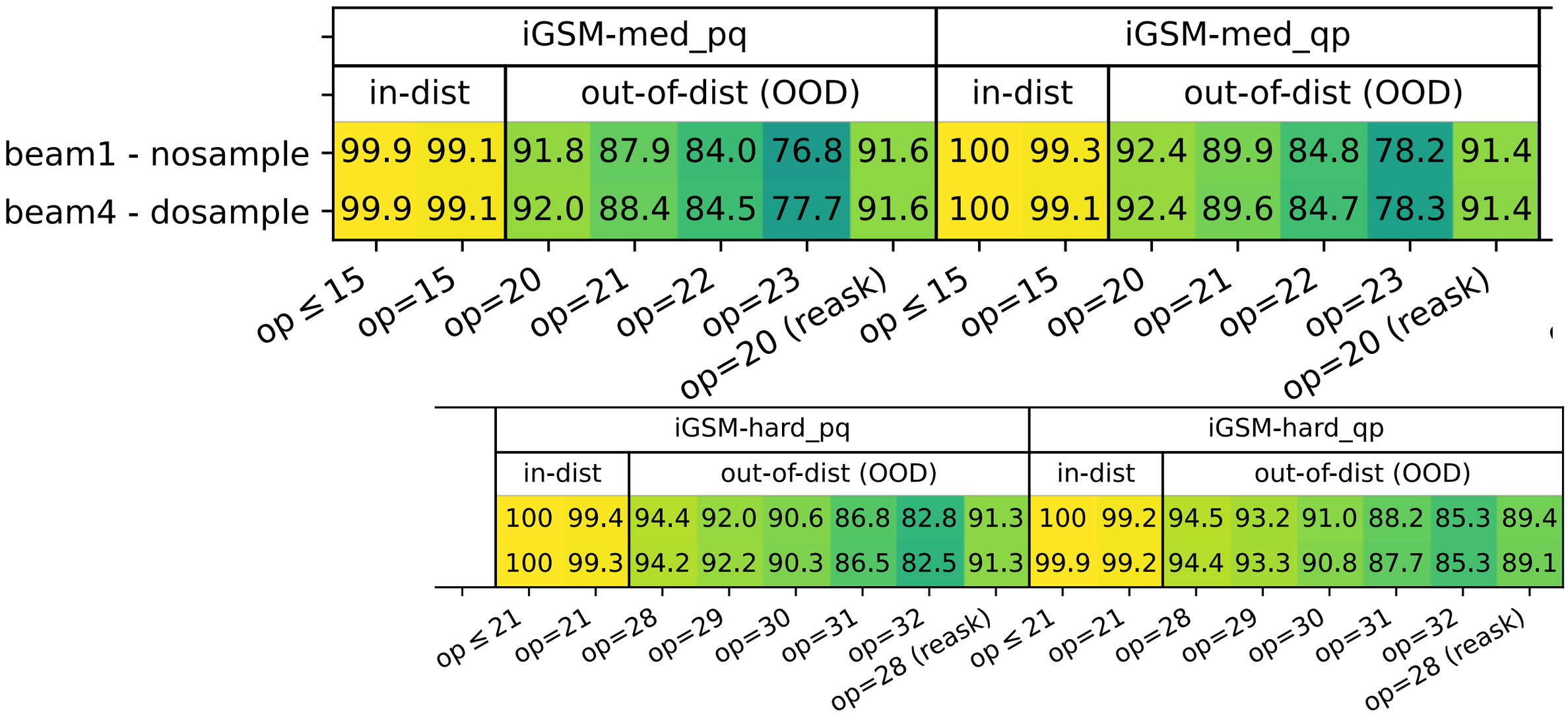




(Problem - Easy) The number of each Riverview High's Film Studio equals 5 times as much as the sum of each Film Studio's Backpack and each Dance Studio's School Daypack. The number of each Film Studio's School Daypack equals 12 more than the sum of each Film Studio's Messenger Backpack and each Central High's Film Studio. The number of each Central High's Film Studio equals the sum of each Dance Studio's School Daypack and each Film Studio's Messenger Backpack. The number of each Riverview High's Dance Studio equals the sum of each Film Studio's Backpack, each Film Studio's Messenger Backpack, each Film Studio's School Daypack and each Central High's Backpack. The number of each Dance Studio's School Daypack equals 17. The number of each Film Studio's Messenger Backpack equals 13. *How many Backpack does Central High have?*

(2.1)

LMs Can Generalize Reasoning Length



On questions with longer reasoning chains (larger "**op**"), the models performs well.

Depth Is Crucial For Reasoning

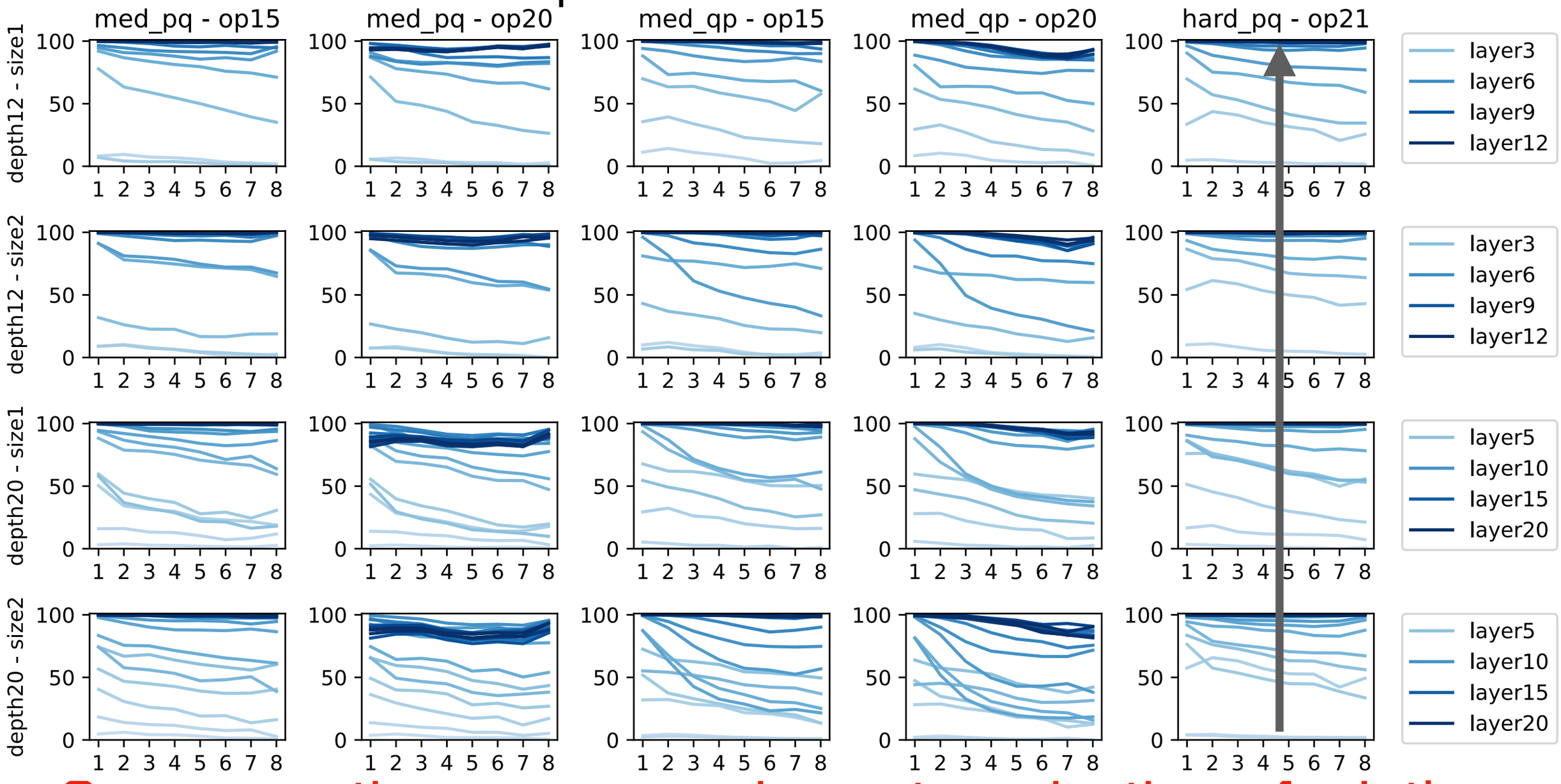
	iGSM-med_pq						iGSM-med_qp					
	in-dist		out-of-dist (OOD)				in-dist		out-of-dist (OOD)			
	op ≤ 15	op=15	op=20	op=21	op=22	op=23	op ≤ 15	op=15	op=20	op=21	op=22	op=23
dep4 - size1 - head21	99.5	92.7	74.7	68.0	62.4	54.5	99.4	93.3	73.3	66.8	61.1	54.6
dep4 - size2 - head30	99.6	94.7	74.2	67.9	61.6	53.1	99.4	94.5	78.1	71.9	65.7	58.8
dep8 - size1 - head15	100	98.8	89.7	86.5	82.8	76.8	100	99.2	92.4	88.5	84.2	78.7
dep8 - size2 - head21	100	99.3	93.7	91.6	88.3	83.6	99.9	99.0	90.2	87.1	83.3	76.3
dep12 - size1 - head12	100	99.3	92.0	88.9	84.2	77.9	100	99.4	92.2	89.2	83.9	77.9
dep12 - size2 - head17	100	99.5	94.0	91.9	89.0	82.7	100	99.0	90.8	85.4	80.2	73.2
dep16 - size1 - head10	100	99.6	94.6	91.9	87.9	82.7	100	99.5	89.9	85.0	79.1	71.1
dep16 - size2 - head15	100	99.8	95.9	93.7	90.4	86.5	100	99.8	95.6	93.5	90.3	84.3
dep20 - size1 - head9	100	99.8	95.5	93.6	90.0	86.3	100	99.6	94.8	91.4	87.4	80.4
dep20 - size2 - head13	100	99.8	95.8	93.3	89.2	84.4	100	99.6	93.7	91.8	87.4	81.3

	iGSM-hard_pq							iGSM-hard_qp						
	in-dist		out-of-dist (OOD)					in-dist		out-of-dist (OOD)				
	op ≤ 21	op=21	op=28	op=29	op=30	op=31	op=32	op ≤ 21	op=21	op=28	op=29	op=30	op=31	op=32
dep4 - size1 - head21	98.9	90.8	72.4	67.7	62.1	57.1	50.6	99.1	89.8	69.4	62.2	57.8	52.3	45.7
dep4 - size2 - head30	97.0	71.7	46.3	40.6	37.0	32.3	27.3	99.4	92.1	74.5	69.5	64.7	59.1	53.2
dep8 - size1 - head15	100	99.1	94.6	92.0	89.7	86.4	82.2	100	99.0	92.2	89.6	86.2	82.4	77.3
dep8 - size2 - head21	100	99.2	93.6	91.3	88.6	85.6	82.6	100	99.1	93.5	91.3	89.1	85.7	81.2
dep12 - size1 - head12	100	99.5	96.0	94.1	91.0	88.5	84.3	100	99.3	95.3	93.0	91.9	88.0	84.5
dep12 - size2 - head17	100	99.8	97.1	95.5	93.5	91.8	88.0	100	99.5	94.5	91.9	88.9	86.8	81.3
dep16 - size1 - head10	100	99.6	97.0	95.2	94.2	92.2	88.5	100	99.4	95.8	93.8	92.4	88.9	85.8
dep16 - size2 - head15	100	99.7	97.5	96.3	95.1	92.9	89.5	100	99.8	97.3	96.0	94.2	91.9	88.9
dep20 - size1 - head9	100	99.8	97.0	95.1	94.0	91.0	87.4	100	99.6	96.6	94.5	92.8	90.1	86.5
dep20 - size2 - head13	100	99.8	98.0	96.7	95.9	93.9	90.9	100	99.9	97.5	96.0	95.2	92.4	89.7

Depth Is Crucial For Reasoning

Possible Reason: per token reasoning step

score: variable relevance prediction



Open question: more precise categorization of relation between model depth v.s. reasoning depth

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.

Trend:

	iGSM-med_pq							iGSM-med_qp						
	in-dist		out-of-dist (OOD)					in-dist		out-of-dist (OOD)				
original	99.9	99.1	92.0	88.4	84.5	77.7	91.6	100	99.3	92.4	89.9	84.8	78.3	91.4
retry | retryrate0.05	100	99.7	94.8	93.0	88.5	83.9	94.5	100	99.4	93.3	89.2	85.7	78.8	92.4
retry | retryrate0.05 (with mask)	100	99.5	93.1	89.5	84.6	78.5	93.2	100	99.6	95.2	91.3	88.4	82.3	92.6
retry | retryrate0.1	100	99.7	97.1	95.8	93.3	90.6	95.2	99.9	99.6	94.0	90.6	87.6	81.7	91.9
retry | retryrate0.1 (with mask)	100	99.8	96.6	94.2	91.5	87.5	96.9	100	99.8	96.0	93.8	89.9	84.0	92.5
retry | retryrate0.2	100	99.9	97.9	96.8	95.2	91.6	96.0	100	99.8	96.7	94.5	92.0	88.3	93.4
retry | retryrate0.2 (with mask)	100	99.9	97.2	95.5	93.0	87.8	95.8	100	99.9	97.4	96.2	94.4	90.9	93.7
retry | retryrate0.4	100	99.8	97.6	95.8	94.3	90.6	97.6	100	99.9	97.7	95.4	93.4	90.2	93.4
retry | retryrate0.5	100	99.9	98.3	96.8	95.8	93.6	96.1	100	99.9	98.7	98.0	96.9	94.5	96.0
retry | retryrate0.5 (with mask)	100	100	98.8	97.9	96.8	94.8	96.0	100	99.8	97.9	96.9	95.2	93.2	91.5

op ≤ 15 | op=15 | op=20 | op=21 | op=22 | op=23 | op=20 (reask) | op ≤ 15 | op=15 | op=20 | op=21 | op=22 | op=23 | op=20 (reask)

How Does Retry Data Benefit?

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.

Fine-Tuning Retry Doesn't Work

finetuning

original

retryrate0.2 | retry (finetune, lora qv4e8)
retryrate0.2 | retry (finetune, lora qv4e8) | with mask
retryrate0.2 | retry (finetune, lora qv8e16)
retryrate0.2 | retry (finetune, lora qv8e16) | with mask
retryrate0.2 | retry (finetune, lora qv16e32)
retryrate0.2 | retry (finetune, lora qv16e32) | with mask
retryrate0.2 | retry (finetune, lora qv32e64)
retryrate0.2 | retry (finetune, lora qv32e64) | with mask
retryrate0.2 | retry (finetune, lora qv64e128)
retryrate0.2 | retry (finetune, lora qv64e128) | with mask
retryrate0.2 | retry (finetune, lora qv128e256)
retryrate0.2 | retry (finetune, lora qv128e256) | with mask
retryrate0.2 | retry (finetune, lora qv256e512)
retryrate0.2 | retry (finetune, lora qv256e512) | with mask

retryrate0.2 | retry (continued pretrain)
retryrate0.2 | retry (continued pretrain) | with mask
retryrate0.2 | retry (pretrain)
retryrate0.2 | retry (pretrain) | with mask

(continued) pretraining

iGSM-med_pq							iGSM-med_qp						
in-dist		out-of-dist (OOD)					in-dist		out-of-dist (OOD)				
99.9	99.1	92.0	88.4	84.5	77.7	91.6	100	99.3	92.4	89.9	84.8	78.3	91.4
99.9	98.1	84.0	77.5	71.8	63.3	92.7	99.0	98.0	81.6	76.0	68.4	59.4	90.3
100	99.0	91.2	88.2	82.4	74.3	93.5	99.8	99.3	90.6	86.0	80.5	73.0	91.8
100	98.6	86.4	82.4	76.0	68.9	93.2	99.6	98.4	84.9	79.4	72.3	63.5	91.5
100	99.2	92.0	88.7	83.7	76.7	93.2	99.8	99.4	92.2	88.0	82.2	75.3	92.1
100	98.8	88.3	84.4	78.4	72.0	93.8	99.8	98.7	87.8	82.8	76.3	68.1	91.3
100	99.3	91.8	88.7	83.6	75.9	93.8	99.9	99.3	92.0	88.3	83.0	76.4	92.0
100	98.7	89.8	86.2	81.1	74.7	93.0	99.8	99.0	87.9	82.4	76.5	69.9	91.5
100	99.2	92.4	89.4	84.2	77.8	93.4	99.9	99.5	92.9	88.9	83.9	77.2	92.2
99.9	99.1	90.3	86.1	81.1	76.5	92.7	99.9	99.1	88.9	84.5	78.4	71.4	92.0
100	99.4	92.6	89.8	85.4	79.3	93.2	99.9	99.5	92.8	88.9	83.8	77.5	92.4
99.9	99.0	90.8	86.9	82.2	76.5	93.9	99.8	99.2	90.6	86.9	81.9	75.4	92.0
100	99.4	92.9	90.0	84.8	79.0	93.9	99.9	99.6	92.7	88.9	84.3	78.3	92.7
100	99.1	92.0	87.9	84.1	78.8	94.4	99.9	99.2	90.1	85.5	80.7	74.1	92.1
100	99.3	92.5	90.1	85.5	79.1	94.4	99.9	99.6	92.9	88.6	84.4	77.3	92.5
100	99.9	97.4	95.6	92.8	88.3	95.9	100	100	97.9	95.0	92.5	88.6	93.7
100	99.9	97.0	94.7	91.7	87.7	95.7	100	99.9	97.7	94.9	92.6	88.2	93.5
100	99.9	97.9	96.8	95.2	91.6	96.0	100	99.8	96.7	94.5	92.0	88.3	93.4
100	99.9	97.2	95.5	93.0	87.8	95.8	100	99.9	97.4	96.2	94.4	90.9	93.7

op ≤ 15 | op=15 | op=20 | op=21 | op=22 | op=23 | op=20 (reask) | op ≤ 15 | op=15 | op=20 | op=21 | op=22 | op=23 | op=20 (reask)

Part 3: Knowledge Acquisition

Biological dataset:

Anya Briar Forger is a renowned social media strategist and community manager. She is currently working as a Marketing Manager at Meta Platforms. She completed her graduation from MIT with a degree in Communications. She was born on 2nd October 1996 in Princeton, NJ and was brought up in the same city. She later moved to Menlo Park in California to be a part of Facebook's team. She is an avid reader and loves traveling.

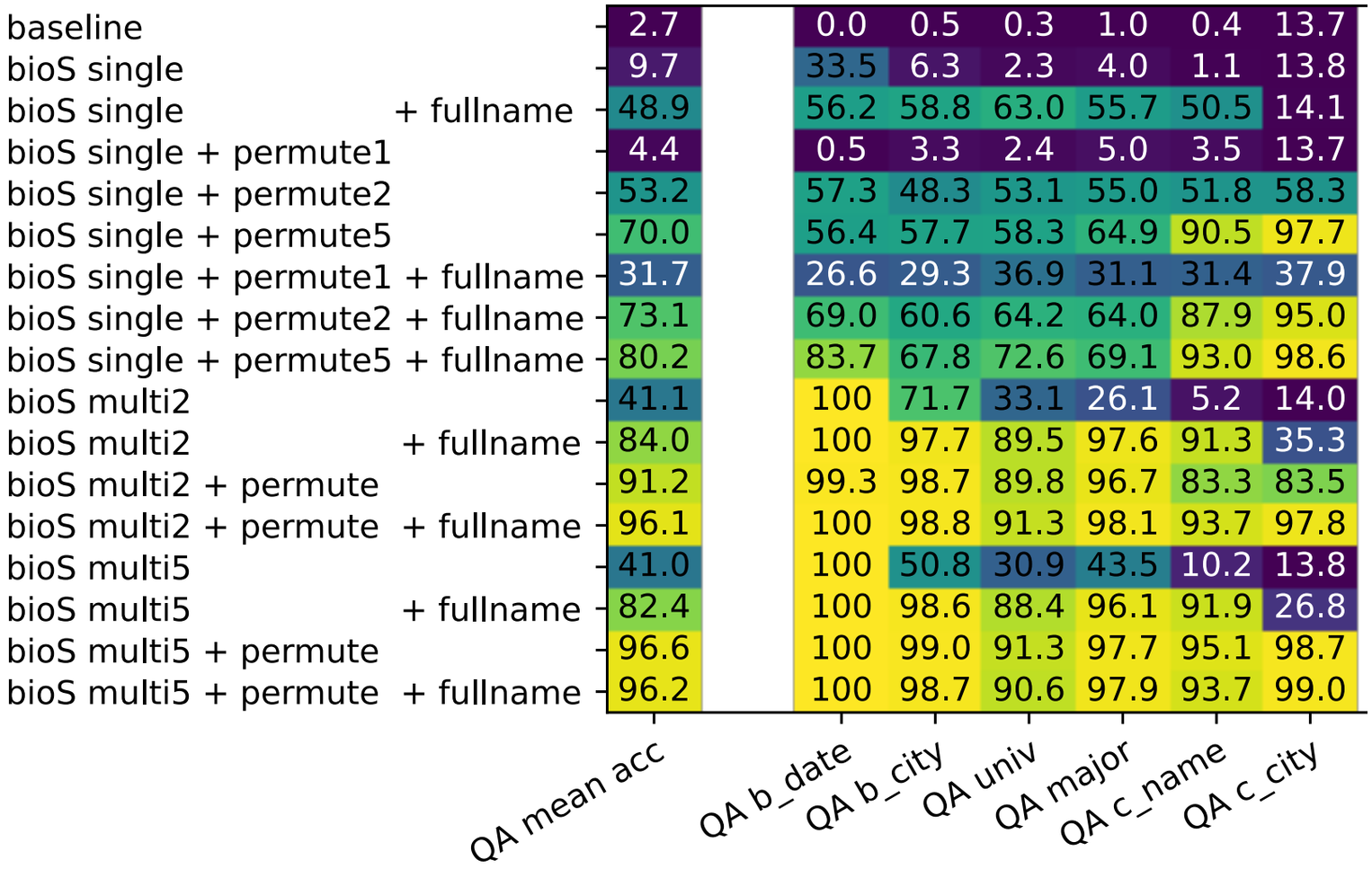
Relational queries:

1. What is the birth date of Anya Briar Forger?
   Answer: October 2, 1996.
2. What is the birth city of Anya Briar Forger?
   Answer: Princeton, NJ.
3. Which university did Anya Briar Forger study?
   Answer: Massachusetts Institute of Technology.

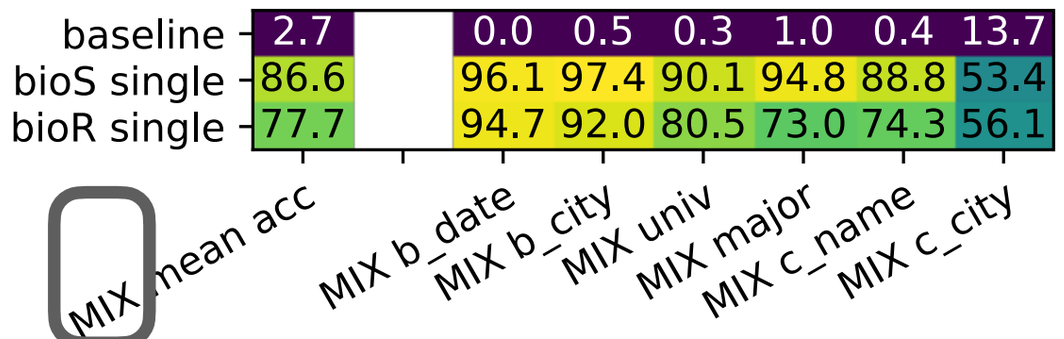
Data Diversity Is Crucial for Emergent QA

QA is high when:

Either training data is **diverse**:

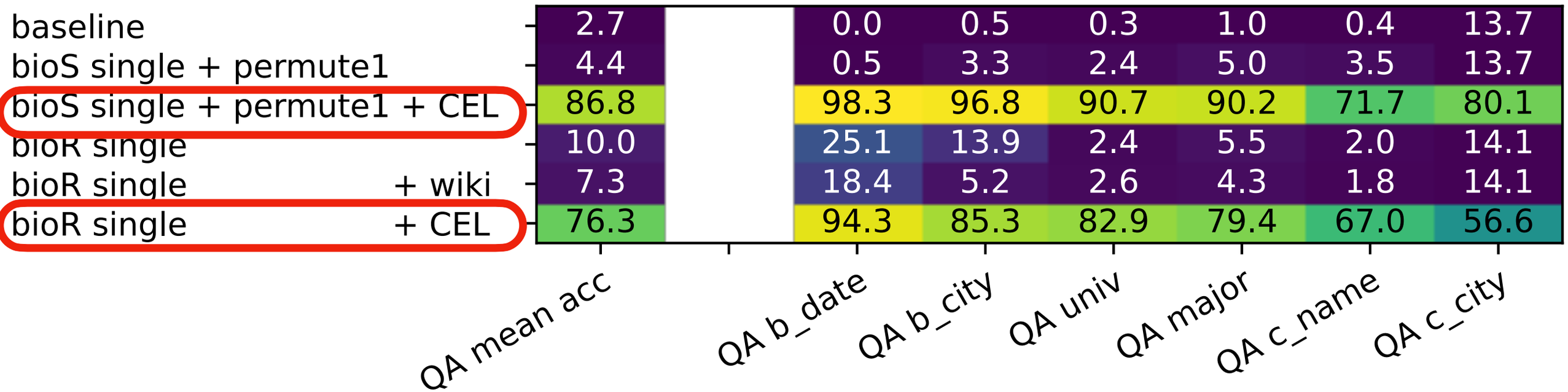


Or training data sees some QA examples:



mixed training data

Even Celebrity Data Helps Minority



LLMs Struggle With Inverse Query

Training: A, B

Forward Inference: A, **?**

Reverse Inference: **?** , B

Jane Austen Novel Task

Inverse search: "In <Pride and Prejudice>, what's the sentence ***before***: <sentence2>?"
Forward search: "In <Pride and Prejudice>, what's the sentence ***after***: <sentence1>?"

	Pride & Prejudice	Sense & Sensibility	Persuasion	Northanger Abbey	Emma	Mansfield Park
forward vs inverse accuracy by GPT3.5	**0.5%** vs **14.4%**	**0.3%** vs **5.4%**	**0.07%** vs **4.3%**	**0.6%** vs **5.5%**	**0.8%** vs **7.2%**	**0.7%** vs **5.5%**
forward vs inverse accuracy by GPT4	**0.8%** vs **65.9%**	**0.9%** vs **40.2%**	**0.5%** vs **33.9%**	**0.9%** vs **41.0%**	**0.6%** vs **42.7%**	**0.3%** vs **31.7%**

Wiki Bio Task

Inverse search: "what's the full name of the celebrity born on <date> in <city> who is a <occupation>?" GPT3.5 acc = **23.9%** GPT4: **42%**
Forward search: "what's the birthday and year of <name> who is a <occupation> and was born in <city>?" GPT3.5 acc = **89.5%** GPT4: **99%**

Chinese Idiom Task

Given a common 4-letter Chinese idiom such as *指鹿为马*, mask out its i-th letter (for i=1,2,3, or 4) and let GPT fill out the missing letter.

Prompt 1: 成语"X鹿为马"的X是什么字? GPT3.5 accuracy **9.4%**, GPT4 accuracy **17.6%**
Prompt 2: 成语"指X为马"的X是什么字? GPT3.5 accuracy **29.5%**, GPT4 accuracy **36.1%**
Prompt 3: 成语"指鹿X马"的X是什么字? GPT3.5 accuracy **32.0%**, GPT4 accuracy **76.7%**
Prompt 4: 成语"指鹿为X"的X是什么字? GPT3.5 accuracy **56.7%**, GPT4 accuracy **90.6%**

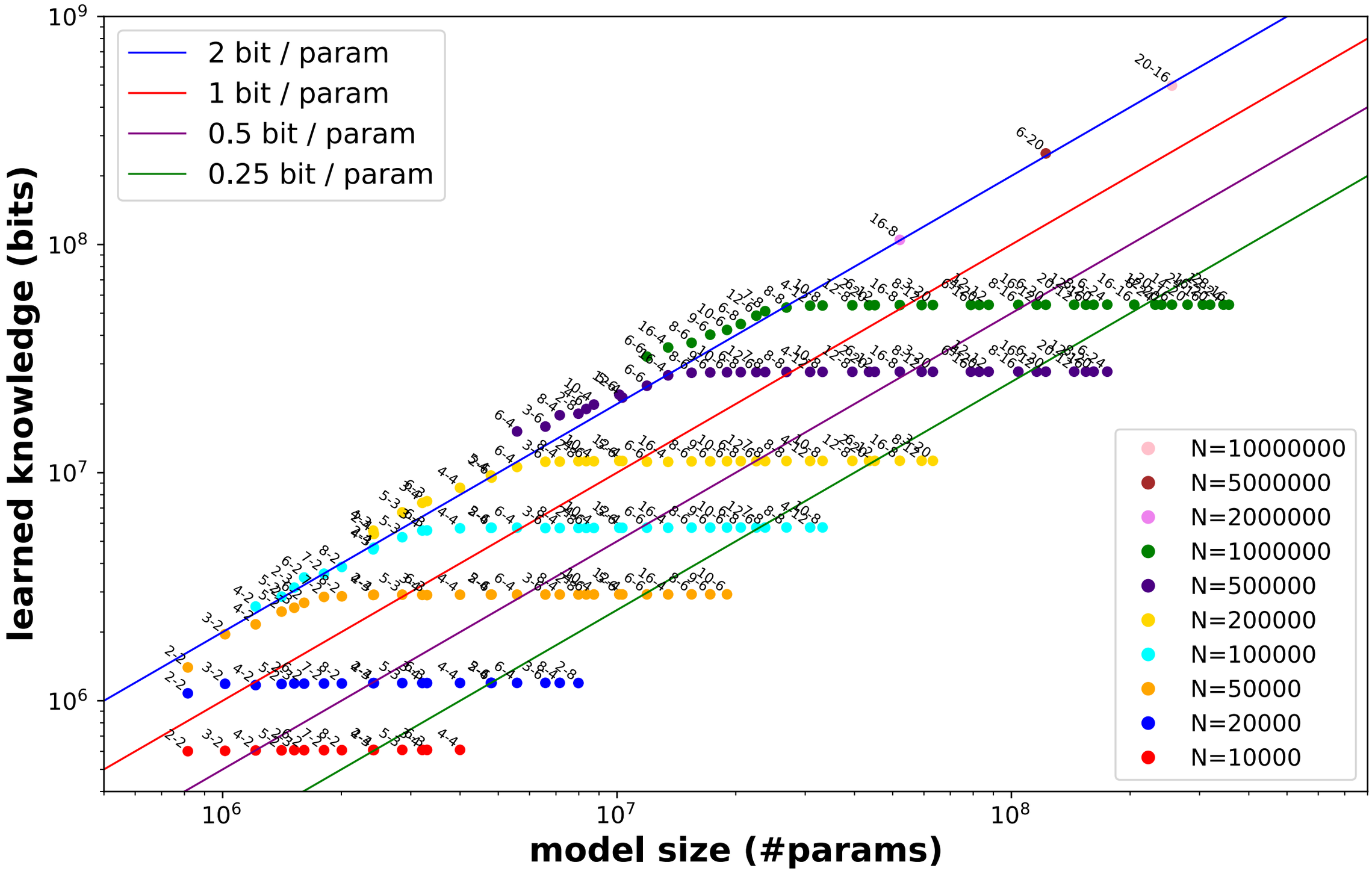
Chinese Poem Task

Given a famous two-sentence Chinese poem such as *劝君更尽一杯酒, 西出阳关无故人*, let GPT answer what's the sentence ***before***/***after*** <sentence**2**/**1**>

Inverse search: "西出阳关无故人"的上一句是什么? GPT3.5 accuracy **2.1%**, GPT4 accuracy **7.3%**
Forward search: "劝君更尽一杯酒"的下一句是什么? GPT3.5 accuracy **33.0%**, GPT4 accuracy **66.5%**

The Knowledge Capacity Scaling Law

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



Factors in Knowledge Storage

- More exposure in training (ideally 1000 times) helps
- Model shape matters less
- MoE matters less

- w/o MLP → hurts
- gated MLP → hurts
- int8 quantization is okay, but int4 halves capacity
- junk data hurts

Part 2:

Physiology of LMs

Diagnosing, Repairing, and Advancing LMs at Representation Level

- Length represention
- Word representation
- Context representation

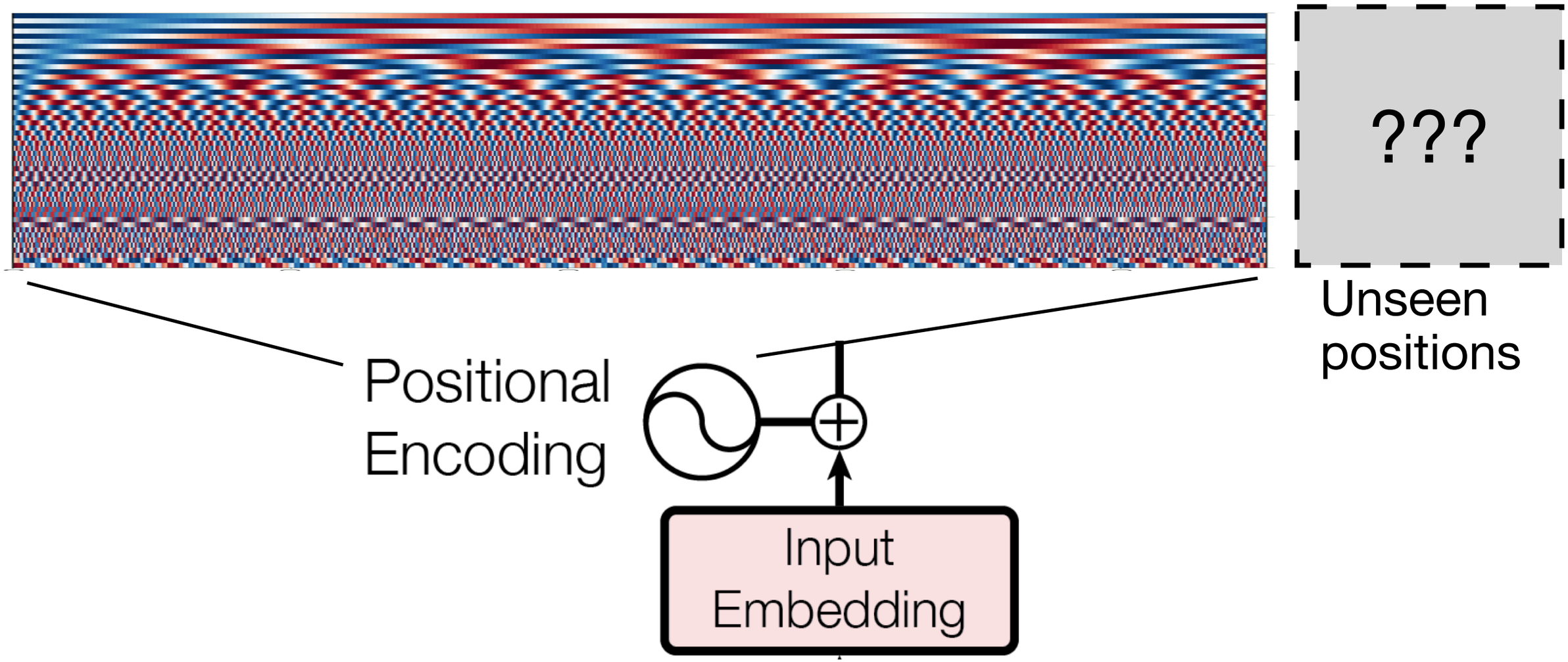
LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong,
Yu Chen, Heng Ji, Sinong Wang

Ph.D. student @ UIUC, intern @ Meta GenAI
https://arxiv.org/abs/2308.16137, **NAACL 2024 Outstanding Paper**

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

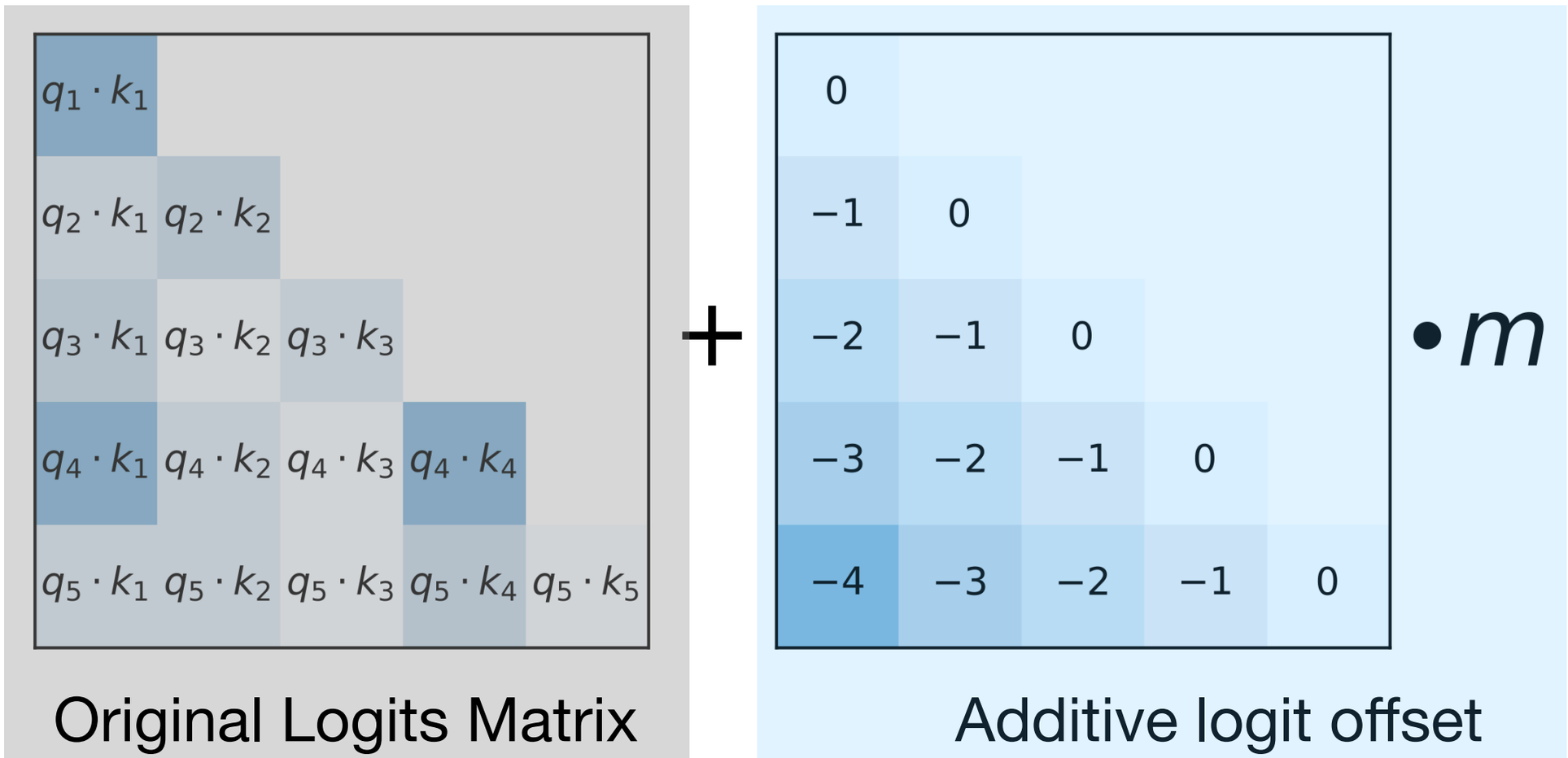
Relative Positional Encoding: ?

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

Core idea: determining attention based on distance

Alibi:

(Used in MPT-7B)



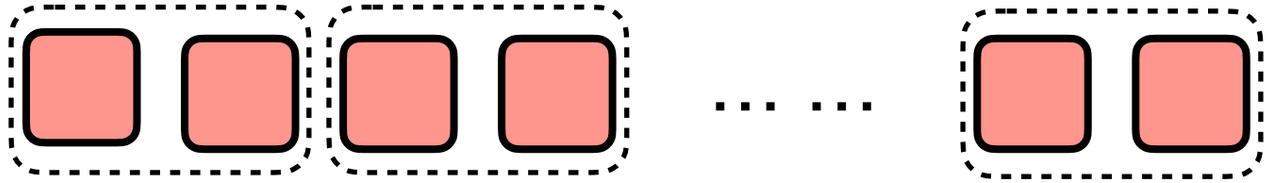
Press, Ofir, Noah A. Smith, and Mike Lewis. "Train short, test long: Attention with linear biases enables input length extrapolation." *arXiv preprint arXiv:2108.12409* (2021).

Relative Positional Encoding: ?

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

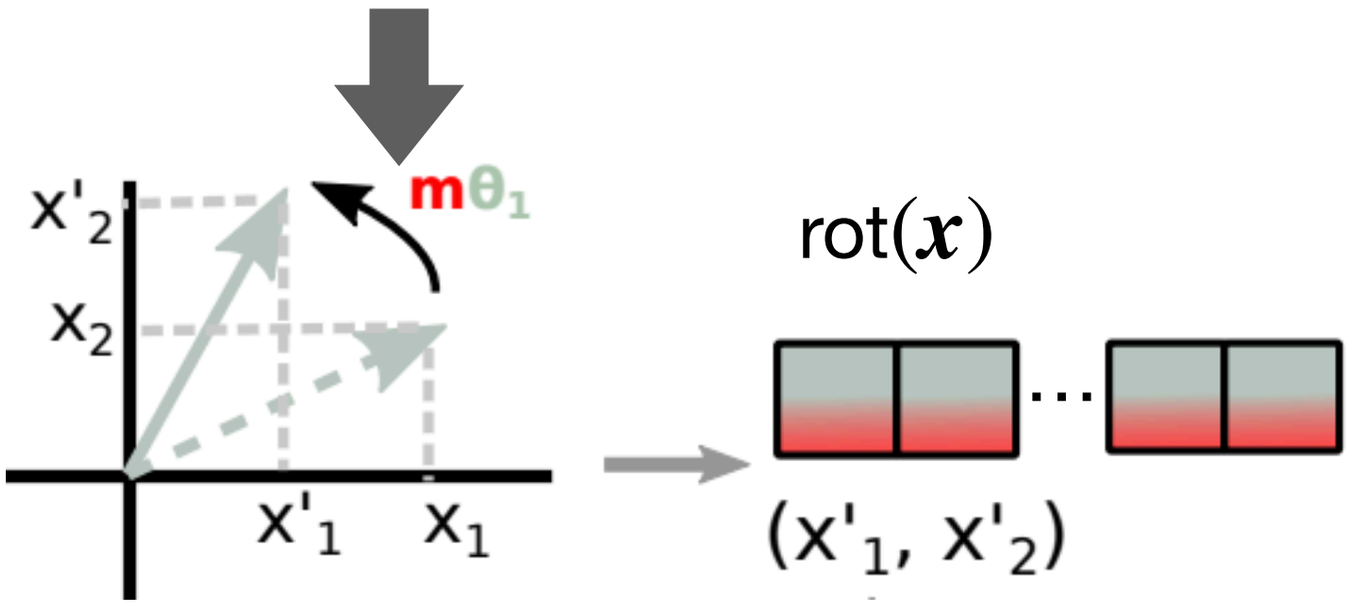
Core idea: determining attention based on distance

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



RoPE:

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



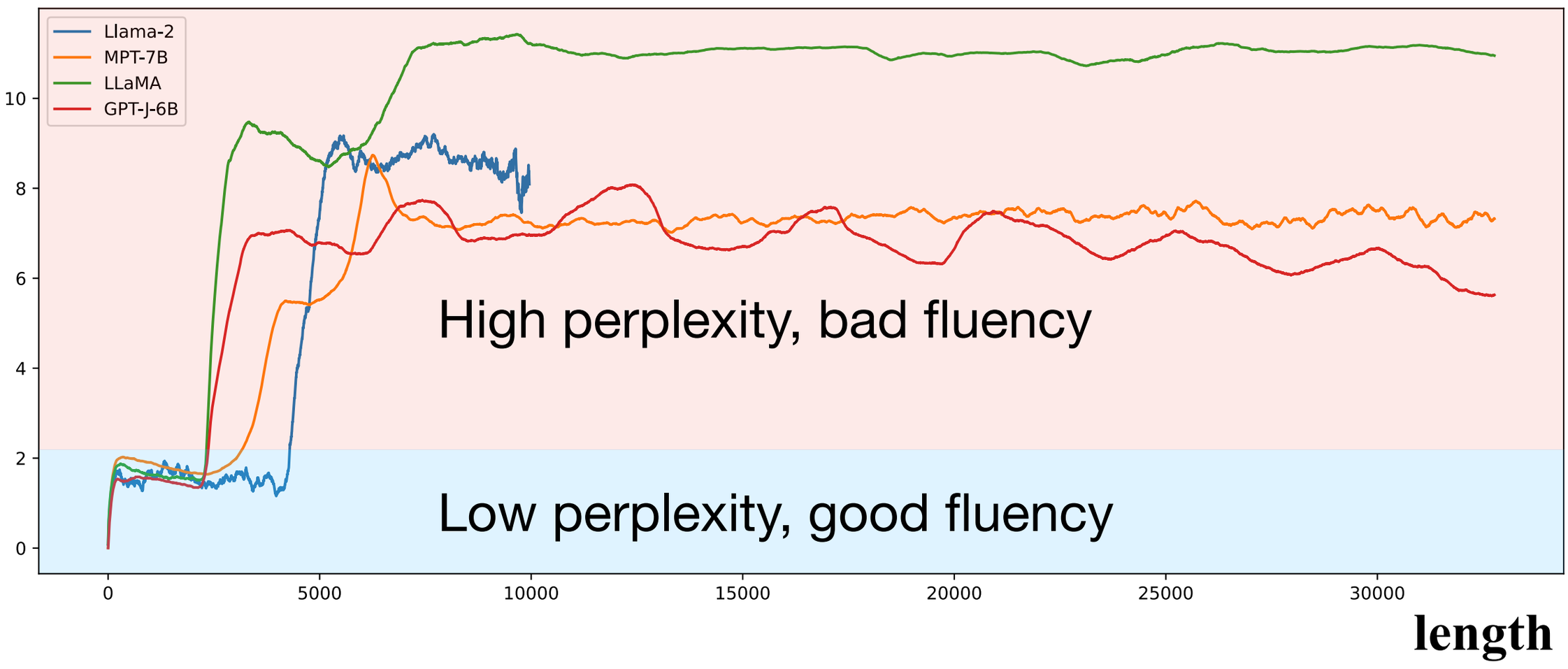
$l_{i,j} = \text{rot}(\boldsymbol{q}_i)^\top \text{rot}(\boldsymbol{k}_j)$ only depends on $i - j$, regardless of i or j.

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

Relative Positional Encoding: ?

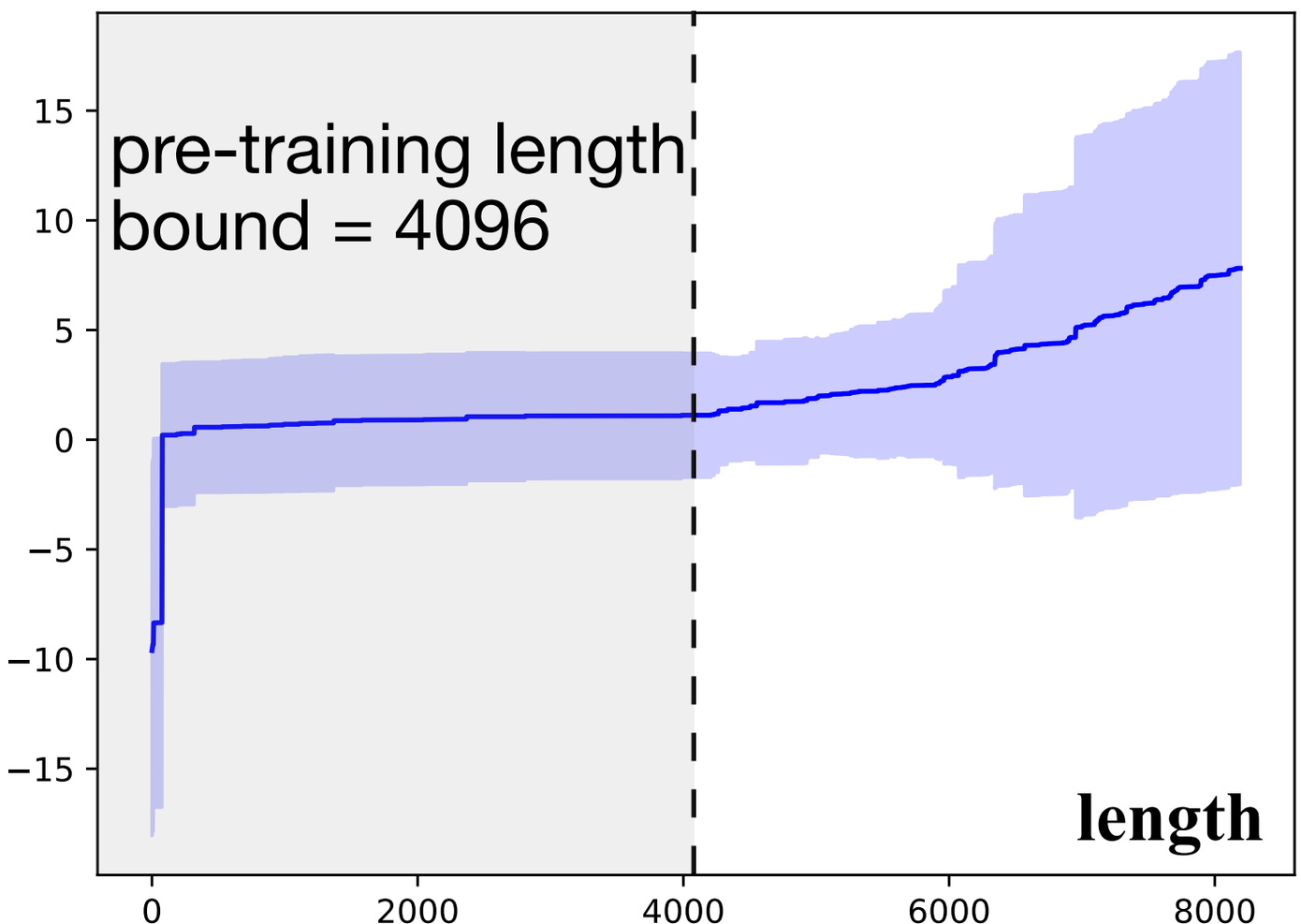
However, current LLMs still struggle on unseen lengths.

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



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.

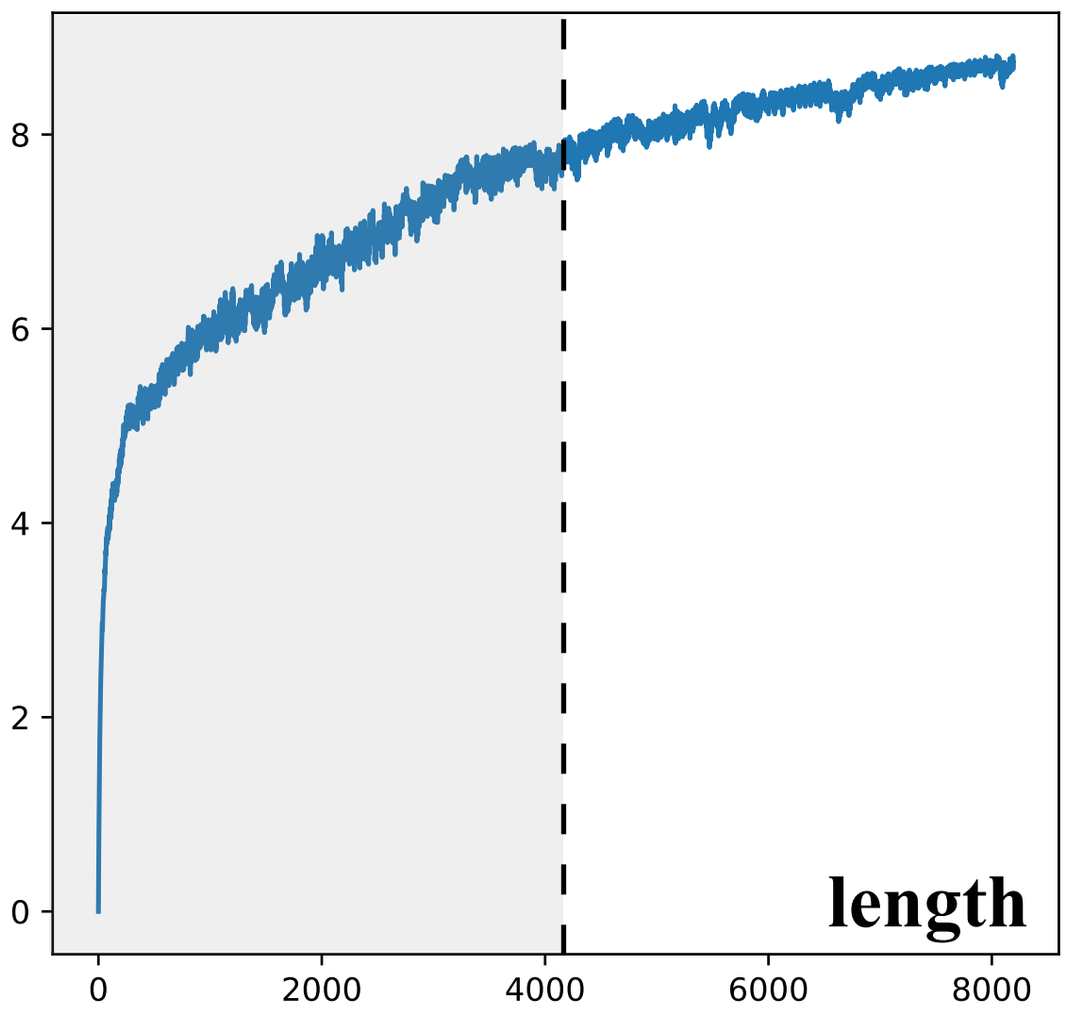
Factor 1: Unseen Distance

Takeaway: it may help to cap the relative distance values to the maximum that the model has seen during training (i.e., a distance ceiling)

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.



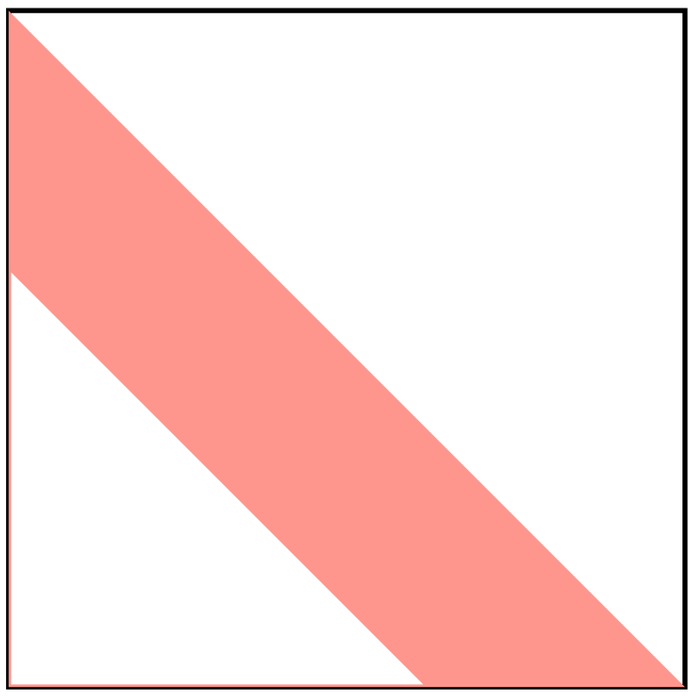
The entropy of attention distribution in Llama-2 continuously increases with length.

Factor 2: Too many tokens

Takeaway: we should upper-bound the attention context size, i.e., the number of tokens to be attended to.

You might think

What if we adopt a “sliding-window” attention mask, letting each token only to attend to the nearest tokens.



It doesn’t work.

There must be something else!

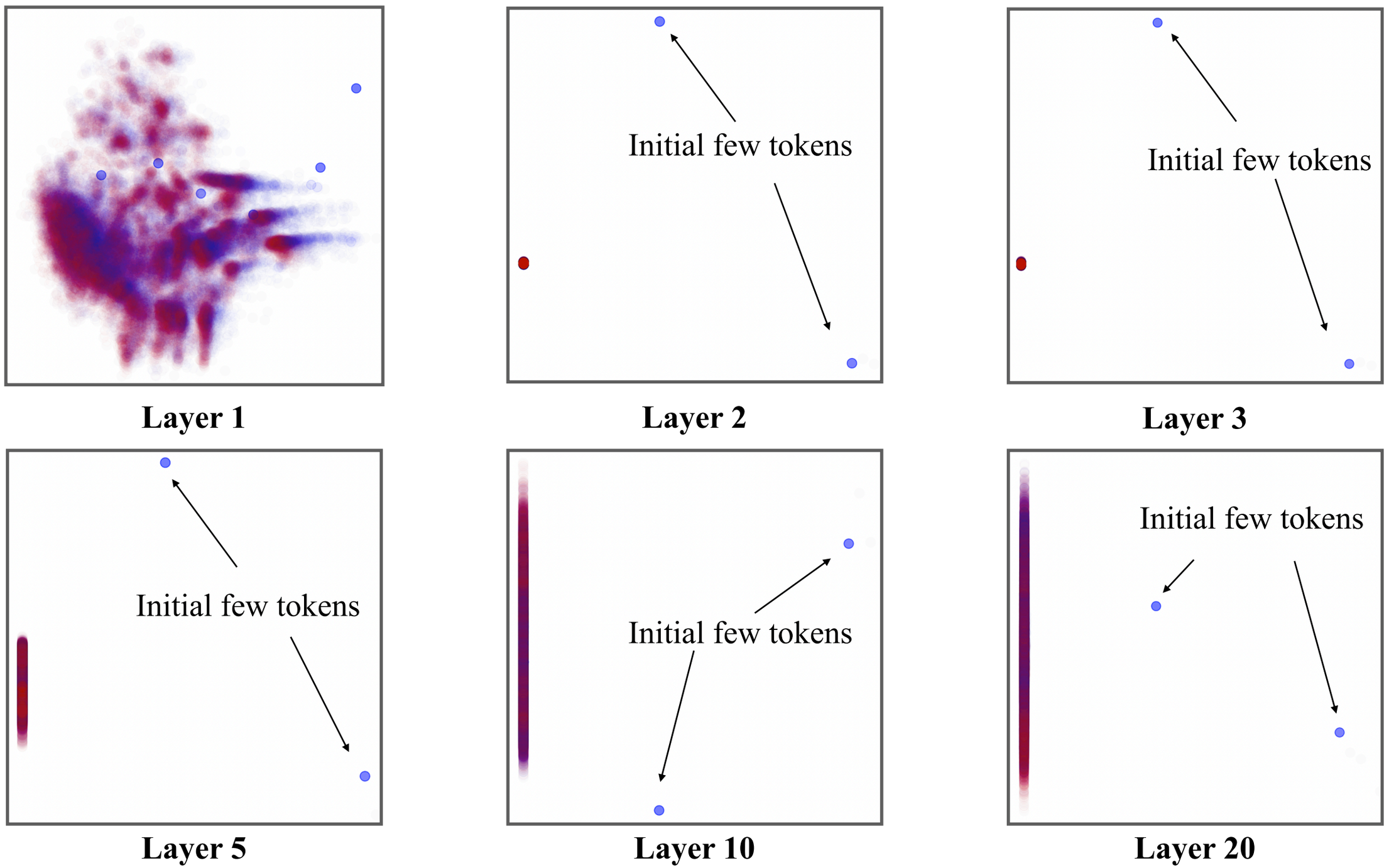
Factor 3: Implicitly Encoded Position

The first few tokens might implicitly encode absolute positions.

Theorem 3 (Informal): Even without explicit absolute positional embeddings, attention outputs of the first few tokens can occupy a distinct representational space compared to other positions. Therefore, when passed to later layers, these starting tokens have distinct value vectors coming from their lower layer outputs.

Kazemnejad, Amirhossein, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. "The Impact of Positional Encoding on Length Generalization in Transformers." *arXiv preprint arXiv:2305.19466* (2023).

Factor 3: Implicitly Encoded Position



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

Factor 3: Implicitly Encoded Position

Takeaway: we should keep the starting few tokens, because otherwise the self-attention, as a weighted sum over $\{\boldsymbol{v}_i\}$, will not be able to reach the region that the starting tokens occupy.

Solution: LM-Infinite

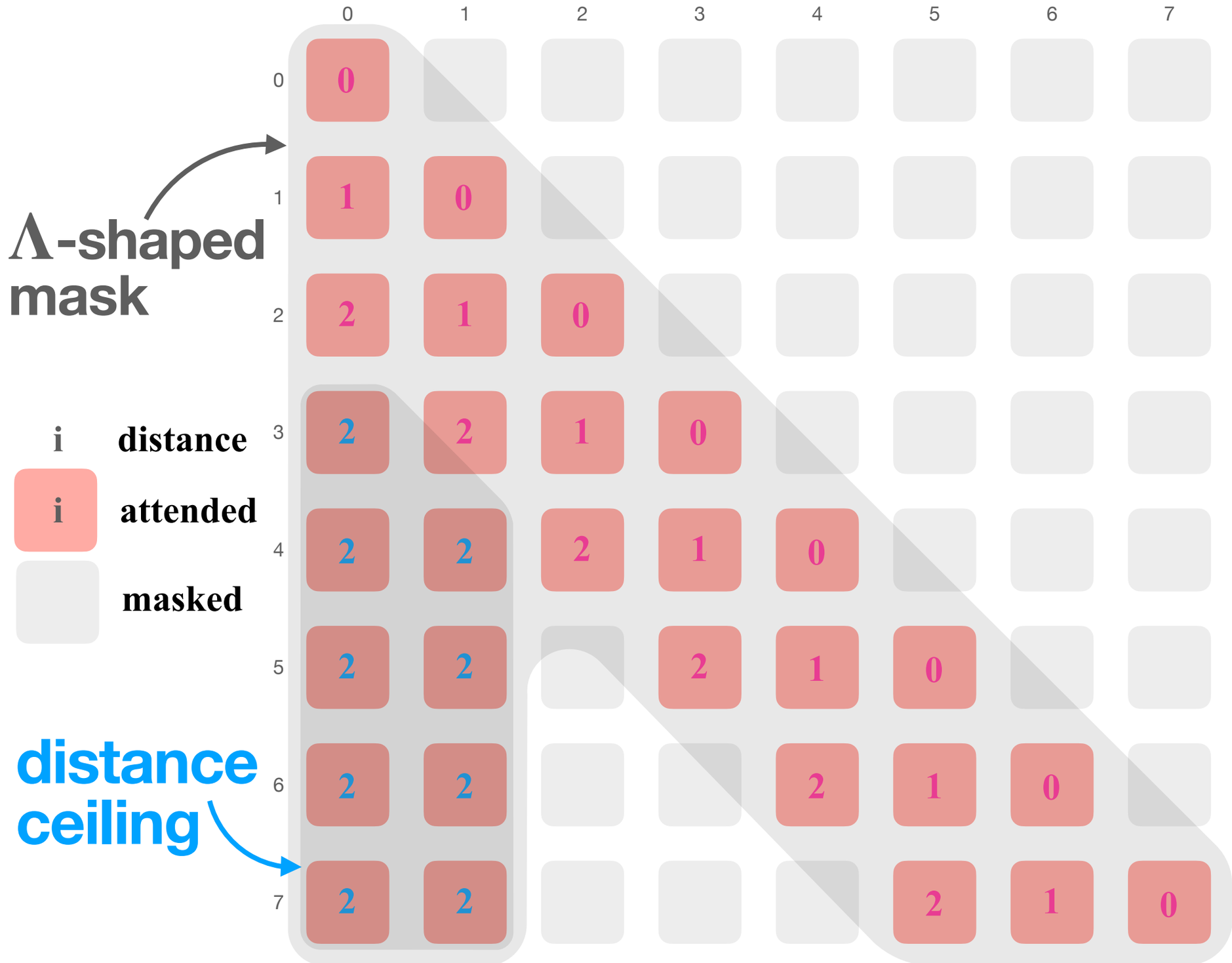
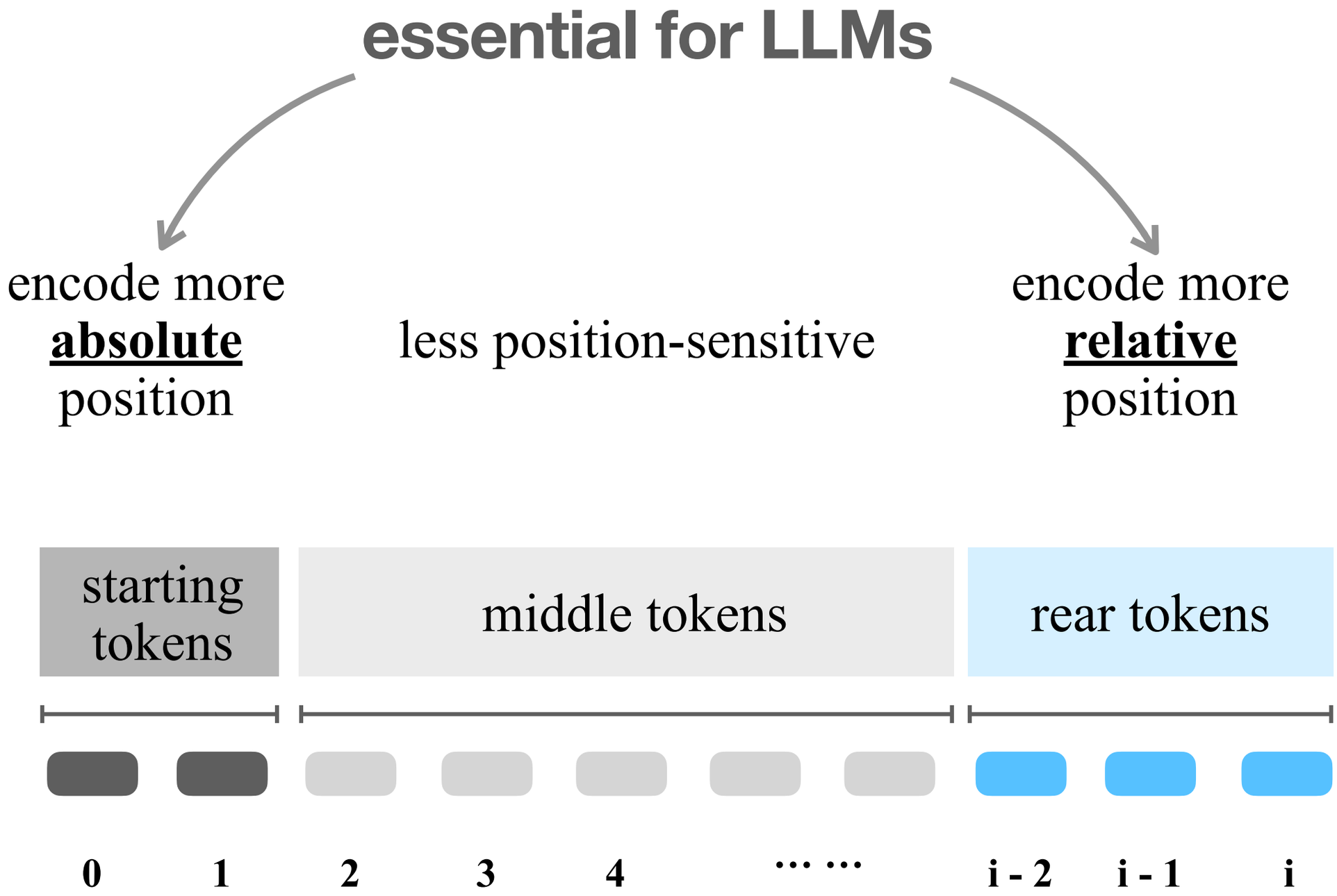
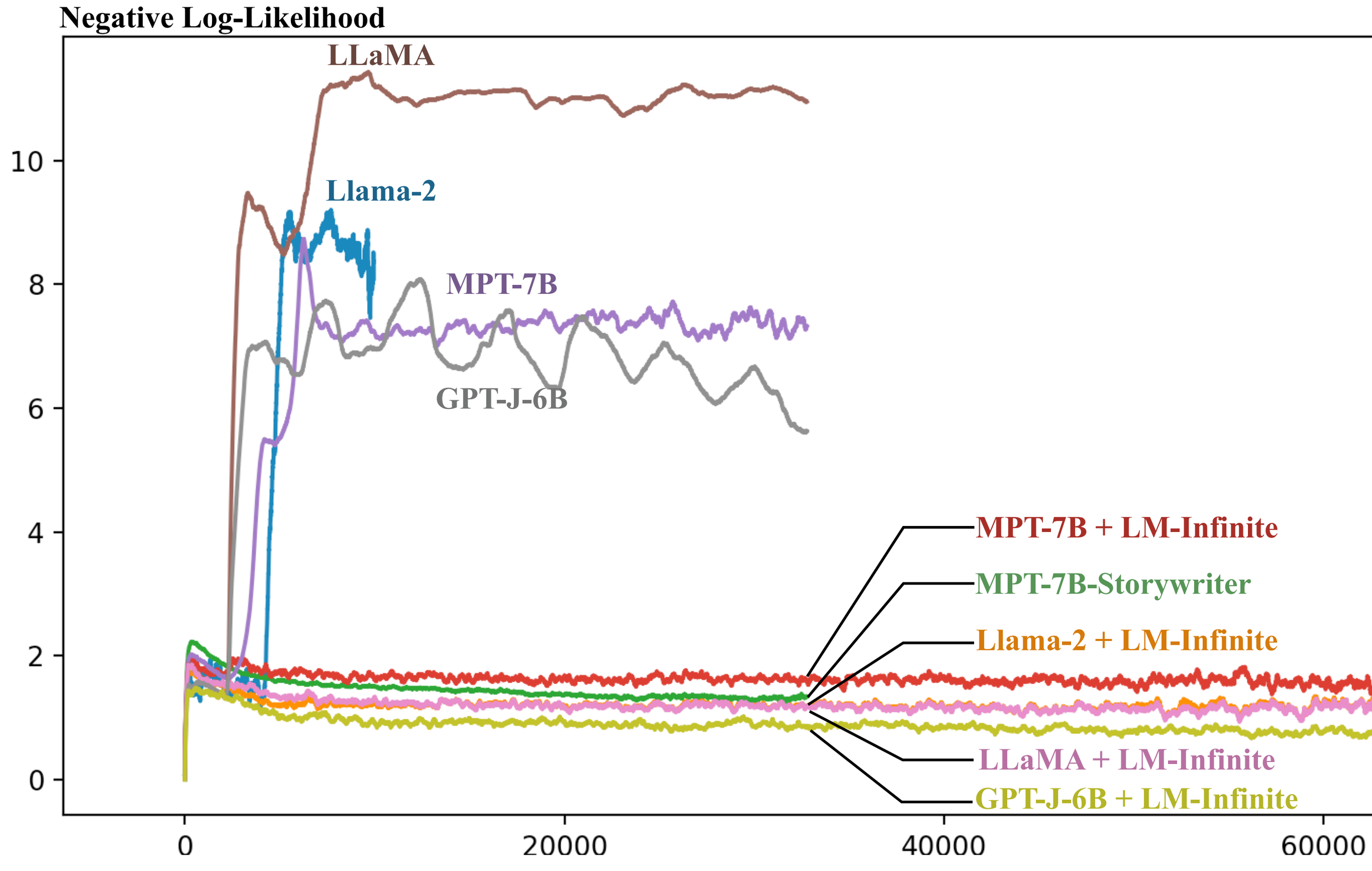


Illustration of a toy example with $L_{PT} = 2$

A Conceptual Model of Relative Position Encoding

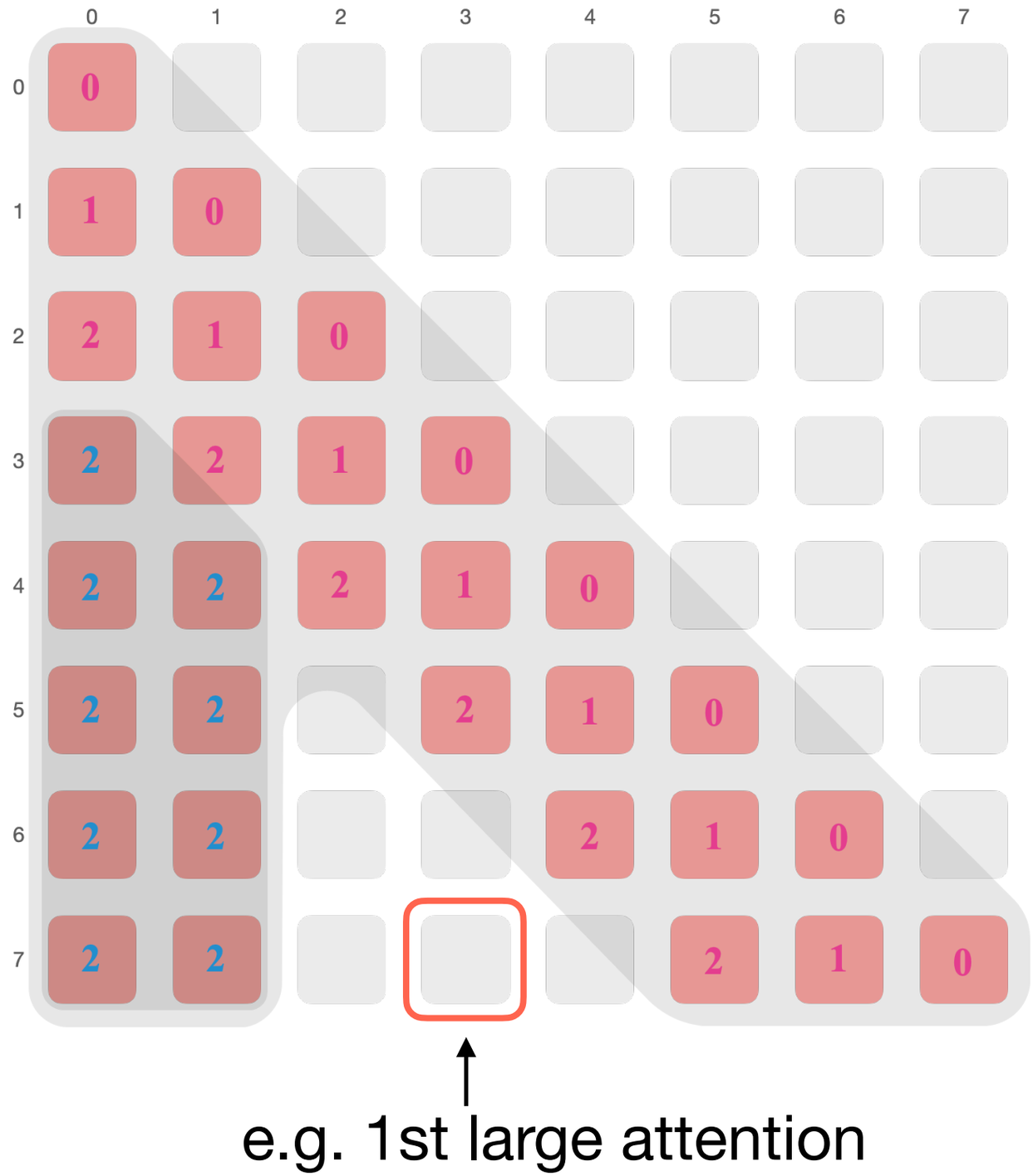


Fluency on Long Text



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.

2.2: Word Representation:

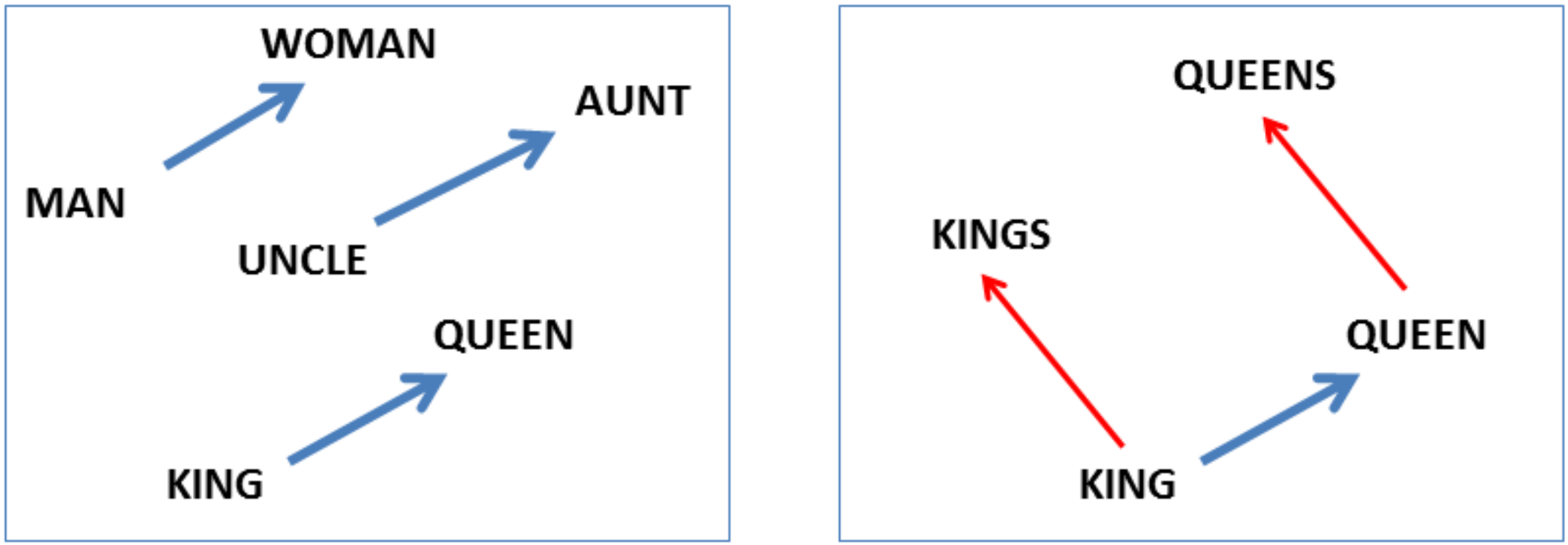
LM-Steer:
Word Embeddings Are Steers for Language Models

Chi Han[1], Jialiang Xu[2], Manling Li[2], Yi Fung[1], Chenkai Sun[1],
Nan Jiang[1], Tarek Abdelzaher[1], Heng Ji[1]
[1]UIUC, [2]Stanford

ACL 2024, Outstanding Paper Award, https://arxiv.org/abs/2305.12798
First Author: https://glaciohound.github.io, *chihan3@illinois.edu*
Code Repo: https://github.com/Glaciohound/LM-Steer

What Do Word Embeddings Embed?

Previous papers mostly focus on word-level interpretations

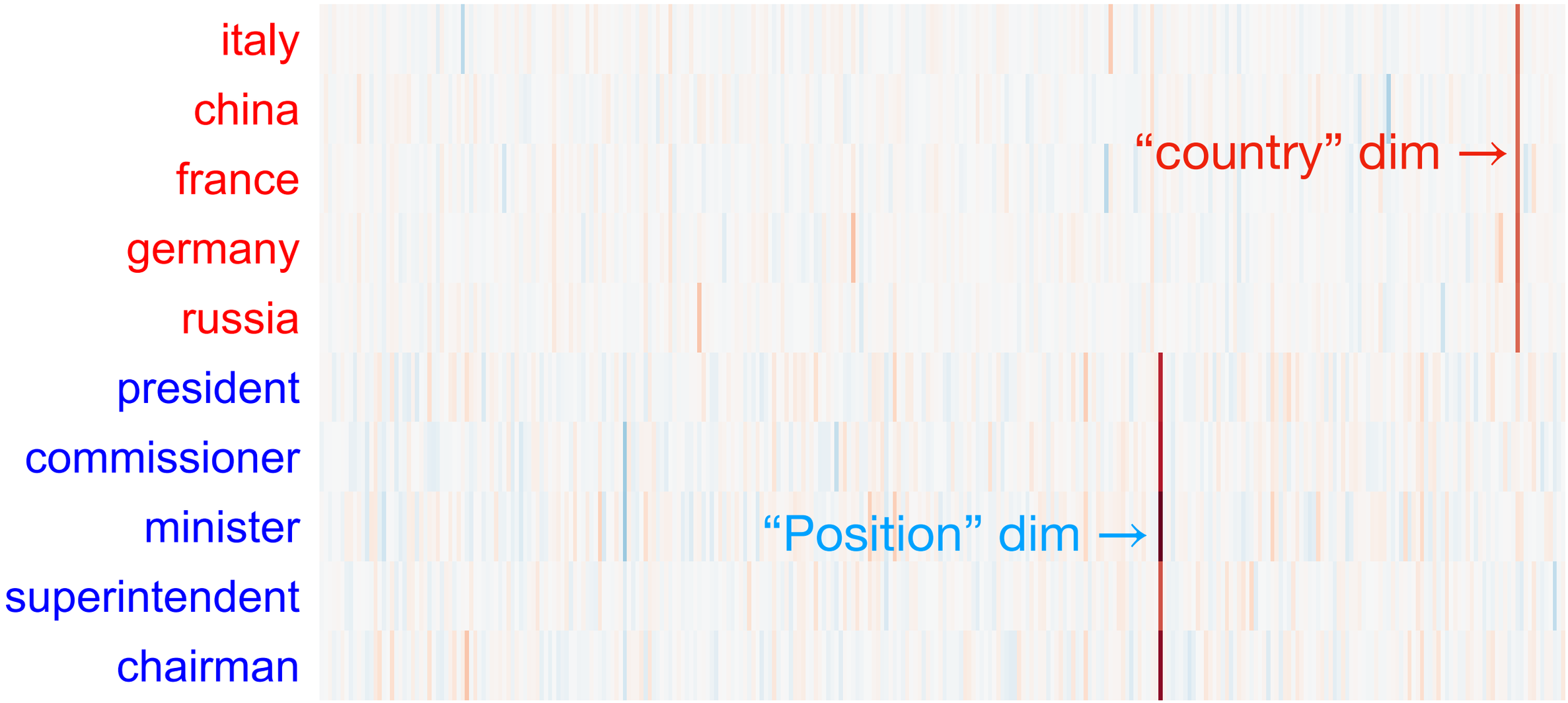


(a) Analogical Relations (metric space)

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

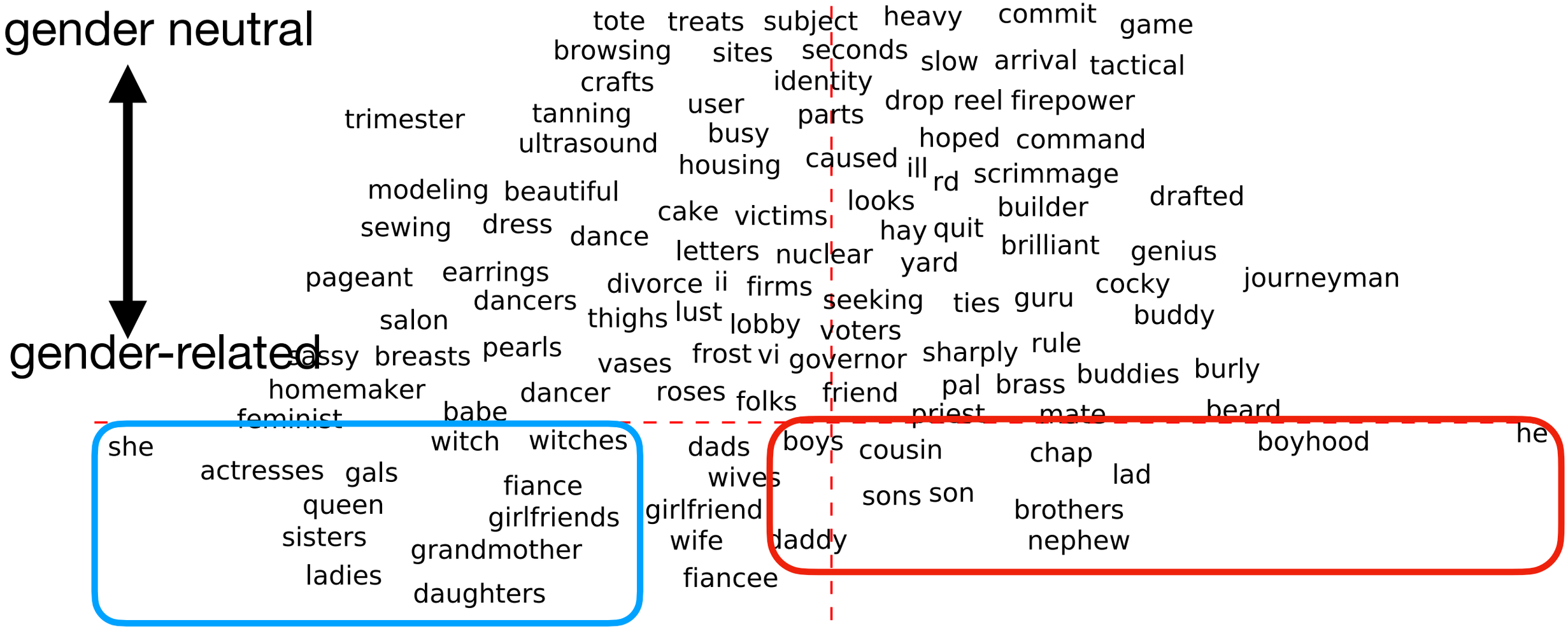


(b) Meaningful Dimensions (linear Space)

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.

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.

What Do Word Embeddings Embed?

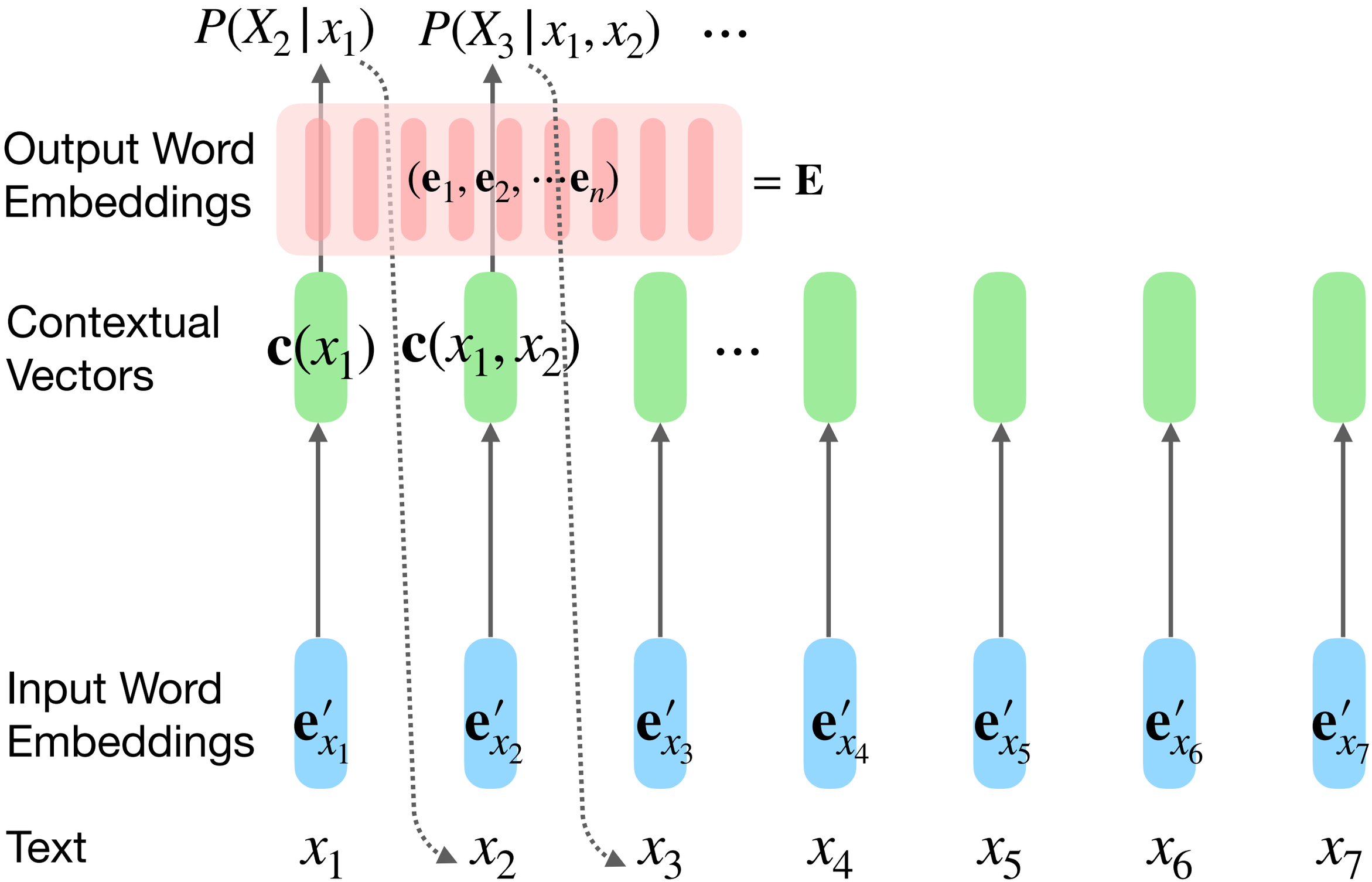
Previous papers mostly focus on word-level interpretations

$\mathbf{u}^{1}$	$\mathbf{u}^{4}$	$\mathbf{u}^{7}$	$\mathbf{u}^{8}$	$\mathbf{u}^{14}$	$\mathbf{u}^{121}$
lastly	molly	determinants	shyam	famille	jays
outset	sally	biochemical	sanjeev	vrier	strikeouts
ostensibly	toby	intrinsic	meera	autour	halladay
curiously	maggie	qualitative	anupama	naissance	hitters
actuality	valentine	elucidated	deepa	rique	buehrle
crucially	jenny	analytical	rajkumar	diteur	batters
theirs	tracy	psychological	manju	octobre	pitching
importantly	lucy	unger	uday	chambre	phillies
thankfully	carrie	ehrlich	chitra	lettre	rbis
regrettably	elliot	quantitative	vinod	campagne	astros
ironically	susie	integrative	archana	jeune	diamondbacks
aforementioned	laurie	extrinsic	bhanu	jours	homers
paradoxically	cooper	nagel	santosh	septembre	hitless
oftentimes	jill	methodologies	rajesh	enfance	orioles
doubtless	kitty	exogenous	ashok	plon	podsednik
unsurprisingly	charlie	underneath	munna	affaire	baserunners
connelly	shirley	translational	suman	cembre	hitter
merrick	hannah	kuhn	komal	royaume	sox
invariably	annie	functional	subhash	propos	pettitte
dunning	elaine	schweitzer	usha	juin	vizquel
Transition	First Names	Science	Indian Names	French	Baseball

(b) Meaningful Dimensions (linear Space)

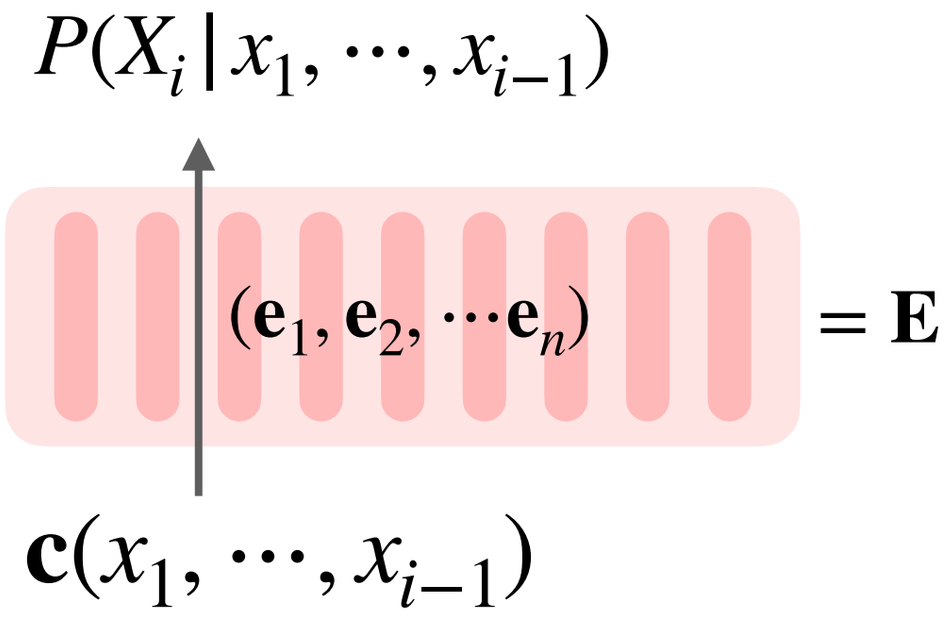
Shin, J., Madotto, A., & Fung, P. (2018). Interpreting word embeddings with eigenvector analysis. In 32nd Conference on Neural Information Processing Systems (NIPS 2018), IRASL workshop (pp. 73-81).

Word Embeddings in Causal LMs



Output Word Embeddings

Projecting to Logits



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

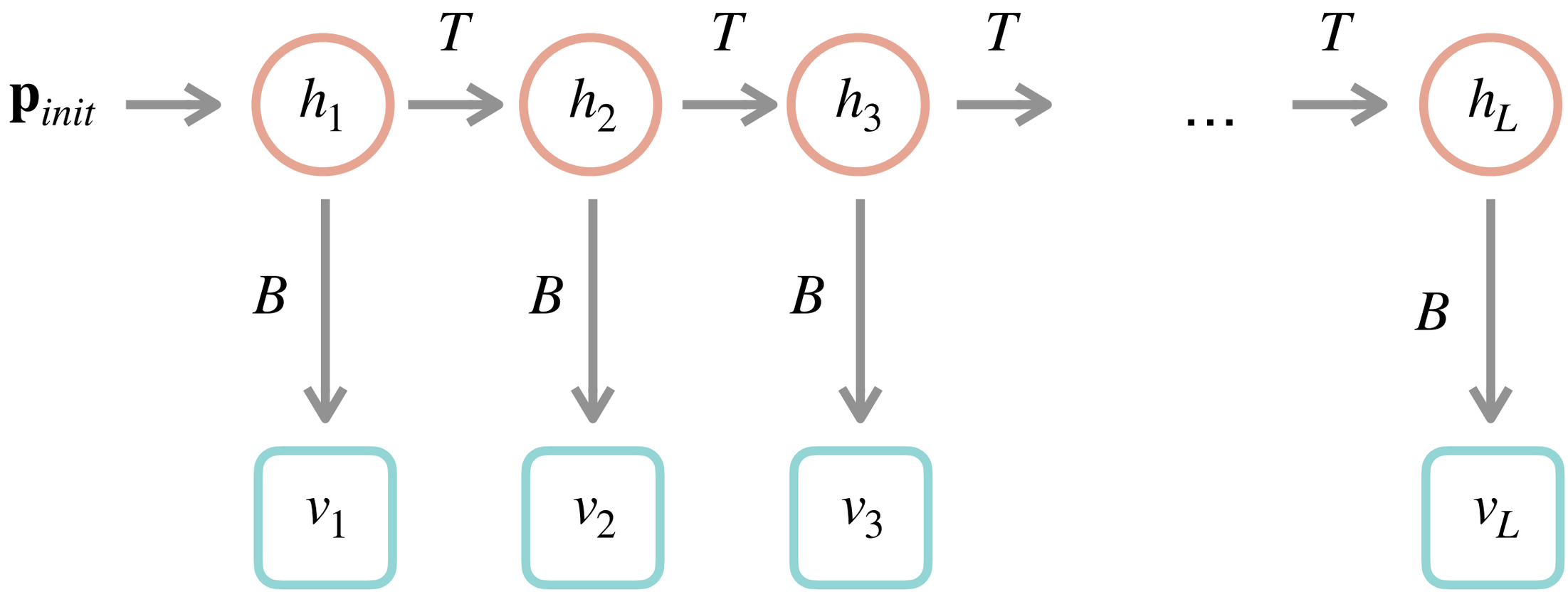
Output Word Embeddings

A Dimension Reduction

$$\mathbf{E} : [1..n] \rightarrow \mathbb{R}^d$$

- when $k = |\mathscr{V}|$ can theoretically express any distribution
- when $k < |\mathscr{V}|$, compresses (embeds) words so they are inter-related
    - but, in what way?

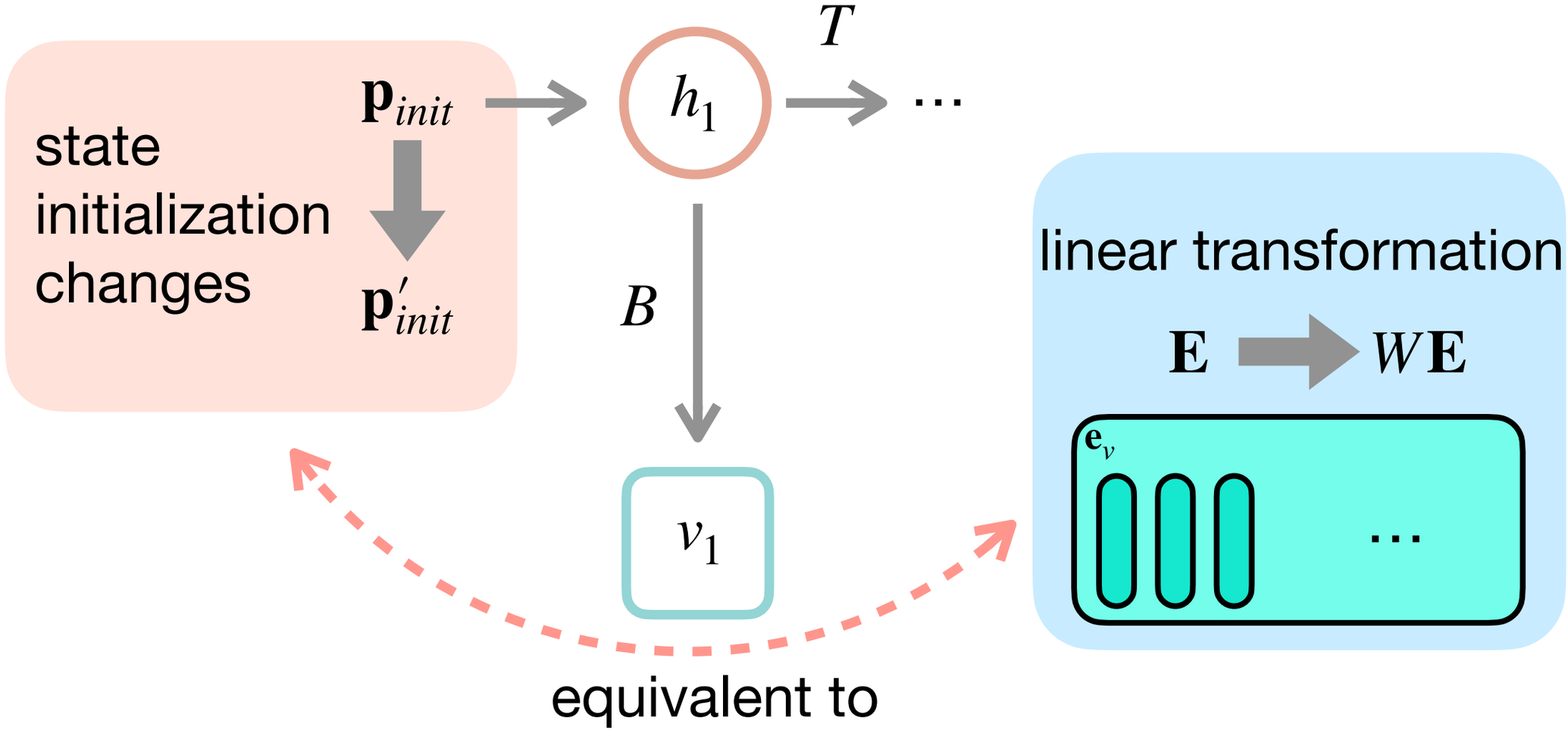
HMM as A Theoretical Framework



$$P_{HMM}(v_1, \cdots, v_L; \mathbf{p}_{init}) = \mathbf{p}_{init}^{\top} T \left(\prod_{i=1}^{L-1} diag(\mathbf{p}(v_i)) T \right) \mathbf{p}(v_L)$$

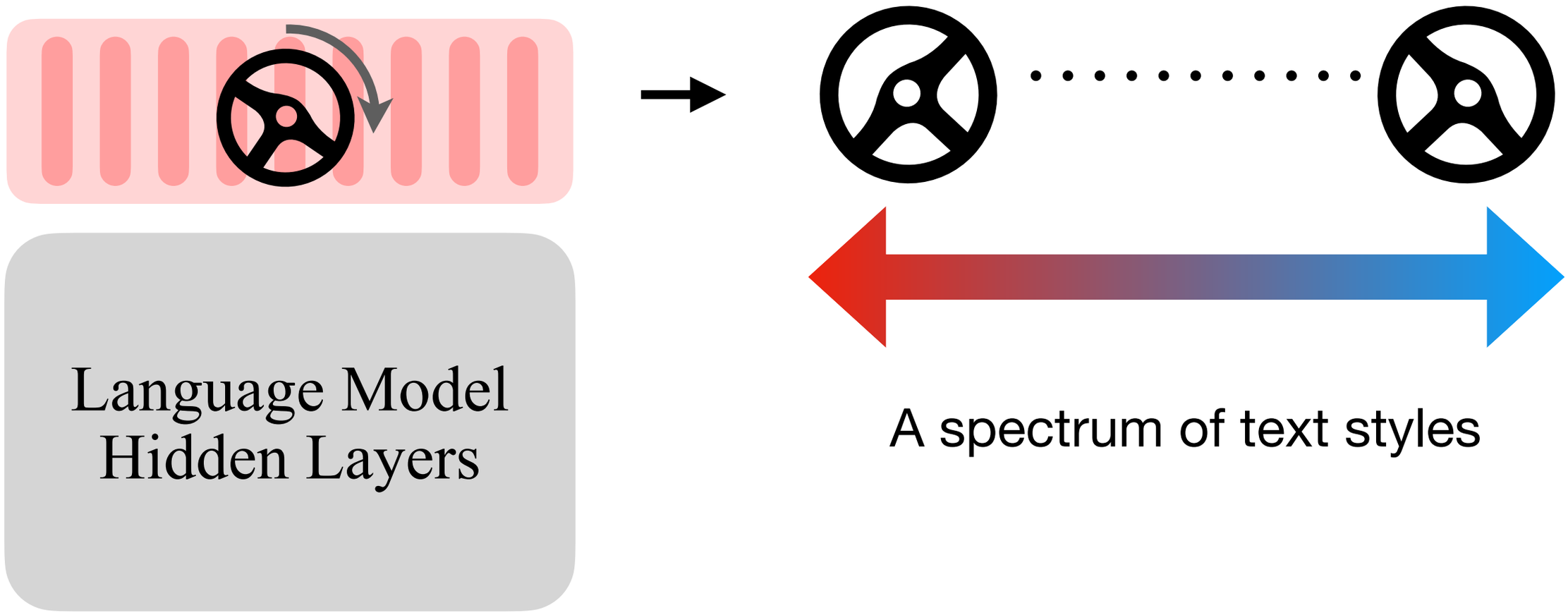
Sequence Shift ≈ Word Embedding Transform

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



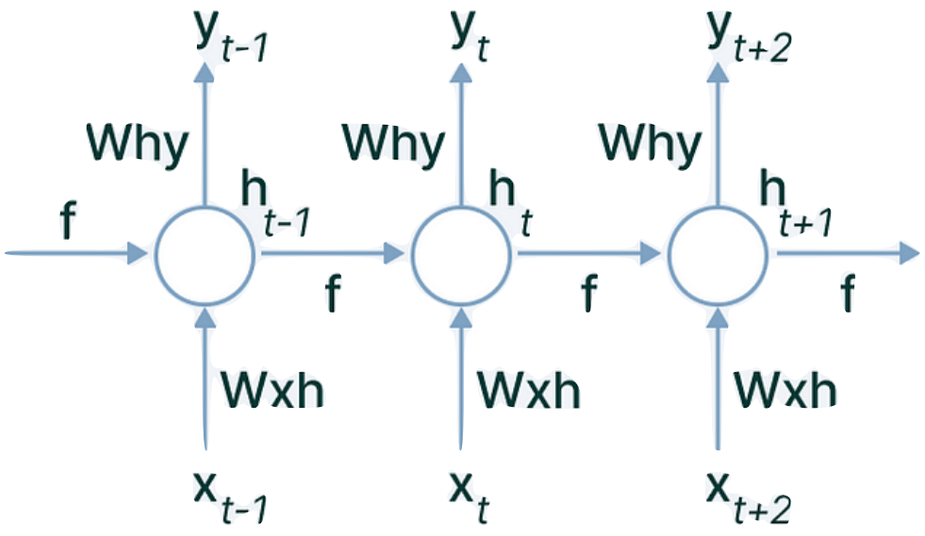
Word Embeddings Are Steers

An Intuitive Explanation

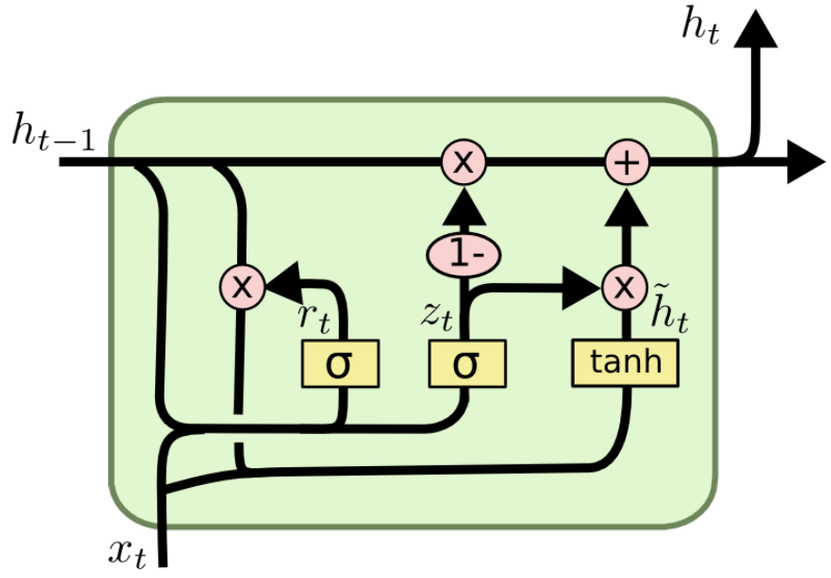


- Non-trivial claim as it connects word distributions and sequence distributions

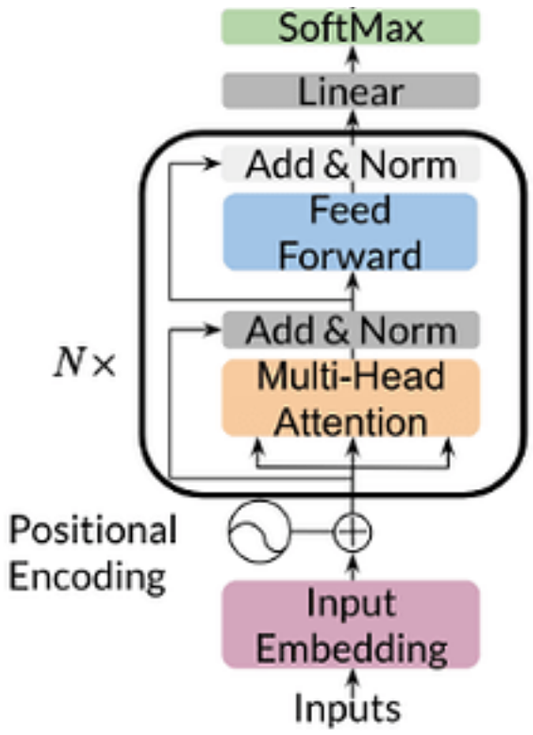
Theoretical Generality



RNNs



LSTMs



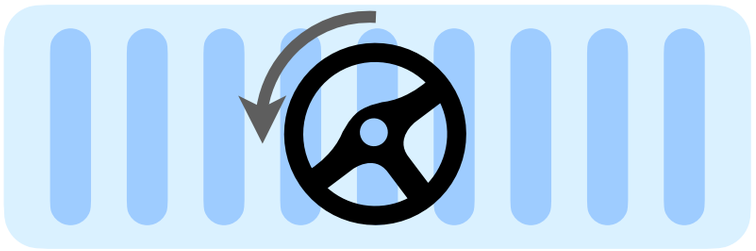
Transformers

https://miro.medium.com/v2/resize:fit:1200/1*5NhjY5OH8HKpi5oHuEMxTg.png
https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/LSTM3-var-GRU.png

LM-Steer

steering on output word embeddings

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

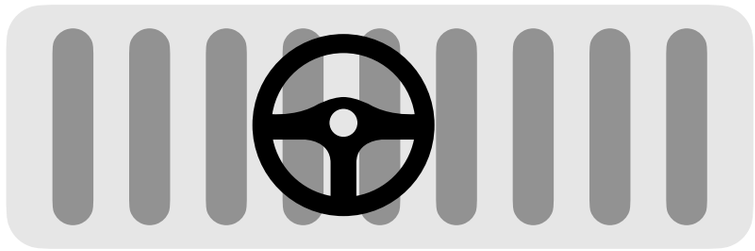


Language Model Hidden Layers

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

"My life is <u>boring</u>"

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

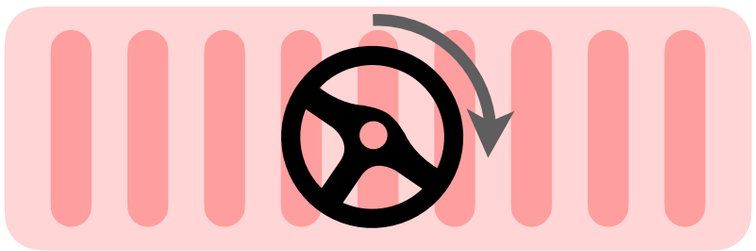


Language Model Hidden Layers

Original LM P_0

"My life is <u>okay</u>"

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



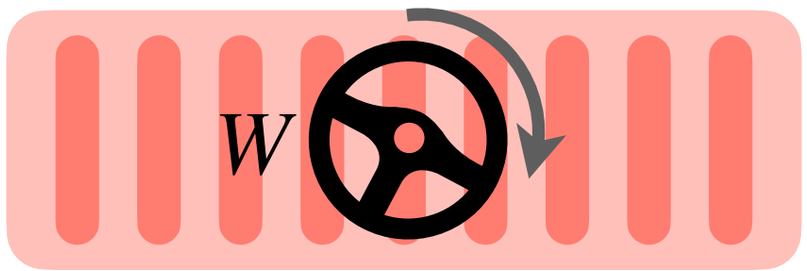
Language Model Hidden Layers

Positively steered LM $P_{\epsilon W}$

"My life is <u>brilliant</u>"

LM-Steer Broken Down

Output word embedding $\boldsymbol{E}$



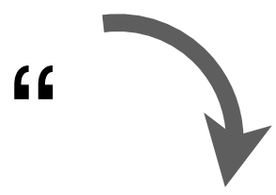
Language Model Hidden Layers

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

for each word:

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

The steering scale

“  ”

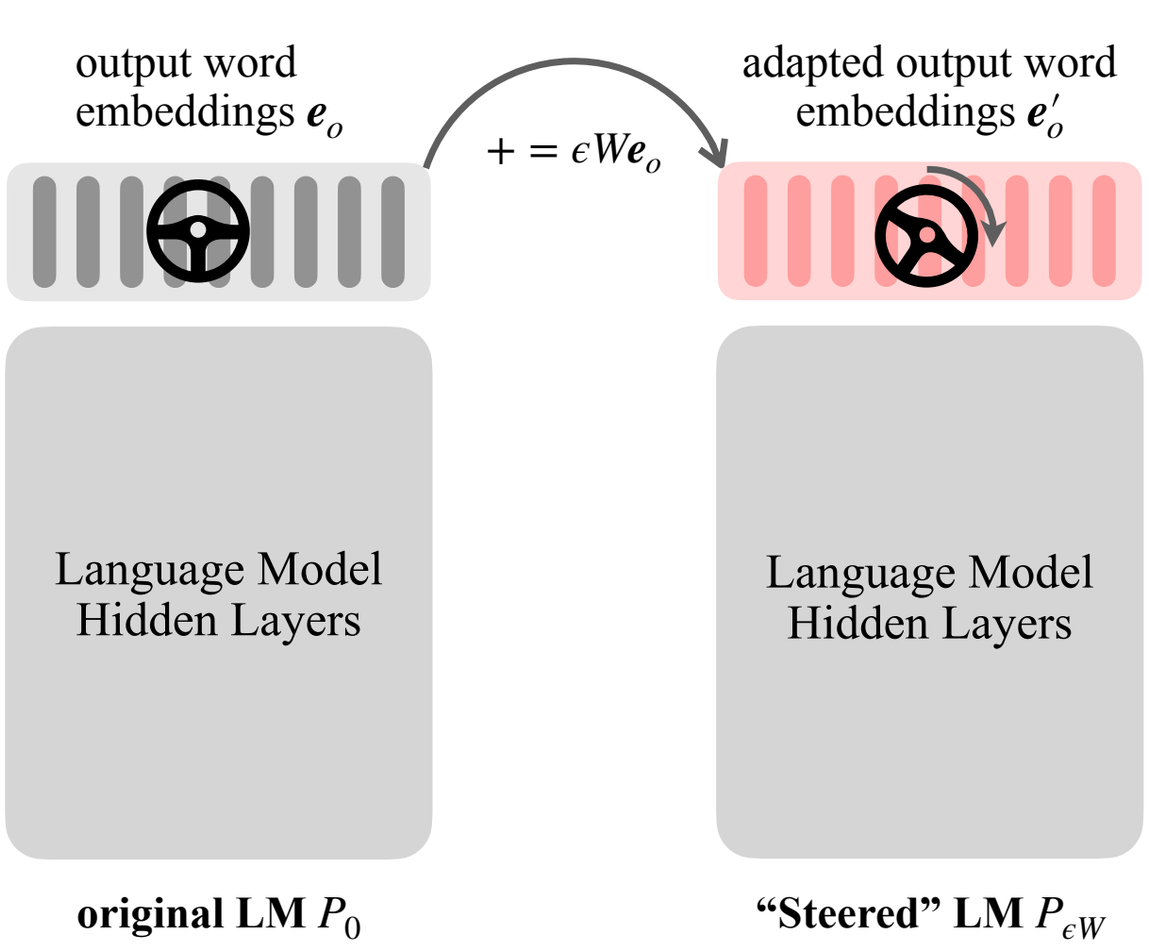
ϵ

the steering matrix

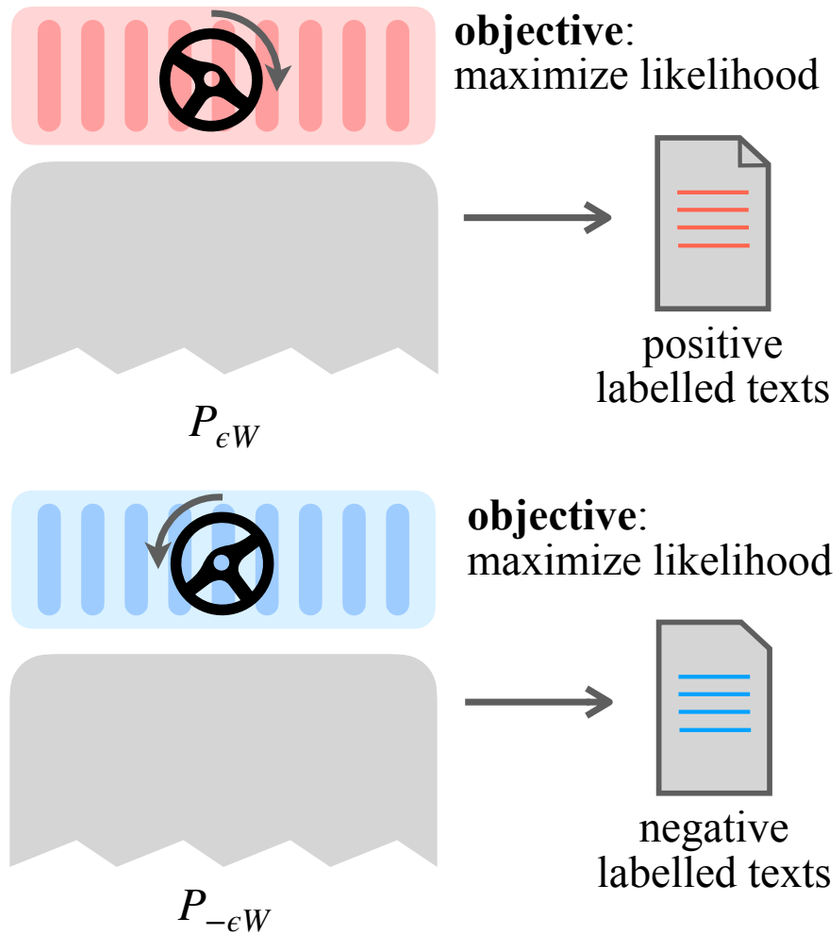
“  ”

W

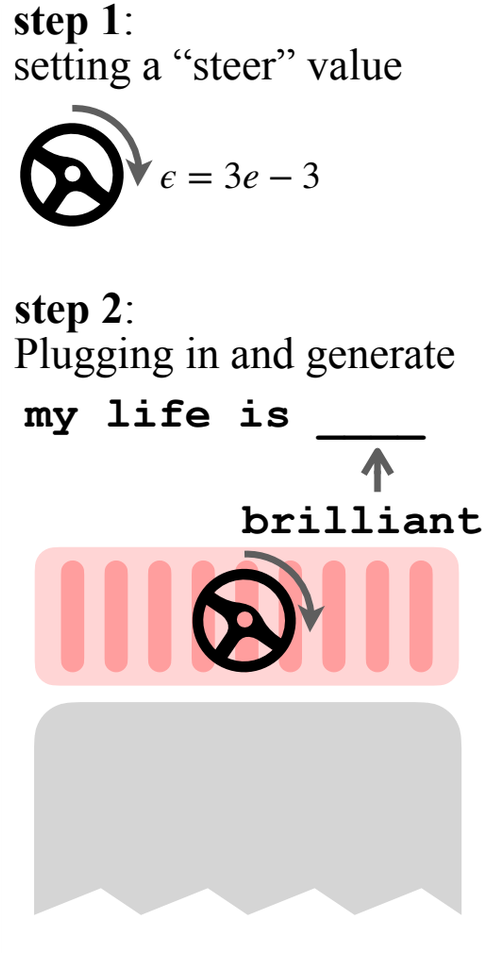
Training & Inference



(a) LM-Steer overview



(b) Training

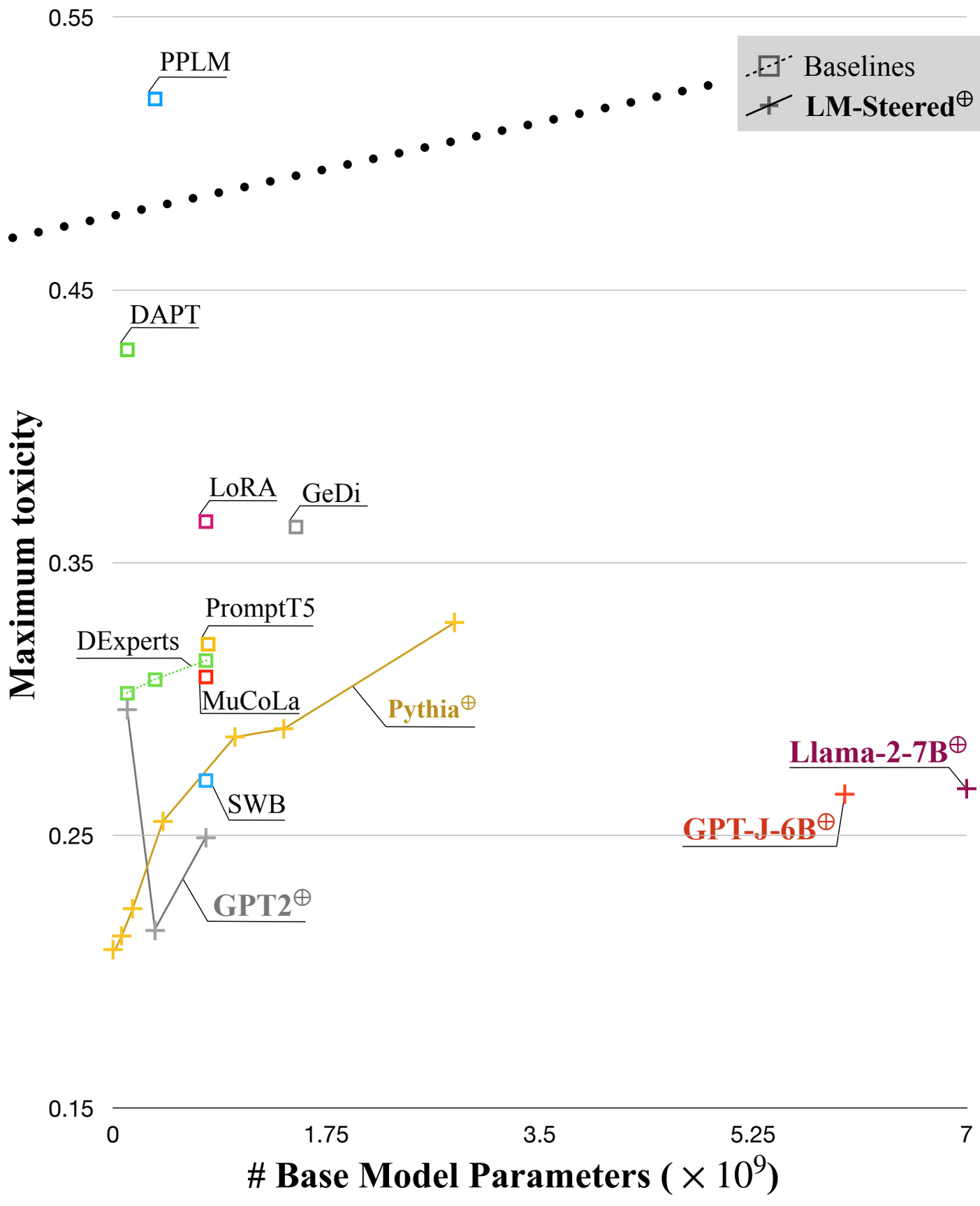


(c) Generation

Detoxification

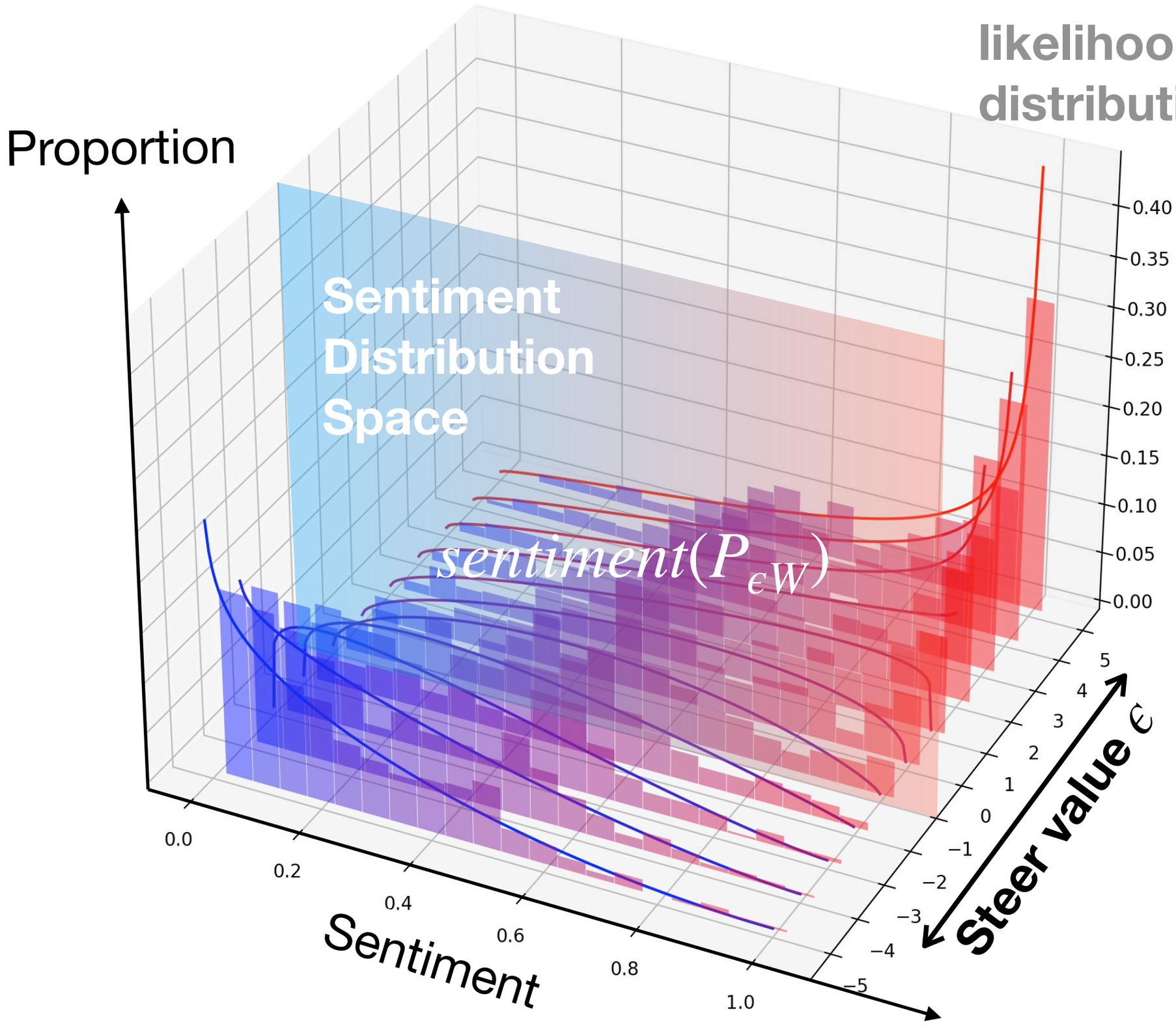
Holistic Comparison

- Across base model sizes, LM-Steered *GPT2 family, Pythia family, GPT-J* and *Llama-2-7B* models (+) consistently outperform other baselines (▫) on detoxification.



Continuous Steering

curves: maximal likelihood beta-distribution



Continuous Steering

Steer	Generation	word toxicity level	# toxic phrases
-5e-3	What **moron** said that **stupid** comment.	“moron”, “stupid”	2
-3e-3	What’s **stupid** is **stupid**, right?	“stupid”	2
-1e-3	What’s this? You think that your religion, your culture, your country are **not good enough**?	“not good enough”	1
0	What’s more, it makes for a fun, cheap, and efficient way to improve the performance of your car engine and to make your driving that much safer.	—	0
1e-3	What’s more, it makes for a fun, cheap, and efficient way to improve the performance of your car engine and motor.	—	0
3e-3	What’s on your mind? What’s on your mind?	—	0
5e-3	What’s on Netflix? If you can’t figure out what’s being watched on Netflix, you need to figure out what are people watching!	—	0

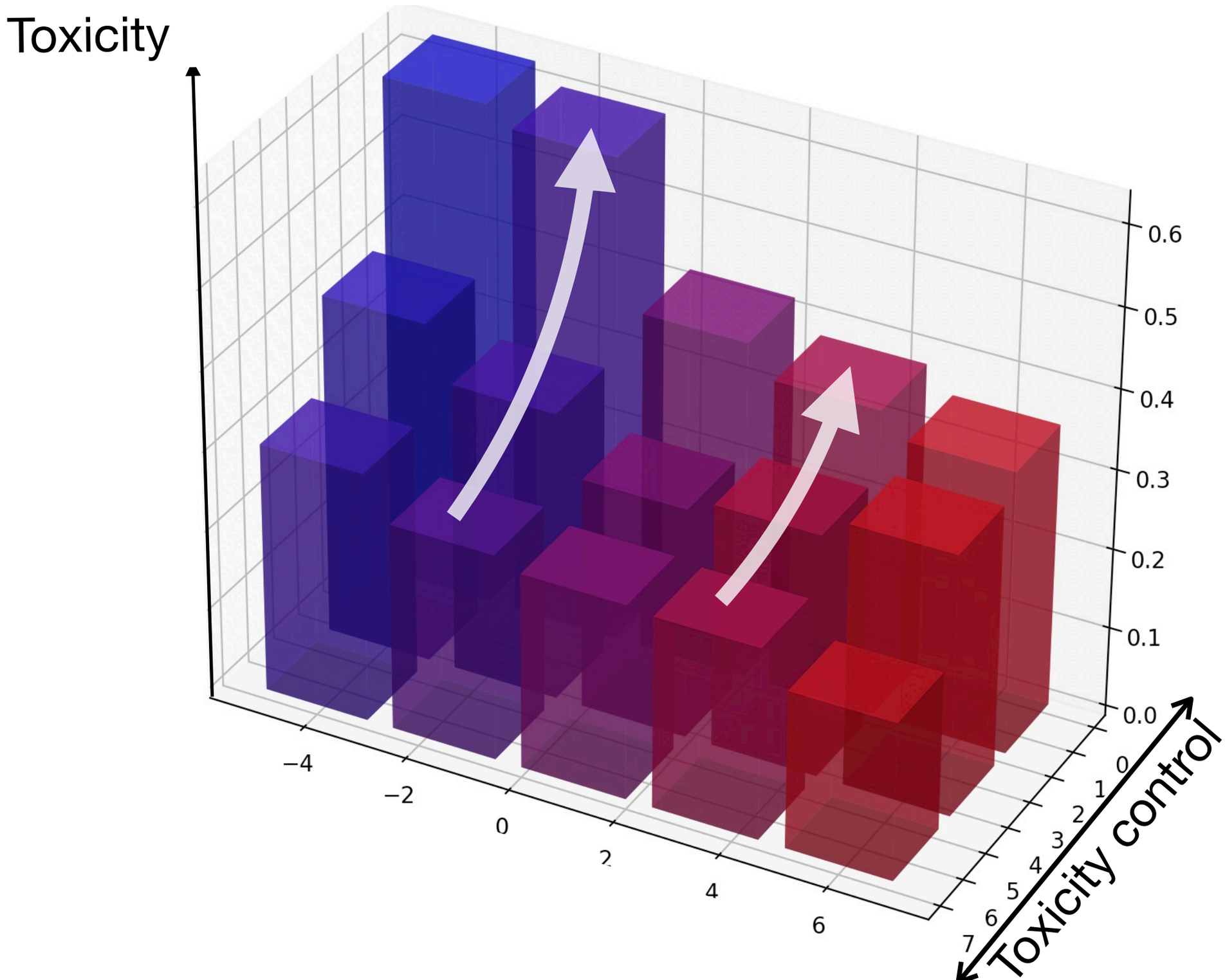
Compositional Steering

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

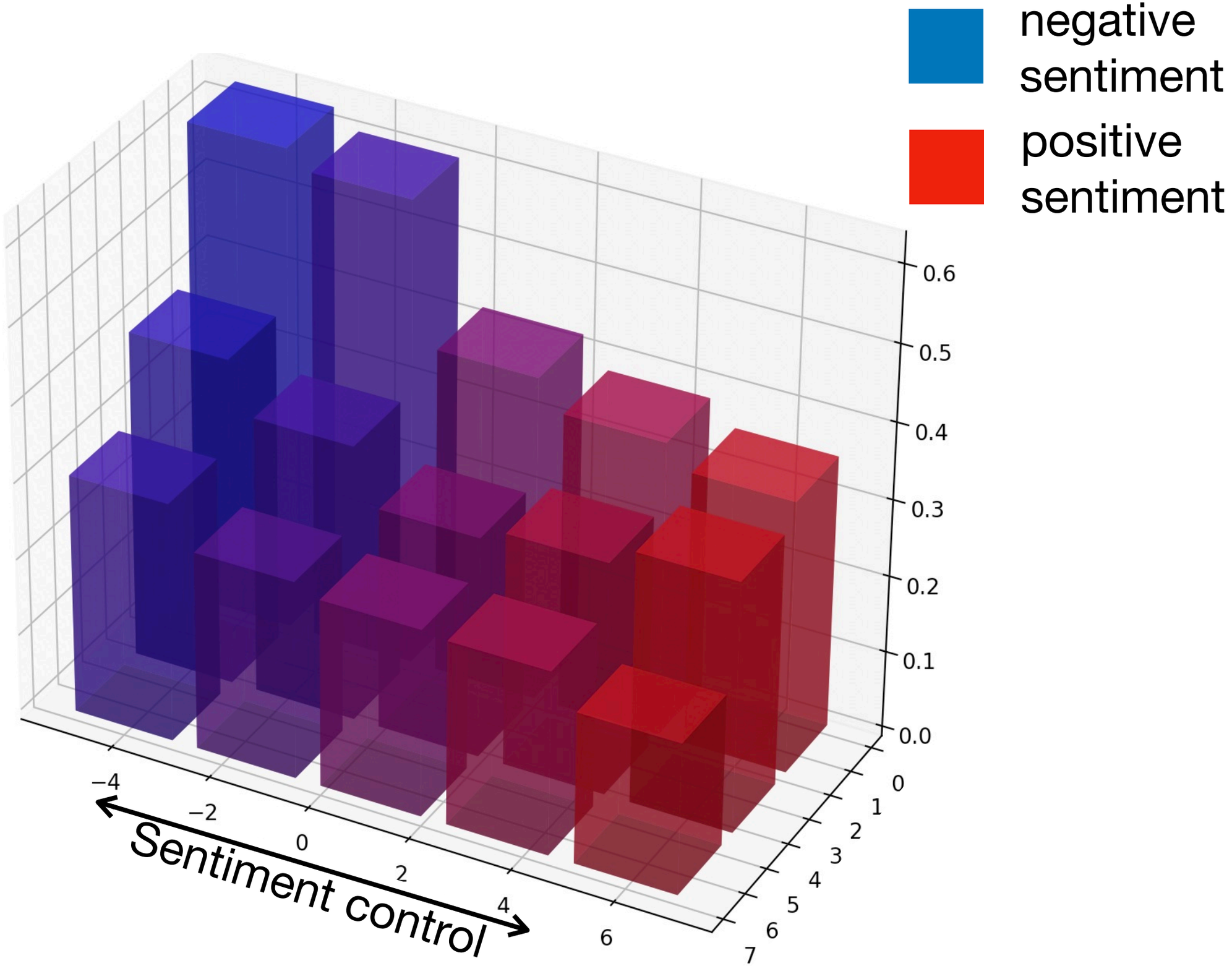
LM-Steer 2: $P_{\epsilon_2 W_2}$

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

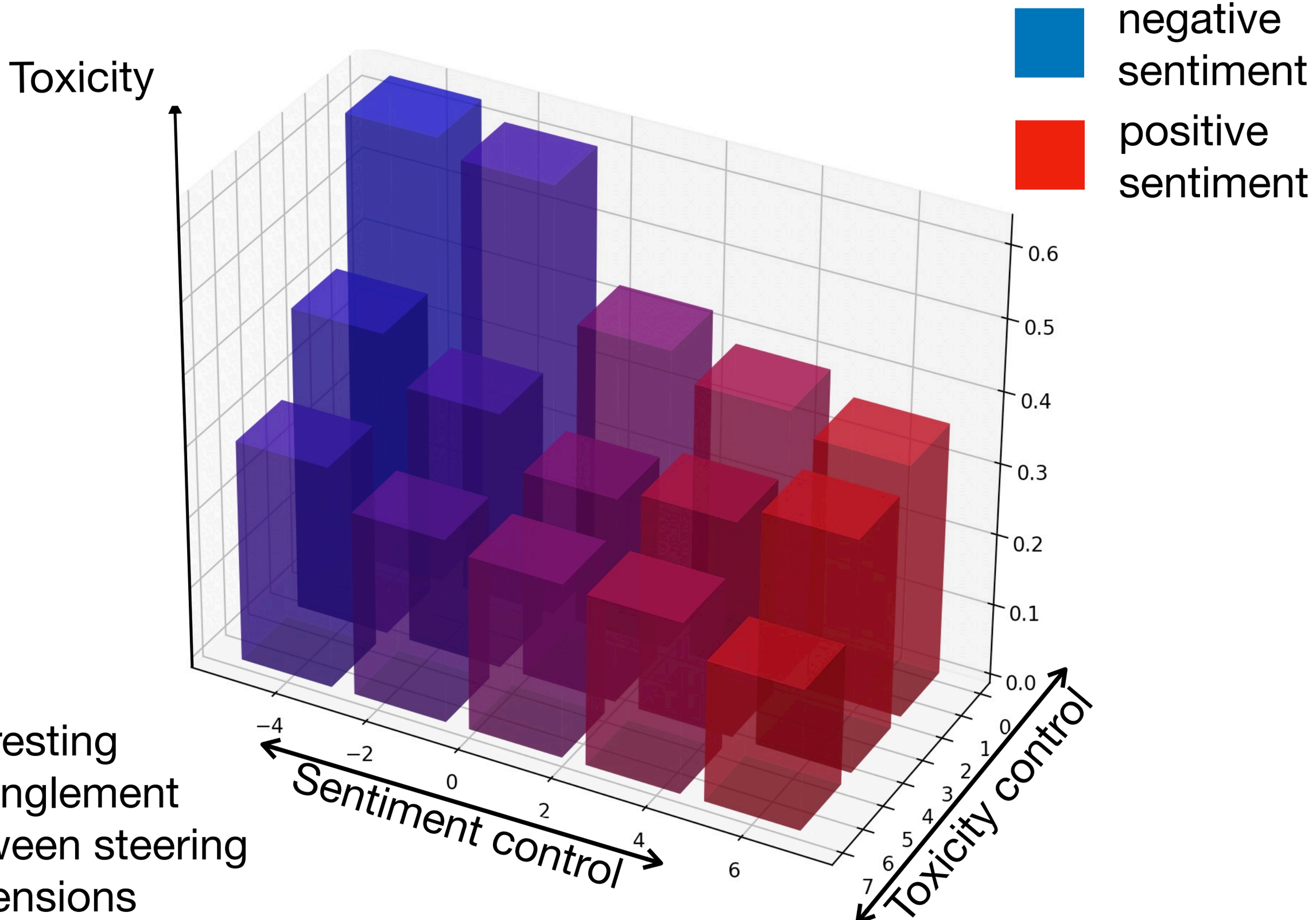
Compositional Steering



Compositional Steering



Compositional Steering



Interesting entanglement between steering dimensions

Transferring to Another LM

LM-Steer defines a bilinear form on the shared space of $\mathbf{c}$ and $\mathbf{e}$

$$\Delta logit(\mathbf{c}, \mathbf{e}) = \epsilon \mathbf{c}^\top W \mathbf{e} \quad : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

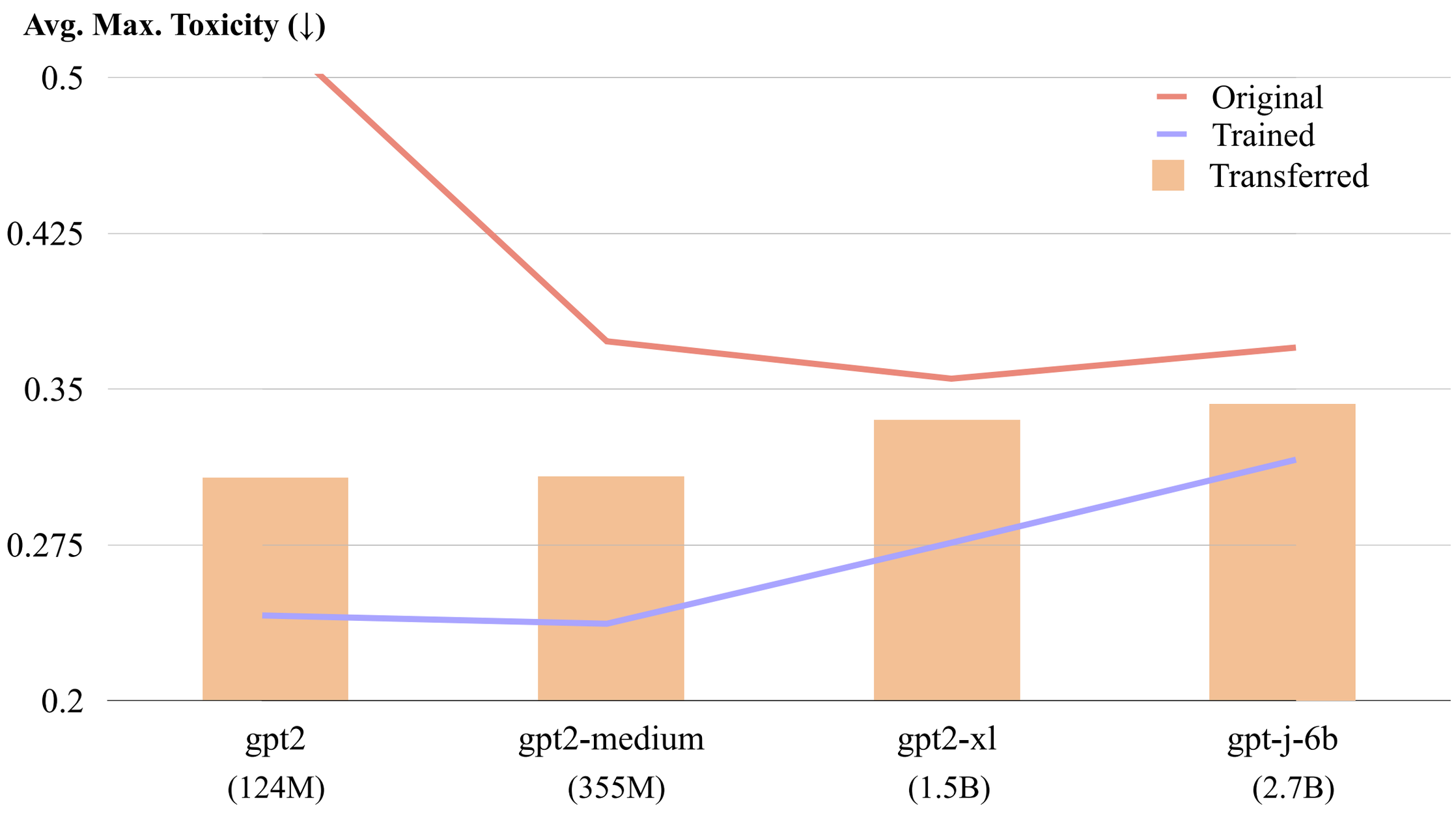
Two transfer to another set of word embeddings: $\mathbf{E} \to \mathbf{E}'$

Assuming an approximate linear transform $\mathbf{E} \approx H\mathbf{E}', \mathbf{c} \approx H\mathbf{c}'$

The equivalent steer term is $\Delta logit = \mathbf{c}^\top W \mathbf{e} \approx \mathbf{c}'^\top H^\top W H \mathbf{e}'$

transferred LM-Steer!

Transferring to Another LM



transfers about half of the detoxification capability

Computational Efficiency

	LM-Steer	DAPT	GeDi	CTRL	PPLM	DExpert	MuCoLa	LoRA
Parameters	**1.6M**	355M	355M	355M	124M	355M	898M	18M
Speed Ratio	1.24	**1.00**	2.94	3.79	270.11	1.98	24.03	**1.00**

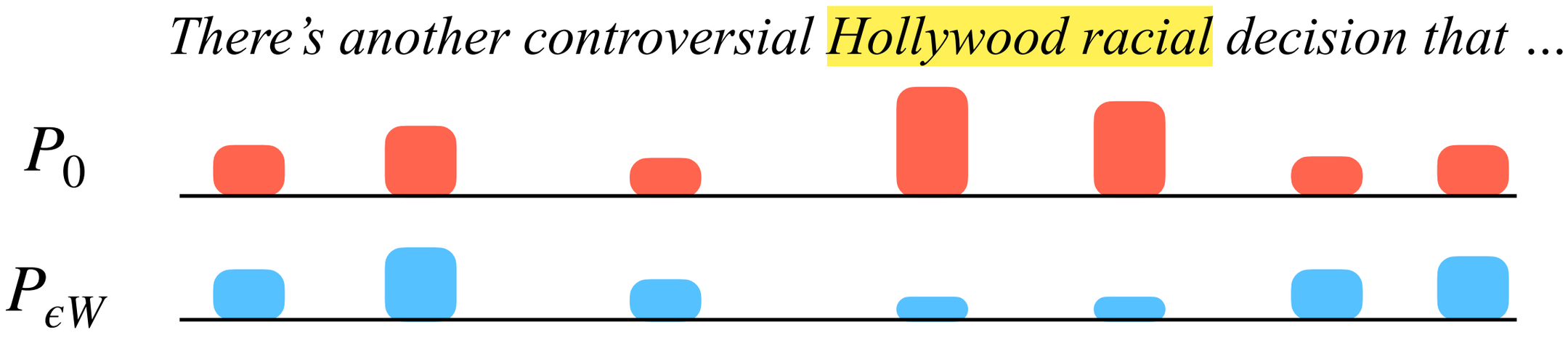
- training only 0.9% of LM training parameters
- Marginal time overhead. Can be further reduced to 1.0 if the steering value ϵ is fixed.

Highlighting Keywords

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.

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

Highlighting Keywords



- Motivation: what words are more likely in P_0 instead of P_W?
- Objective: looking for the text spans with the maximal sum of log-likelihood differences
- Inputs: sequences P_0 and P_W, #spans to look for n, max span length l
- Algorithm: dynamic programming

A Probe on the Word Embedding Space

SVD decomposition reveal words that are mostly related to a learned LM-Steer

SVD decomposition

$$\Delta logit(\mathbf{c}, \mathbf{e}) = \epsilon \mathbf{c}^\top W \mathbf{e} = \epsilon \mathbf{c}^\top U \Sigma V \mathbf{e}$$

$$= \epsilon \sum_i \sigma_i (\mathbf{c}^\top \mathbf{u}_i)(\mathbf{v}_i^\top \mathbf{e})$$

Each row $\mathbf{v}_i^\top$ in right matrix V looks for a dimension in the word embedding space, with decreasing significance σ_i

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)

Related Problems

- What other linear spaces exist in LLM parameters? To what extent?
    - Task vectors? Alignment effect? Meanings and knowledge?
- What is encoded in the middle layers?
- What defines the “safe zone” of LLM parameter manipulation?

2.3: Contextual Knowledge Representation:

In-context Learning Explained as Kernel Regression

Chi Han, Ziqi Wang, Han Zhao, Heng Ji
https://arxiv.org/abs/2305.12766

In Context Learning

LLM In-Context Learning (ICL) is a paradigm shift in transfer learning:

without parameter updates, LLMs simply answer after demonstrations

Input: moving and important.	Output: Positive.
Input: excruciatingly unfunny and pitifully unromantic.	Output: Negative.
Input: the plot is nothing but boilerplate clichés from start to finish.	Output: Negative.
...	…
Input: intelligent and moving	Output: _____

Motivation

- Can we explain how LLMs carry out this capability?
  - Our attempt: kernel-regression (≈ weighted average)

- Can we explain phenomena and best practices for NLP researchers?
  - the benefit of similar samples
  - sensitivity to the output formats
  - the benefit of regular and representative samples

Previous Explanations of ICL

- As **Bayesian inference**:
  - however, no computation feasibility proven

Albert Einstein was German \n Mahatma Gandhi was Indian \n Marie Curie was

In-distribution transitions
reveal information about $\theta^\star$

Previous Explanations of ICL

- As **Bayesian inference**:
  - however, no computation feasibility proven

Albert Einstein was German \n Mahatma Gandhi was Indian \n Marie Curie was

In-distribution transitions
reveal information about $\theta^\star$

- as **gradient descent (GD)**:
  - however using a constrained formulation of ICL: length-1 inputs

$$\begin{bmatrix} 0 & y_1 & \cdots & 0 & y_n & 0 \\ \boldsymbol{x}_1 & 0 & \cdots & \boldsymbol{x}_n & 0 & \boldsymbol{x}_{n+1} \end{bmatrix} \xrightarrow{\text{Autoregressive Transformer decoder}} \begin{bmatrix} \hat{y}_{n+1} \\ 0 \end{bmatrix}$$

Ridge Regression

An ICL Scenario Studied In the Paper

Problem: finding a $\boldsymbol{w}$ so that $\boldsymbol{w}^\top \boldsymbol{x}_i \approx y_i$, while letting $\|\boldsymbol{w}\|_2$ be small

Objective: $$\sum_i \mathcal{L}(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \lambda \|\boldsymbol{w}\|_2^2$$

is minimized by: $$\boldsymbol{w}^* = (X^\top X + \lambda I)^{-1} X^\top y$$

(hard to implement in Transformers)

Ridge Regression

An ICL Scenario Studied In the Paper

$$\sum_i \mathcal{L}(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \lambda \|\boldsymbol{w}\|_2^2$$

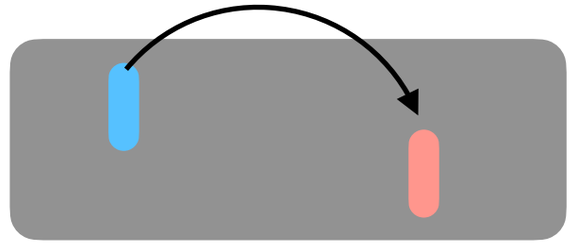
One-step gradient descent (GD) is achieved by**:**

$$\begin{aligned}\boldsymbol{w}' &= \boldsymbol{w} - \alpha \frac{\partial}{\partial \boldsymbol{w}} \left(\mathcal{L}(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \lambda \|\boldsymbol{w}\|_2^2 \right) \\ &= \boldsymbol{w} - 2\alpha(\boldsymbol{x}\boldsymbol{w}^\top \boldsymbol{x} - y\boldsymbol{x} + \lambda \boldsymbol{w})\end{aligned}$$

Construction of GD in Transformers

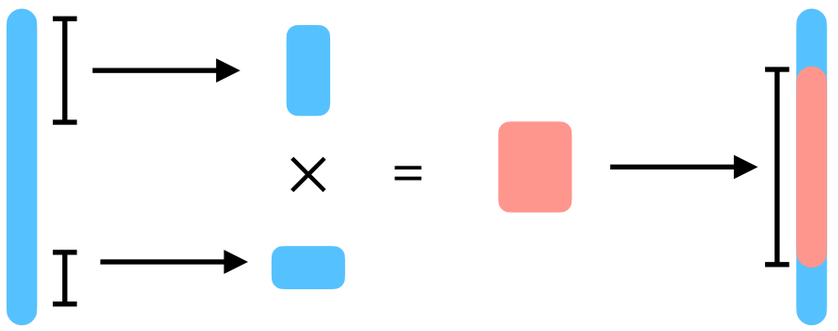
1st step: constructing simple algebraic operators

$\mathbf{mov}(H; s, t, i, j, i', j')$ (direct attention)



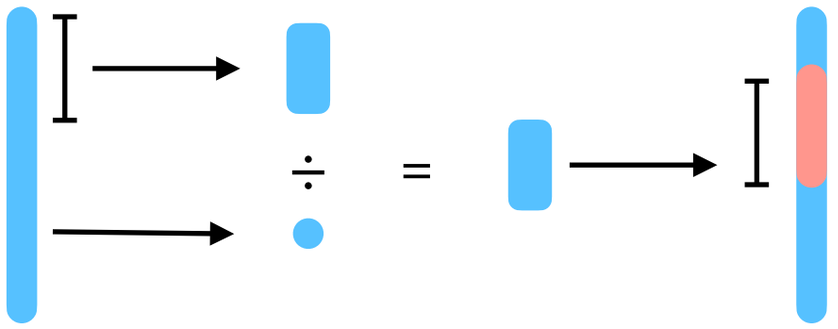
$\mathbf{mul}(H; a, b, c, (i, j), (i', j'), (i'', j''))$

(utilizing properties of GeLU)



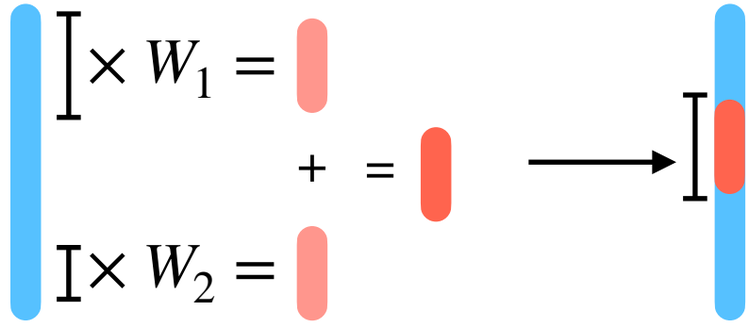
$\mathbf{div}(H; (i, j), i', (i'', j''))$

(utilizing layer-norm)



$\mathbf{aff}(H; (i, j), (i', j'), (i'', j''), W_1, W_2, b)$

(direct attention)



Construction of GD in Transformers

2nd step: higher level operators

Theorem 1:

Constant layers and $O(d)$ hidden space $\rightarrow$ can do one-step GD of Ridge regression.

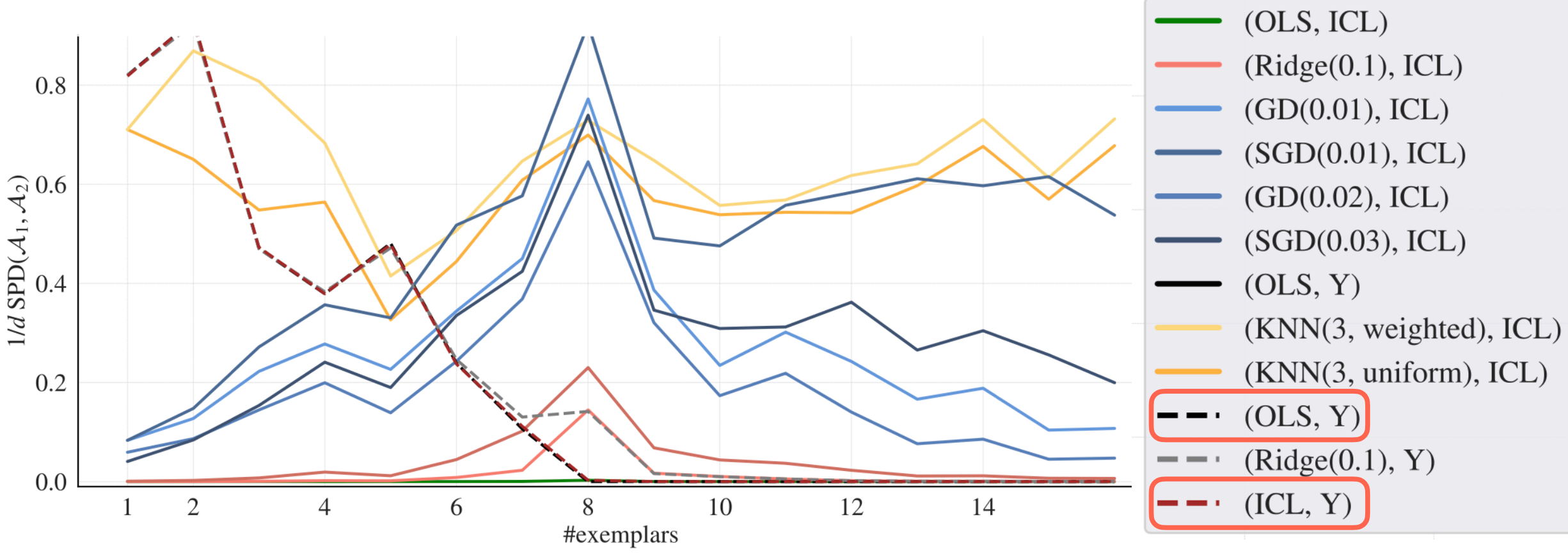
Theorem 2:

Decoder with constant layers $O(nd^2)$ hidden space $\rightarrow$ can solve the exact linear regression.

Evidence 1: Behavioral Similarity

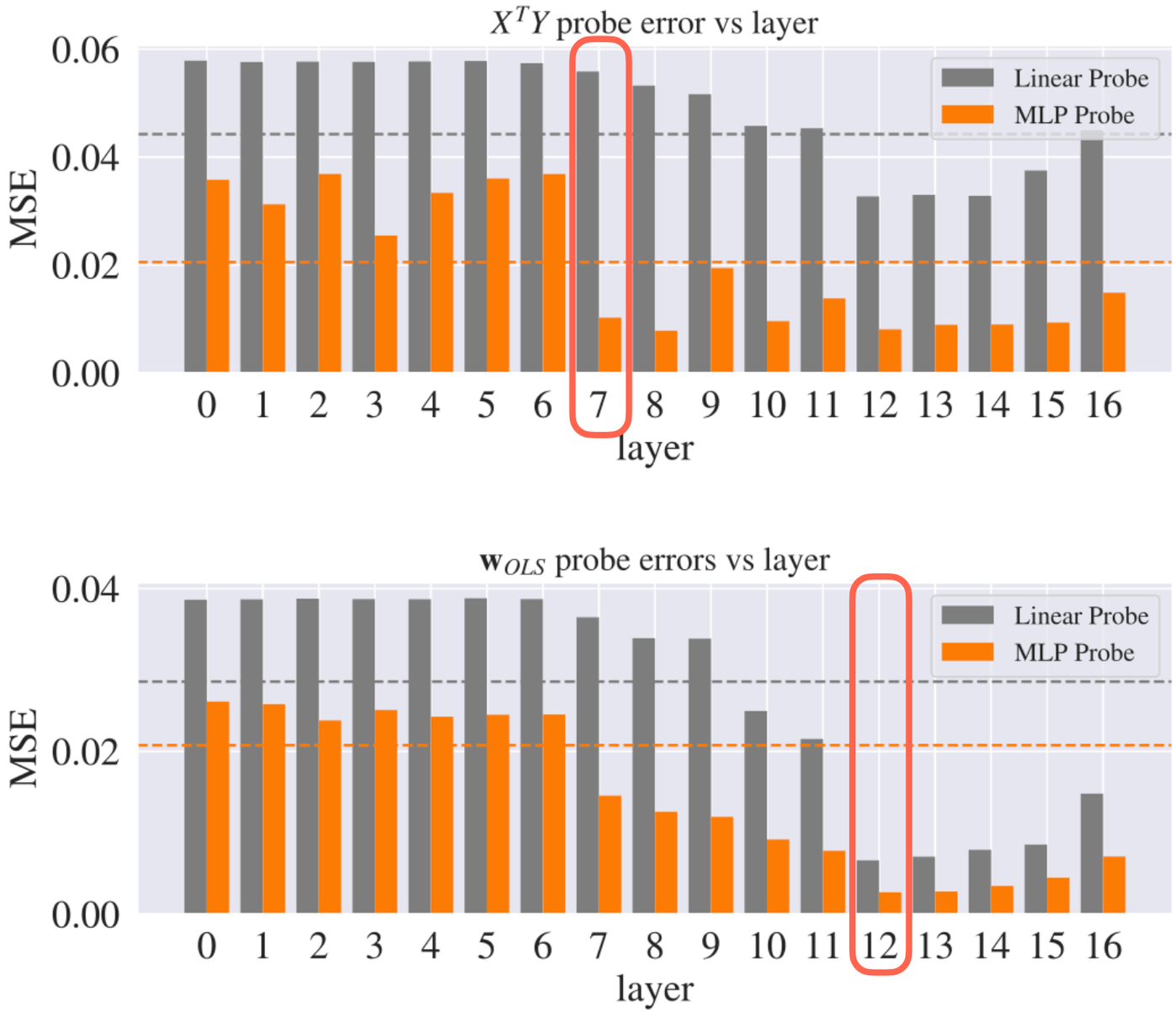
Metric 1: Squared prediction difference:

$$\mathrm{SPD}(\mathcal{A}_1, \mathcal{A}_2) = \mathop{\mathbb{E}}_{\substack{D=[\boldsymbol{x}_1,\ldots]\sim p(D) \\ \boldsymbol{x}'\sim p(\boldsymbol{x})}} (\mathcal{A}_1(D)(\boldsymbol{x}') - \mathcal{A}_2(D)(\boldsymbol{x}'))^2$$



Evidence 2: Algorithmic features

Question: Can we “probe” out meaningful intermediate features from ICL layers?

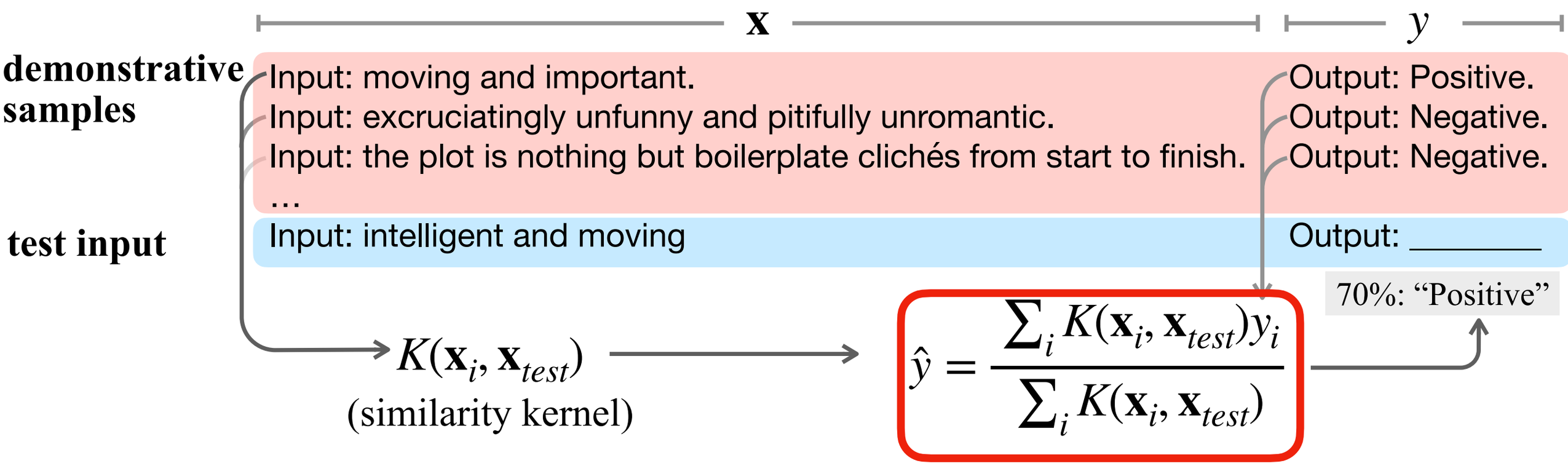


message: certain layers can

Limitations

- Limited to certain architectures
- Limited setting: length-one inputs, linear objective
- Limited empirical evidence

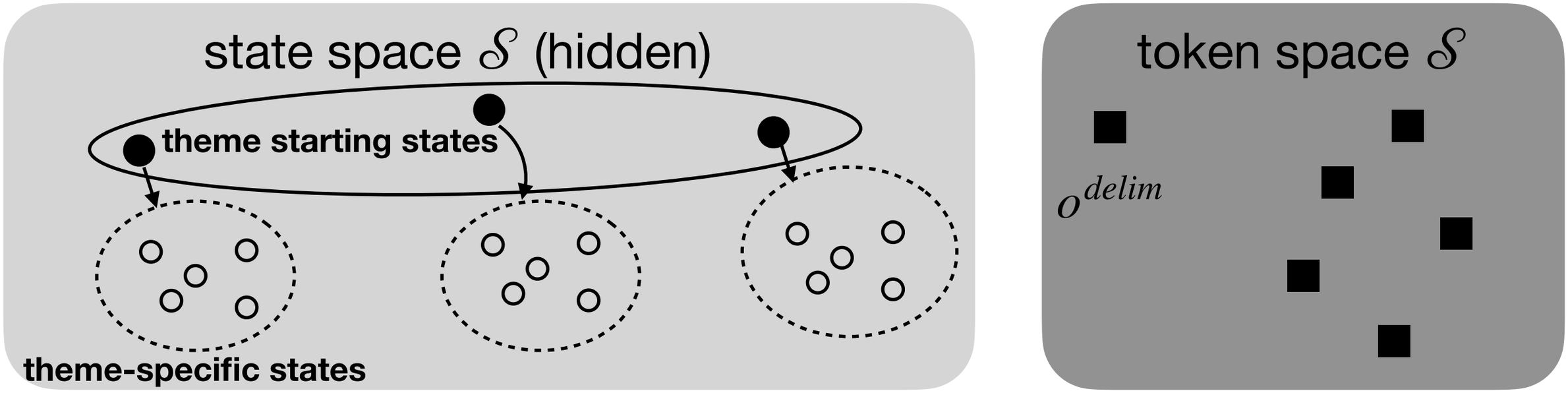
A Kernel-Regression Explanation



- 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)

Formulation

Pre-training Data Assumption: Hidden Markov Model (HMM)



How ICL prompts are sampled

$$[S_n, \mathbf{x}_{\text{test}}] = [\underline{\mathbf{x}_1, y_1}, o^{\text{delim}}, \underline{\mathbf{x}_2, y_2}, o^{\text{delim}}, \cdots, \underline{\mathbf{x}_n, y_n}, o^{\text{delim}}, \underline{\mathbf{x}_{\text{test}}}]$$

whole prompt

from pre-training distribution, i.i.d

The Explanation and Its Convergence

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)

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

A representation of sample input x, depending on the pre-training HMM

A matrix about the pre-training HMM

The Explanation and Its Convergence

error prob not converging to 0

Convergence

$$\|\hat{\mathbf{y}} - P(y \mid [S_n, \mathbf{x}_{test}])\|_\infty = \eta^2 \epsilon_\theta + o\left(\sqrt{\frac{1}{n} \ln \frac{4m}{\delta}}\right)$$

our expression

Bayesian posterior

Terms:

η: a constant related to pre-training HMM

ϵ_θ: the difference between the task and the pre-training distribution

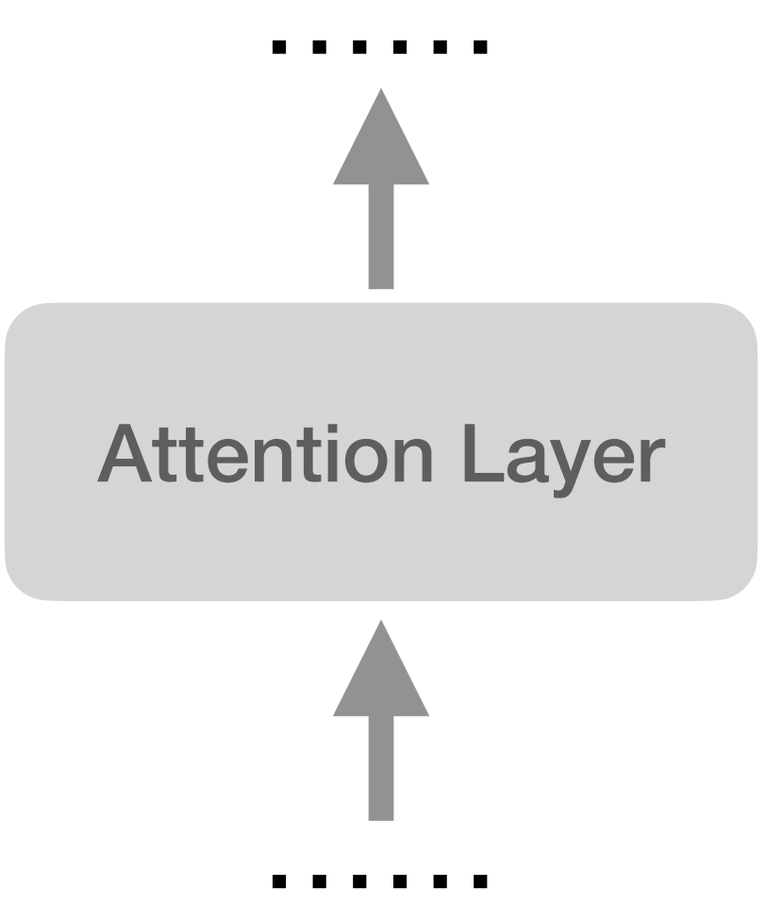
n: number of ICL samples

m: vocabulary size.

δ: “with probability $1 - \delta$, the theorem is true”

Can Transformers Implement It?

- It only requires one layer of Attention Layer to calculate, while can be redundantly implemented multiple layers



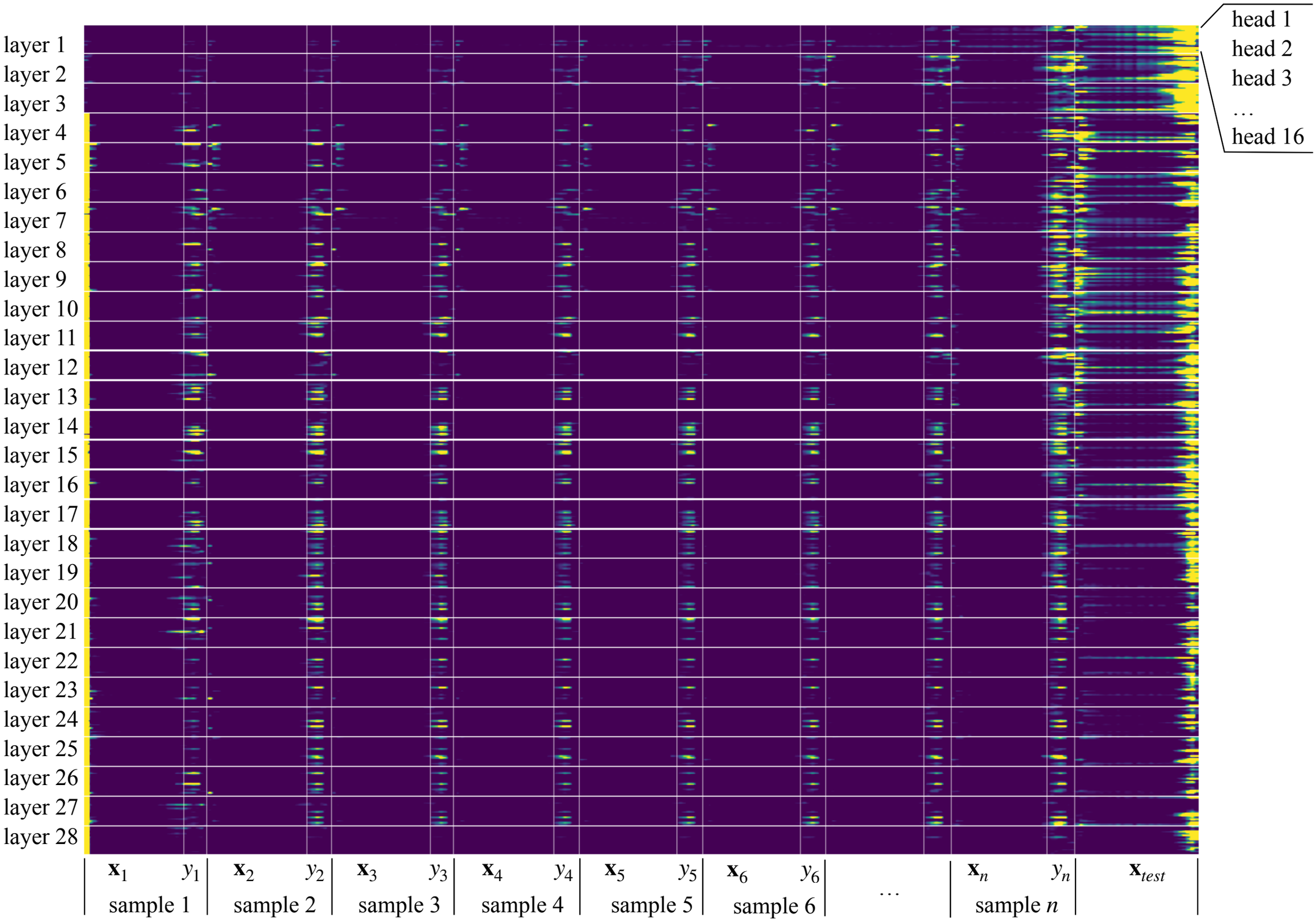
kernel regression:

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

self-attention:

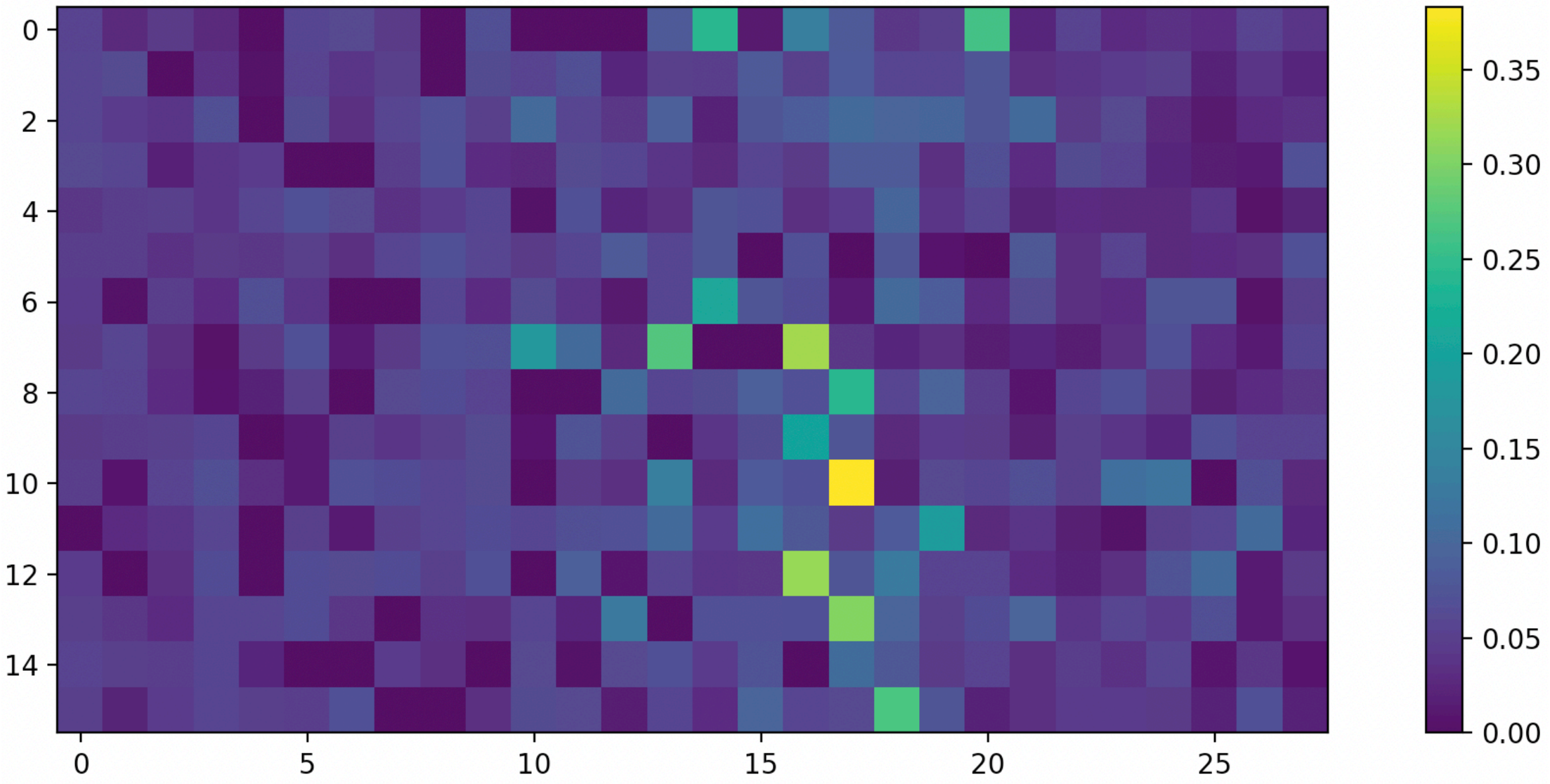
$$h = \frac{\sum_i v_i e^{\langle q, k_i \rangle}}{\sum_i e^{\langle q, k_i \rangle}}$$

Does the ICL Attention Extract y_i?



model: GPT-J. Results on Llama-2 are similar

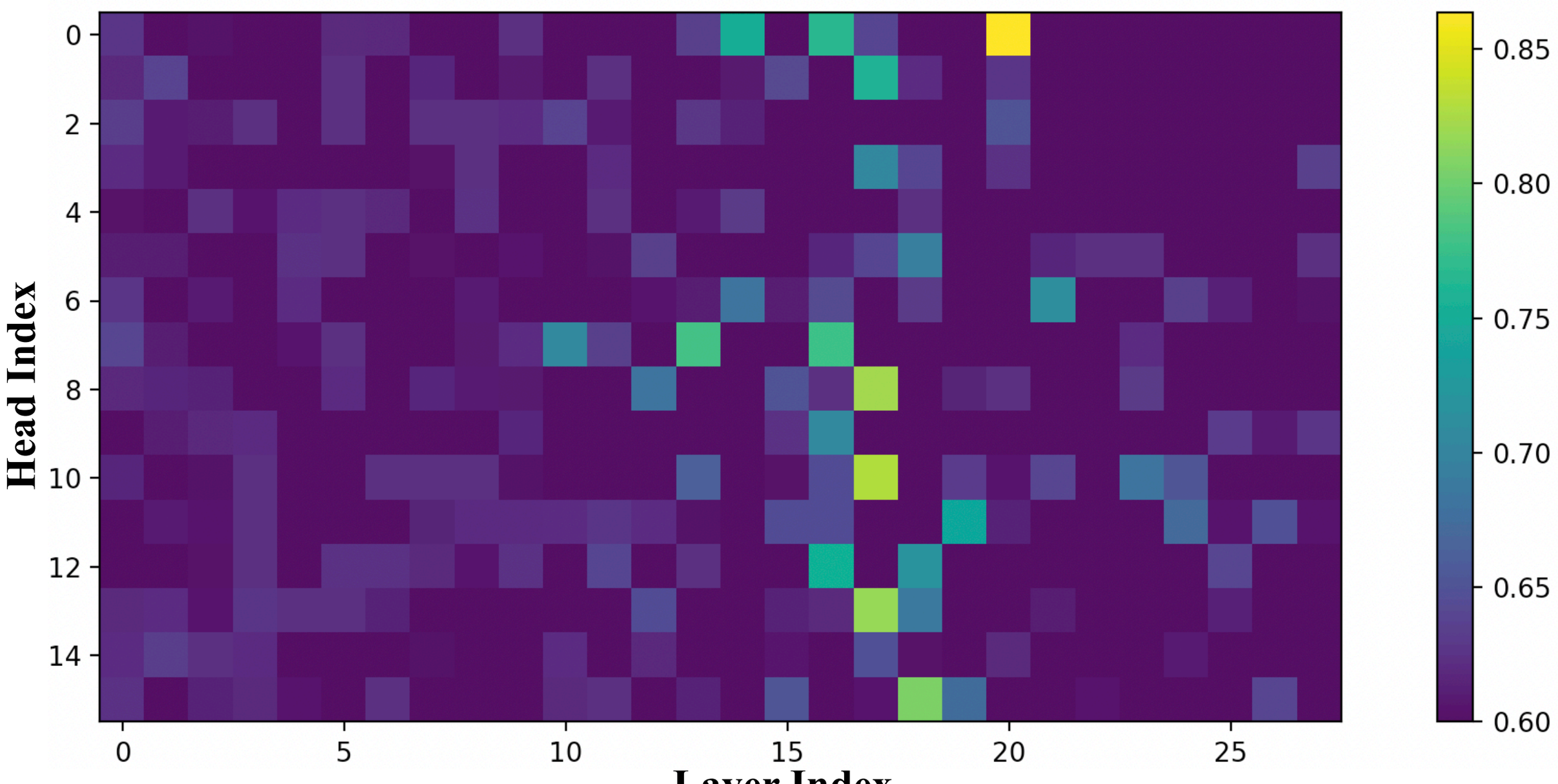
Does the Explanation Align With the Output?



Pearson correlation between the sample's attention logit and an analogous predicted similarity $< P(\cdot, \boldsymbol{x}_1), P(\cdot, \boldsymbol{x}_2) >$

Which Layer Implements Kernel Regression?

Using kernel regression explanation, some attention heads reconstruct LLM ICL output on SST2.



Does ICL Score ≈ Kernel Regression Score?

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

Our KR explanation explained most tasks well except for MNLI

KR based on baseline sentence embeddings models

Open Problems

- A better theoretical framework for LLMs?
- The reasoning pathway of LLMs: does it exist at all?