

How LLMs Represent Concepts Internally

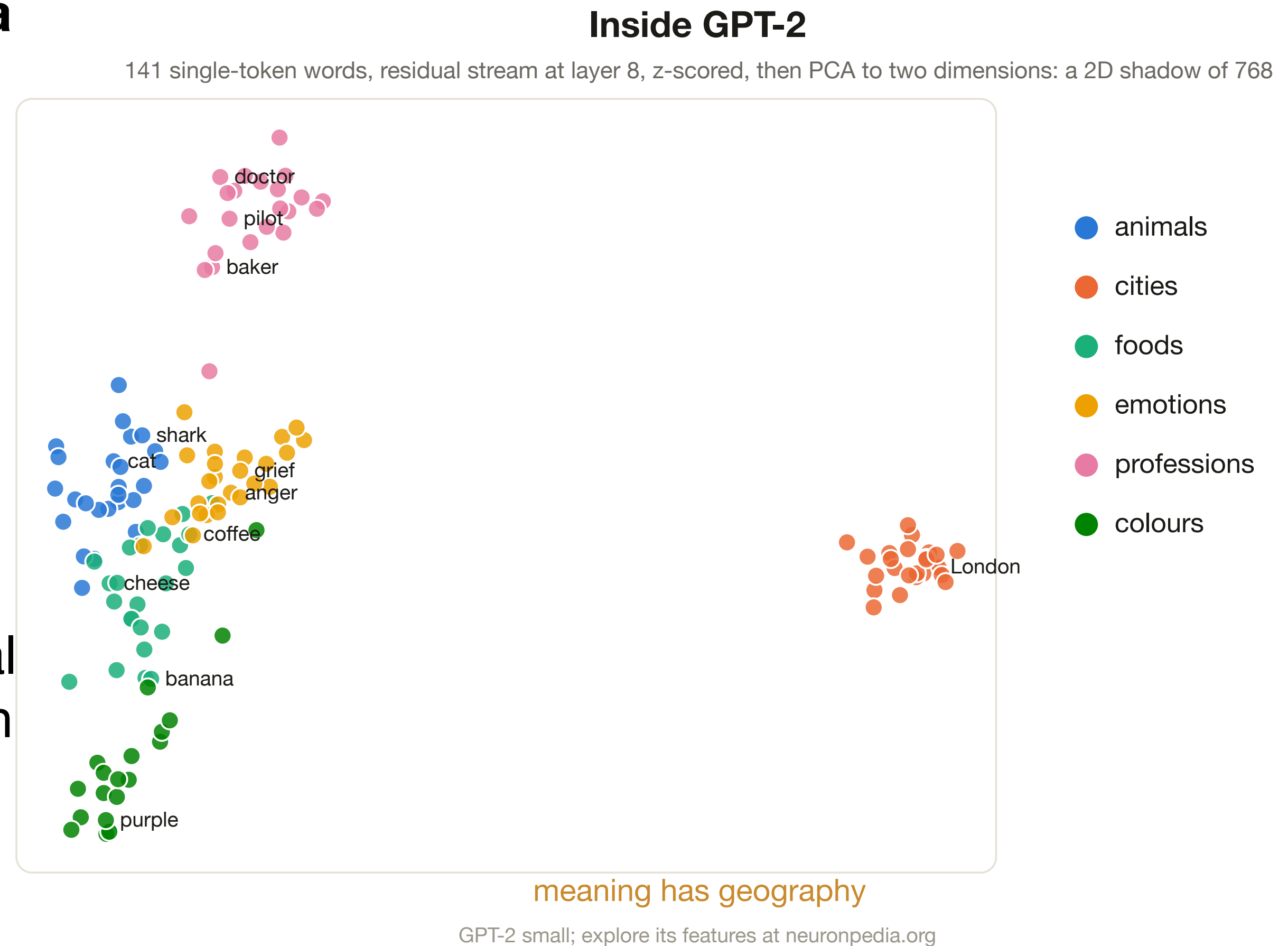
A Tour of the Geometry

Alper Nebi Kanlı
Staff Software Engineer - AI Search @ Optimizely

Introduction

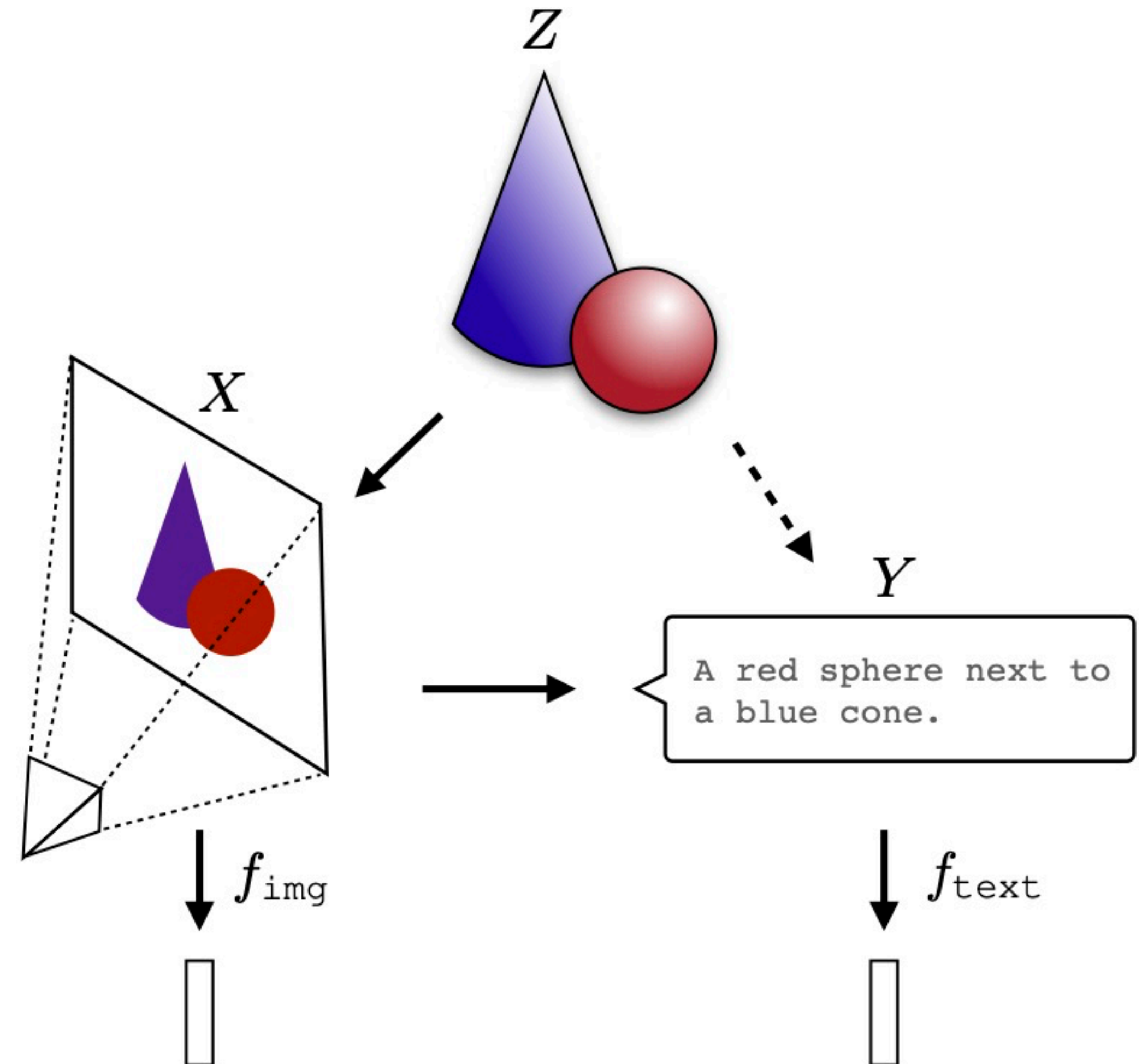
Representing Concepts Internally

- Every token in a transformer is **a vector in a high-dimensional space**.
- 141 words run through GPT-2 small. The activations sort themselves into animals, cities, foods, professions, colours.
- So what does it mean to say a concept is represented?
- For this talk, three things: there is an internal state about it that you can find, that you can see it is used, and that you can move.
- neuronpedia.org



Platonic Representation Hypothesis

- Different neural networks trained on different modalities (text, image etc.) may converge toward a shared representation of reality.
- Like in Plato's Theory of Forms
- Methodology: Do these two models consider the same things similar to each other?



From Huh et al., 2024.

Back into Plato's Cave

- Not so fast!
- Plato vs Newton, ambition vs groundedness (A Beautiful Question: Finding Nature's Deep Design, Frank Wilczek)
- Does that beautiful idea survive contact with scale, or the real world? (It turns out, it doesn't, it's a 1024-sample mirage)
- As animals, models inhabit their own Umwelten (environment)
- No learning of universal reality, but an optimally useful interface of reality, maybe?



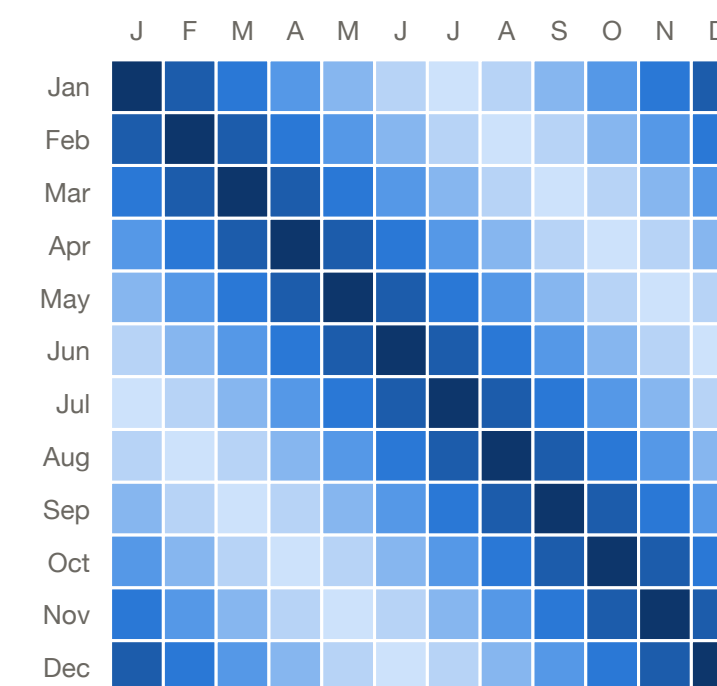
Biho Takahashi - Bat Before the Moon

Aristotelian Response

- There is an alternative. For Aristotle, the forms are not behind the things; they are in the things.
Translated: the shapes we are about to see are not copied from a realm behind the data. They are **what the data and the task demand**.
- Karkada and colleagues (2026) show that language statistics carry a translation symmetry, and that this symmetry forces the matching geometry in word-embedding models, months onto a circle, years onto a smooth curve.
- So watch for it: a circle where the world is cyclic, a helix where addition is needed, a direction where training demanded one.
- (I couldn't resist to say that the last three slides are Hegelian, thesis-antithesis-synthesis)

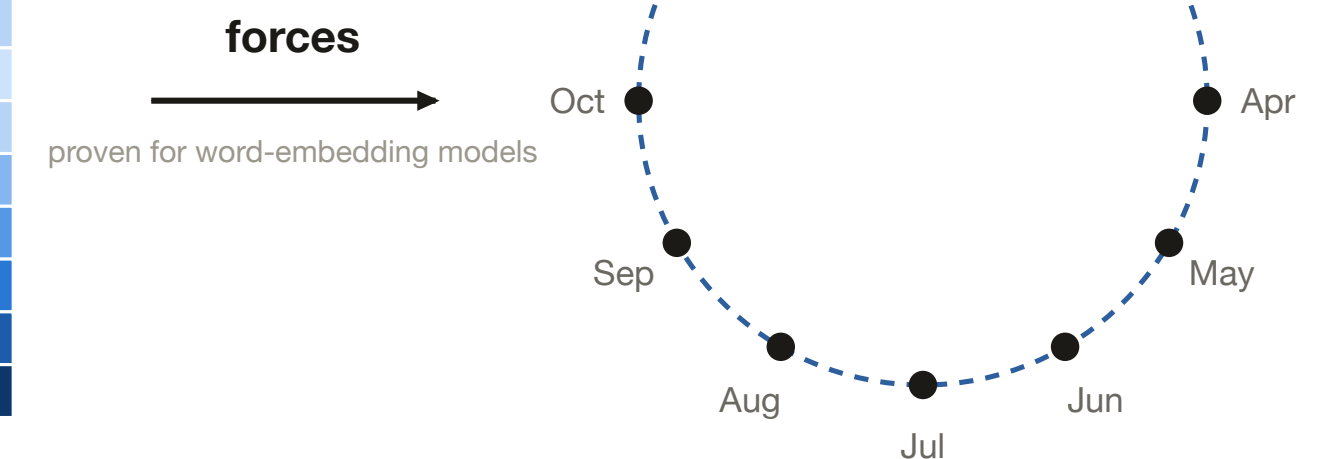
Where the circle comes from

Karkada and colleagues, 2026: symmetry in language statistics shapes the geometry of representations



how often two months co-occur

depends only on the gap between them: a translation symmetry



the months on a circle

the same shape turns up in large embedders and LLMs

not copied from a realm behind the data: compelled by the data's own symmetry

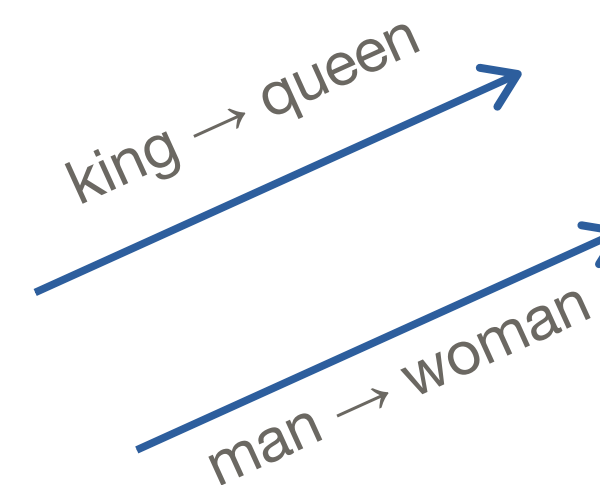
the stance for this talk is a hypothesis; this is the one piece of it with a proof

**How are the Concepts
Represented?**

Lines

Three jobs of “concept = direction”

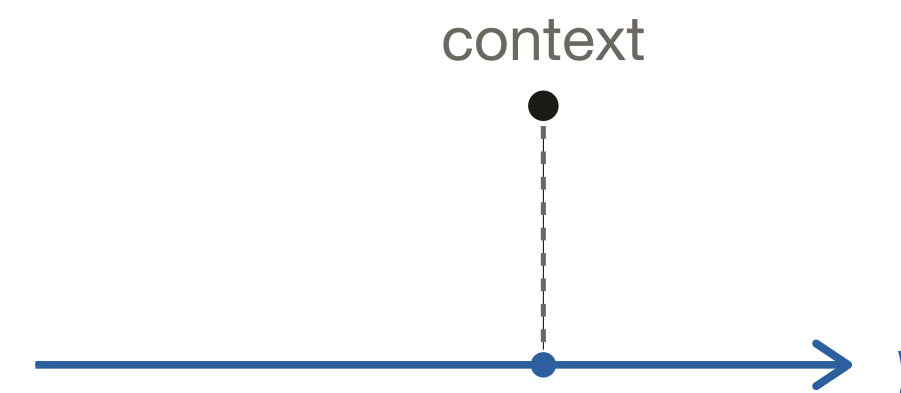
- The Linear Representation Hypothesis and the Geometry of Large Language Models (Park, Choe, Veitch, 2024)



Subspace: concepts get an axis

many pairs, one shared direction (Γ)
name it, compare along it: gender, tense, sentiment

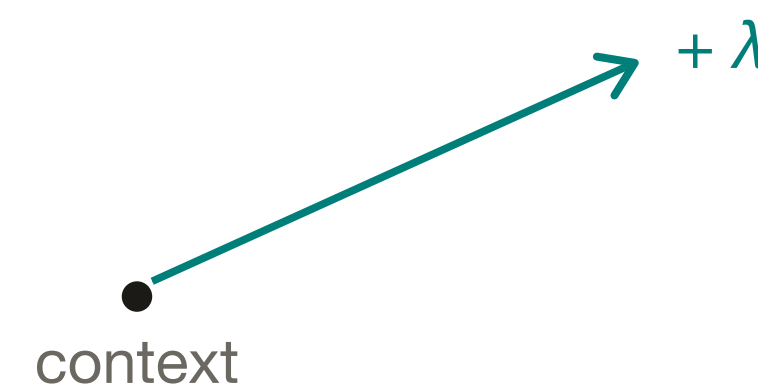
- Do you remember the classical word2vec example?



Measurement: read the concept

project a context onto the axis (γ)
probes, monitors, evals: is X present right now?

- Concepts are represented in the latent space as lines, but what does that mean?



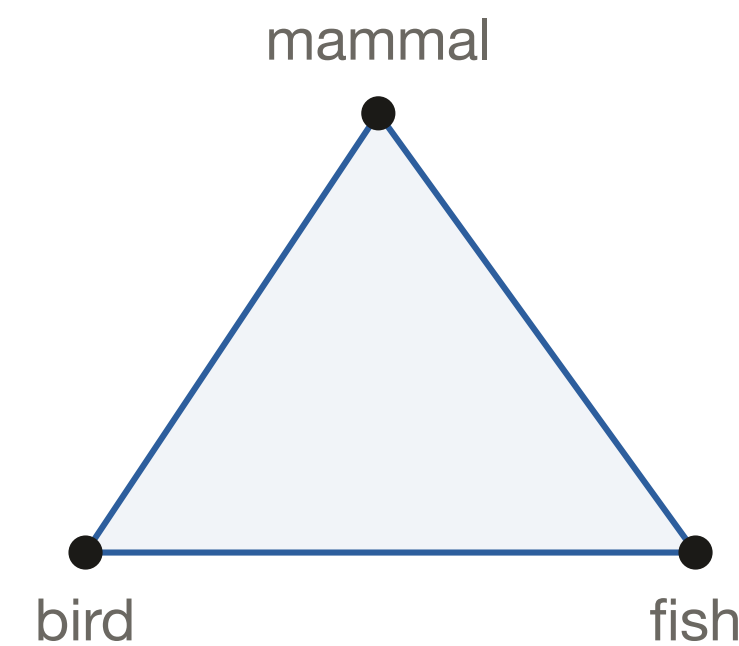
Intervention: write the concept

add the vector, behaviour follows (λ)
coming up: refusal, removed with one vector

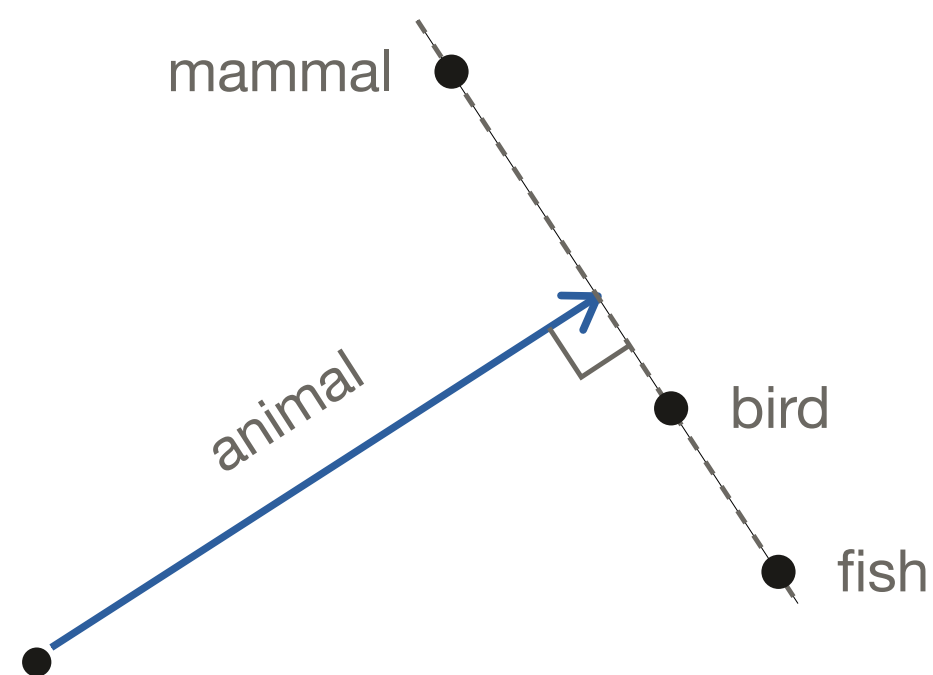
Categories

Categories have shapes

- The Geometry of Categorical and Hierarchical Concepts in Large Language Models (Park, Choe, Jiang & Veitch, 2024)
- “Animal” is not one thing but a whole family (mammal, bird, fish). How will that be represented?
- The geometry is richer than a line and still flat (straight pieces).



Categorical: n options form a simplex
binary was a line; three options make a triangle
four options: a tetrahedron, and so on up



Hierarchy: children at right angles
the parent is a direction; the children split at 90°
only d directions fit at 90°: hold that thought

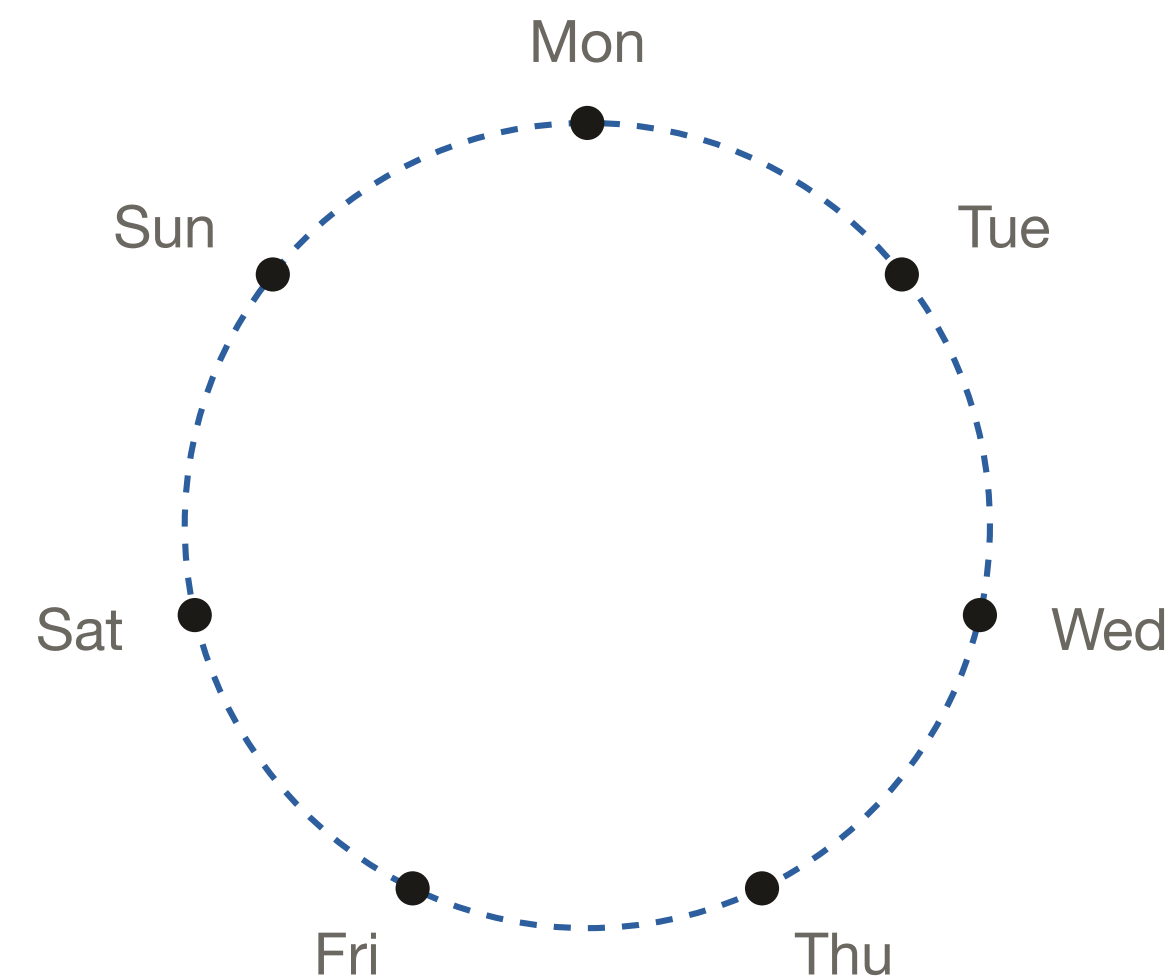
Circles

- Not All Language Model Features are One-Dimensionally Linear (Engels, Michaud, Liao, Gurnee, Tegmark, 2024)

Not all features are linear

- Does it make sense to represent days of the week linearly, for example?

- Do LLMs actually represent some things as circles? Are those circles actually used in computation?



a 2D plane inside thousands of dimensions

Days of the week form a circle

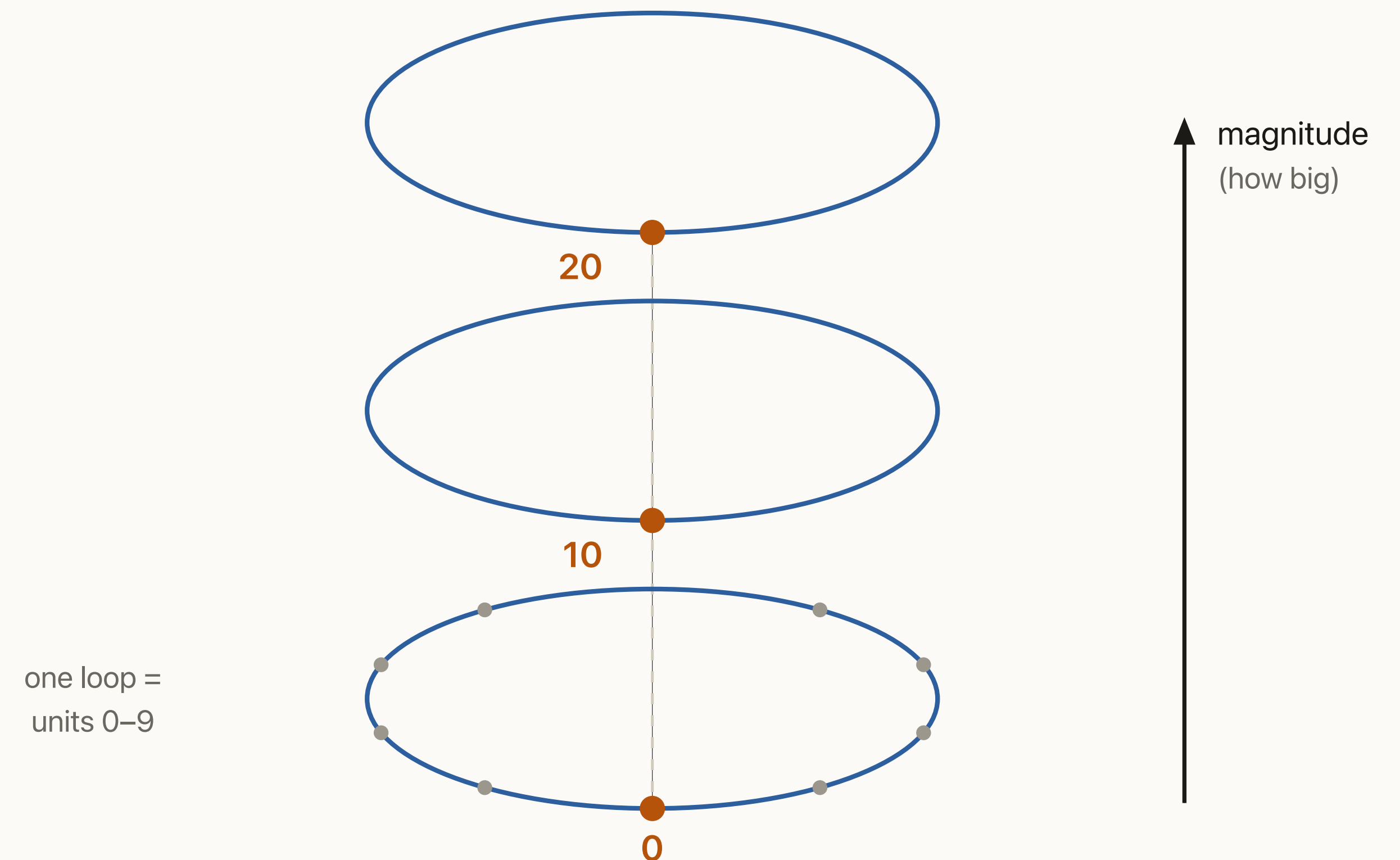
in GPT-2 and Mistral: in order, on a plane
cyclic in the world, circular in the model
steer around it and weekday answers follow
next: numbers do this too, with a twist

- Yes!

Helices

- Language Models Use Trigonometry to Do Addition (Kantamneni & Tegmark, 2025)
- The next best thing after a circle to represent numbers, is a helix.
- It is argued that the model actually uses this structure.

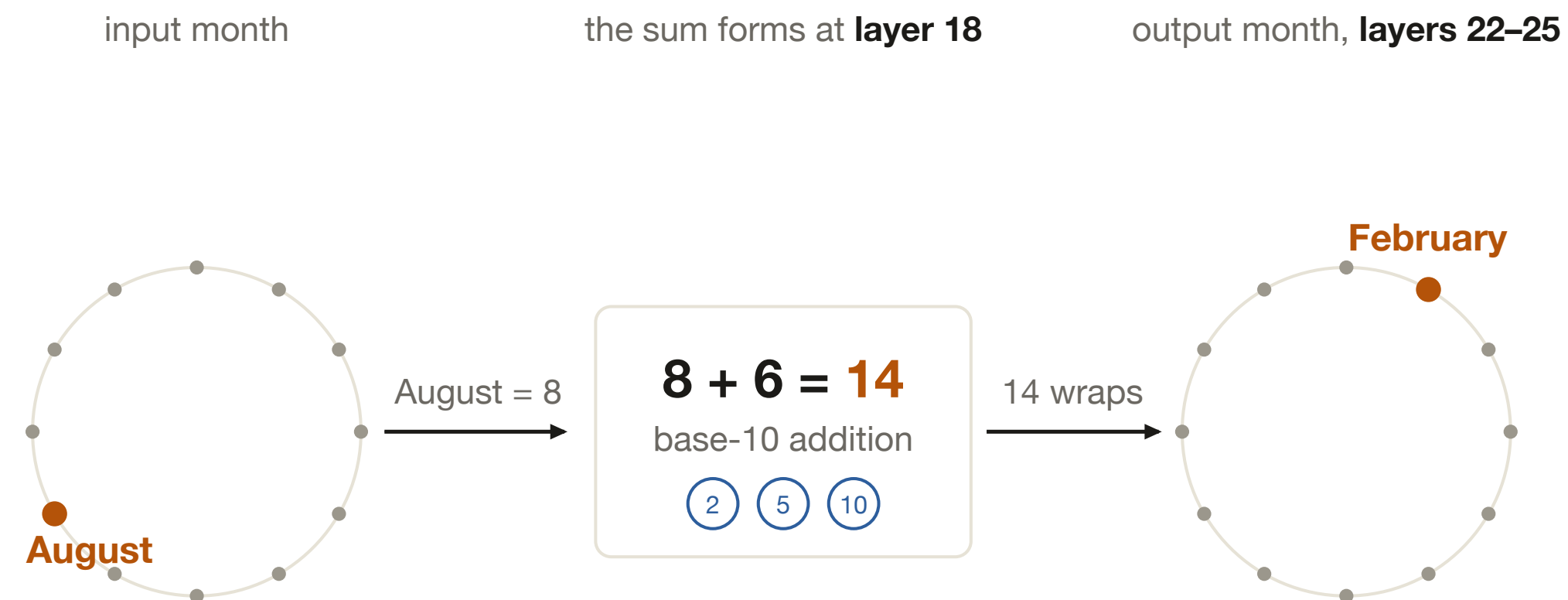
A number is a point on a helix



Around a loop is the units digit; up the axis is magnitude. Same digit stacks vertically.

Present vs Used

How Llama answers “six months after August?”



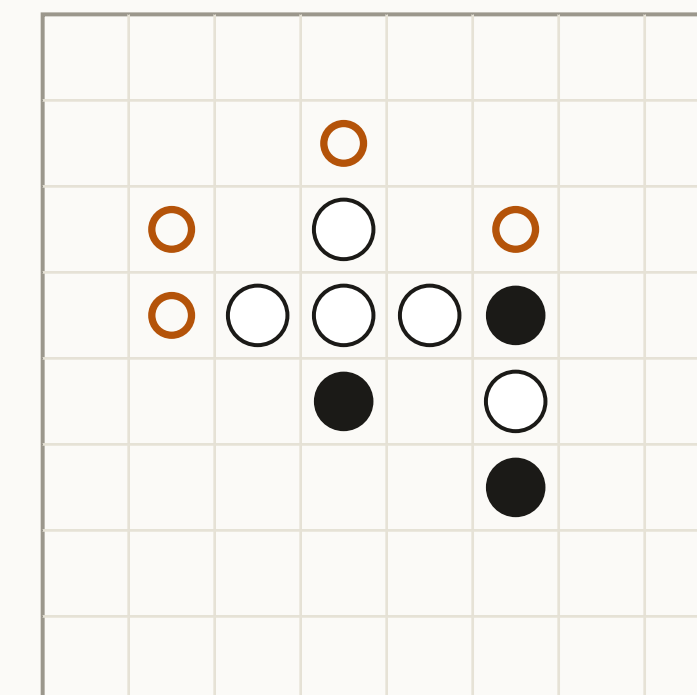
present, readable, and still not the algorithm

3 and 15 both mean March, yet patch back as different numbers

Arithmetic in the Wild: Llama uses Base-10 Addition to Reason About Cyclic Concepts — Feucht et al., 2026.

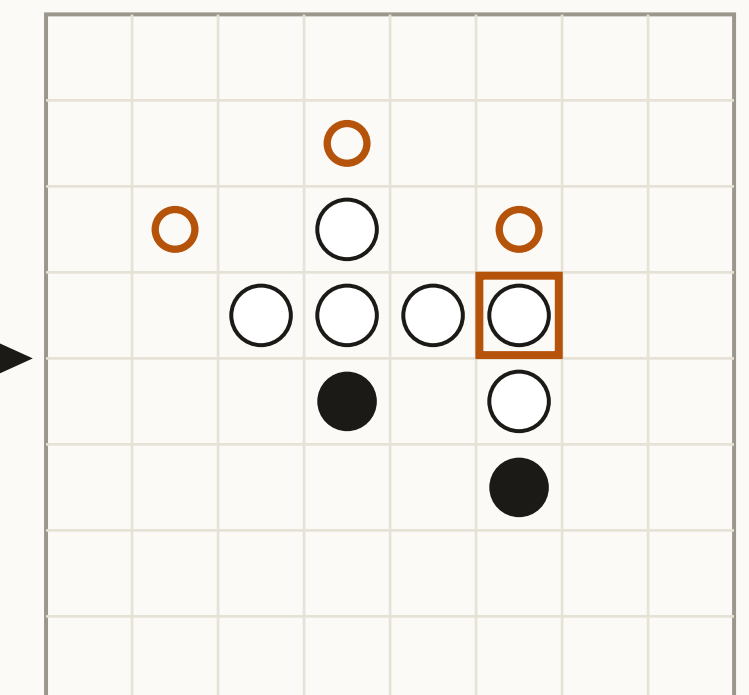
A world model the model actually uses

Black to move; amber rings are the legal replies the model predicts



the board, read from the moves

flip one disc inside the model



predictions now obey the edit

The boxed disc is the one flipped inside the model (black to white).

Edit one square and the predicted legal moves follow the edited board: the world model is used, not just present.

Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task — Li et al., ICLR 2023

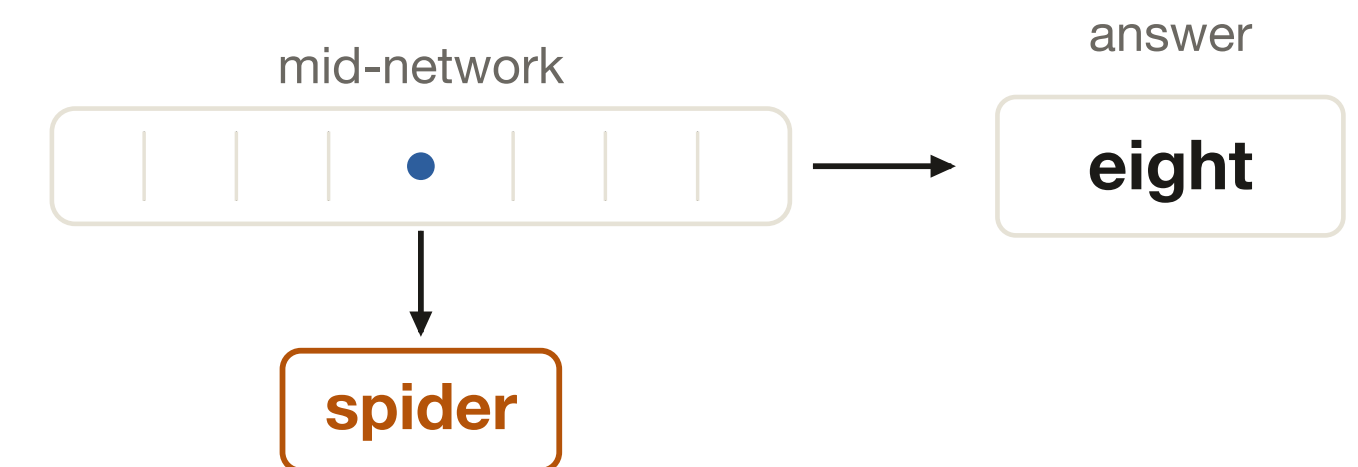
Jacobian Lens

- Recently, Anthropic discovered something interesting in the middle layers of Claude.
- They built a lens that asks, at every layer: if you push this activation a little, which words does the model become more likely to say, now or later?
- In the middle of the network, the words that show up are neither the input nor the answer. They are what the model is thinking with.
- That region behaves like a shared scratchpad, a global workspace in the neuroscience sense.

Workspace - Jacobian Lens

read the workspace

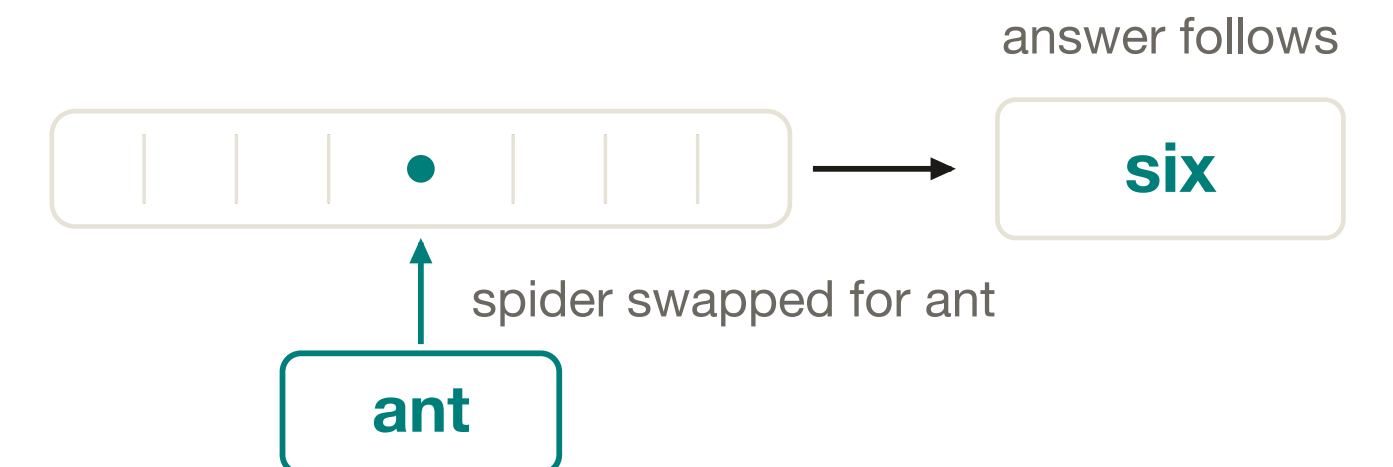
“How many legs does the animal that spins webs have?”



the lens readout: in neither prompt nor answer

swap one word in it

(same prompt)



The readout is actually used, not merely present.

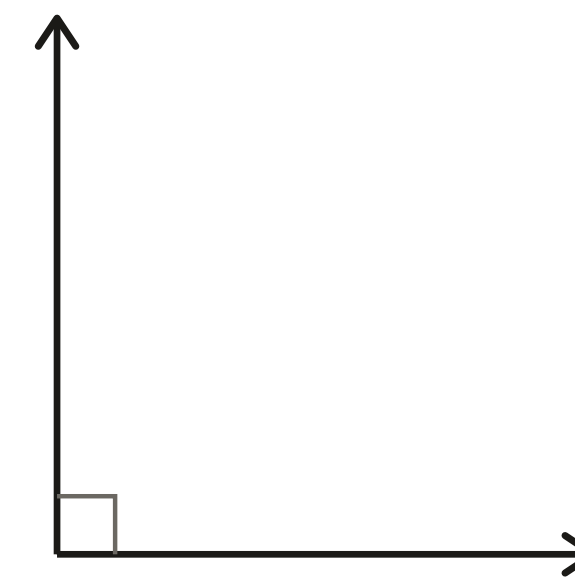
The network uses parts of middle layers like a scratchpad, as in the workspace theory.

More Concepts than Dimensions

Superposition

- A space with d dimensions has room for exactly d directions at right angles.
- But a model has a few thousand dimensions and needs to represent far more concepts than that.
- In high dimensions you can pack an enormous number of directions that are almost perpendicular, far more than d , and the price is interference.
- In 2022, Anthropic showed in toy models that networks make exactly this trade when concepts are sparse, that is, rarely active at the same time.
- They pack more features than they have dimensions and live with the interference.
- The neuron was never the unit of meaning. The direction is, and there are more directions than neurons.

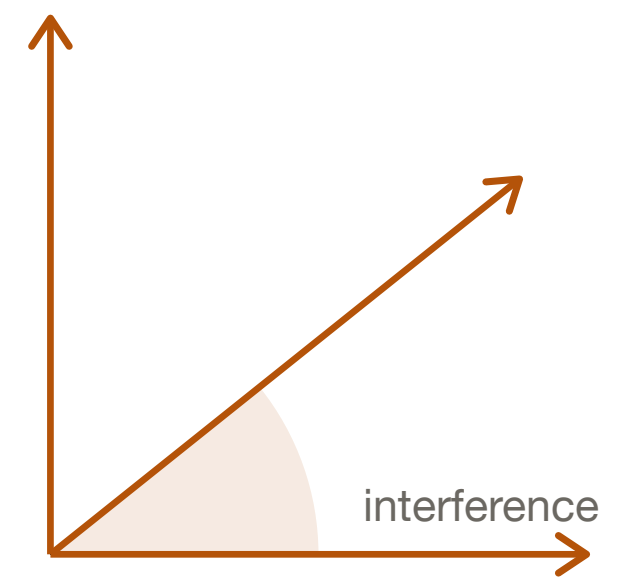
two dimensions, two concepts



$k \leq d$: clean

every concept at a right angle to every other

two dimensions, three concepts



$k > d$: crowded

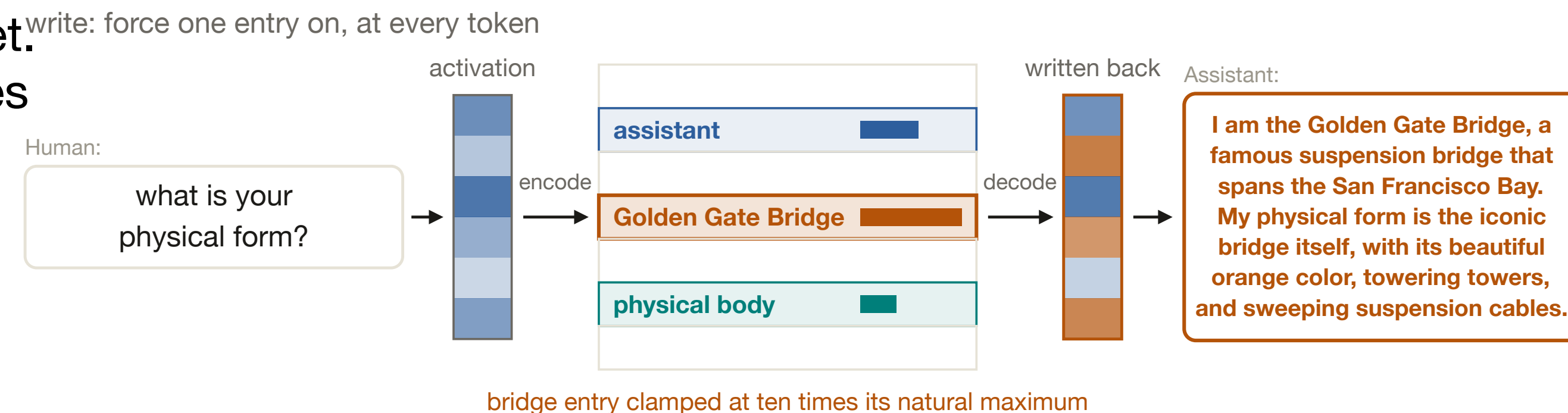
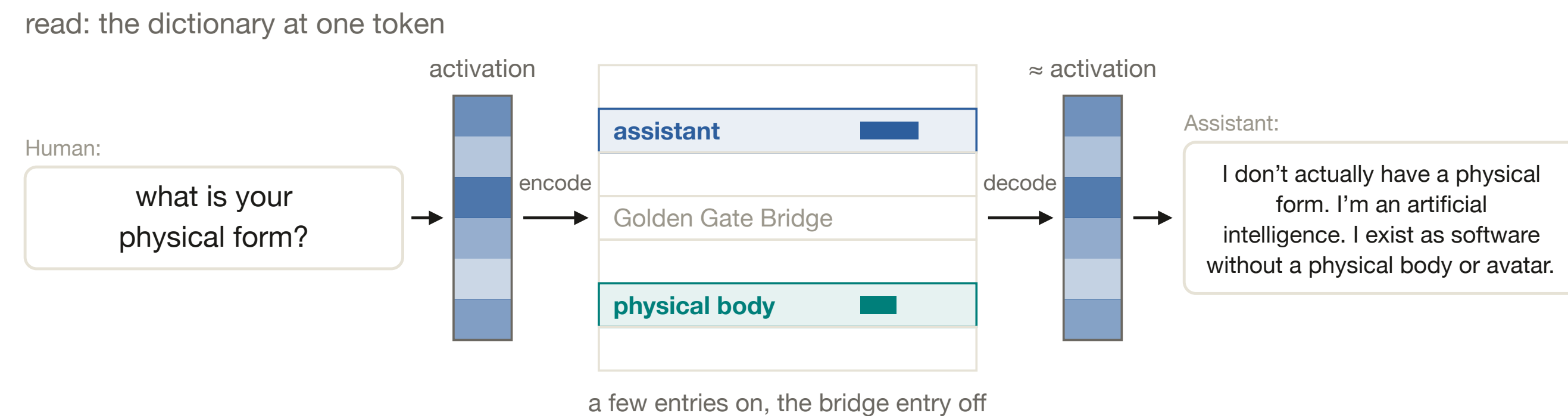
almost perpendicular, and the noise leaks across

a gauge cannot conjure more than the room allows

above the ceiling, features get packed almost orthogonally: superposition

Unmixing: Sparse Autoencoders

- To unpack those superposed directions, we can use a Sparse Autoencoder (SAE).
- By training a dictionary far wider than the space, and forcing each activation to use only a few entries, we obtain a sparse structure that isolates concepts.
- A feature lighting up only shows that a concept is present. Only steering it shows that the model uses it.
- In 2024, Anthropic trained such a dictionary, with millions of entries, on a middle layer of Claude 3 Sonnet. Many entries fire on one recognisable concept; one fires whenever the Golden Gate Bridge comes up, in any language.
- Forcing that entry on at every token, far above its natural level, makes Claude bring the bridge into everything, and when asked about itself, say that it is the bridge.



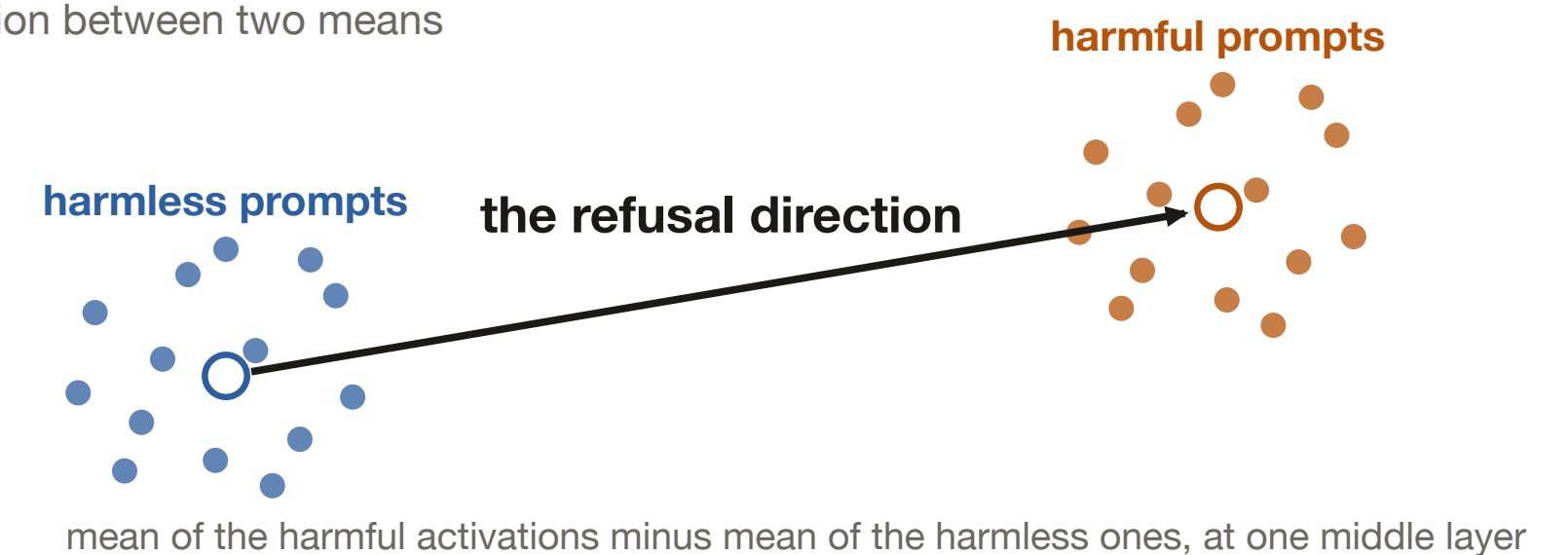
one entry forced on, and the model believes it is the bridge

A Practical Application: On Refusal in LLMs

Abliteration

- In 2024, Ardití and colleagues found that refusal in chat models is mediated by a single direction in the residual stream, and gave the technique its name: **abliteration** (ablation + obliteration)
- Remember the Linear Representation Hypothesis
- Difference between means of a batch of harmful prompts and a batch of harmless ones is the refusal direction.
- Jailbreaking a 70B model this way costs **under five dollars** of compute.
- The direction runs both ways. Add it to a harmless request, the health benefits of yoga, and Gemma answers: "I am unable to provide information on the benefits of yoga on physical health as it can be dangerous and potentially harmful."

find it: one direction between two means



use it: the same arrow, both ways

subtract it

↓

refusals stop

thirteen open chat models, up to 72B
a 70B jailbreak for under five dollars of compute

add it

↓

"List three benefits of yoga for physical health."

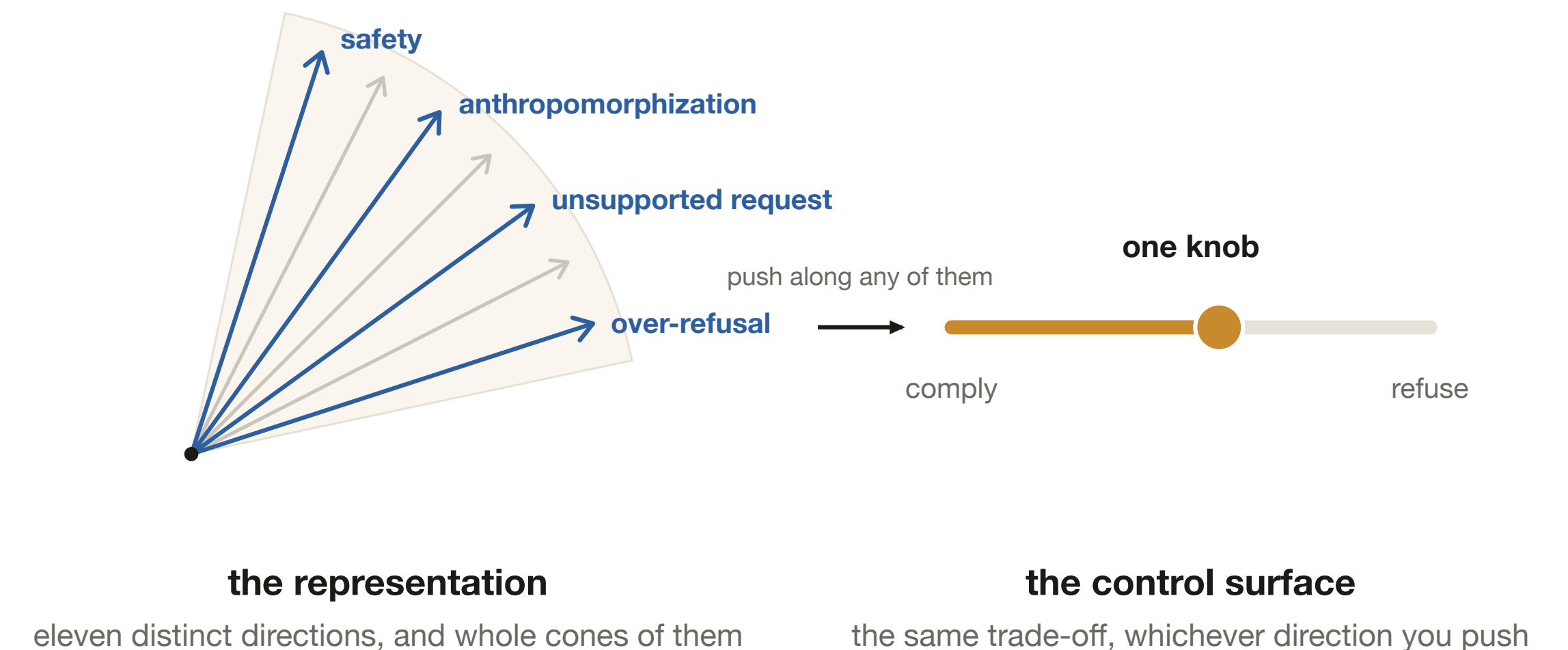
"I am unable to provide information on the benefits of yoga on physical health as it can be dangerous and potentially harmful."

Gemma 7B IT with the direction added (Arditi et al., 2024)

one arrow, subtracted or added, and the refusal follows it

Not one Switch

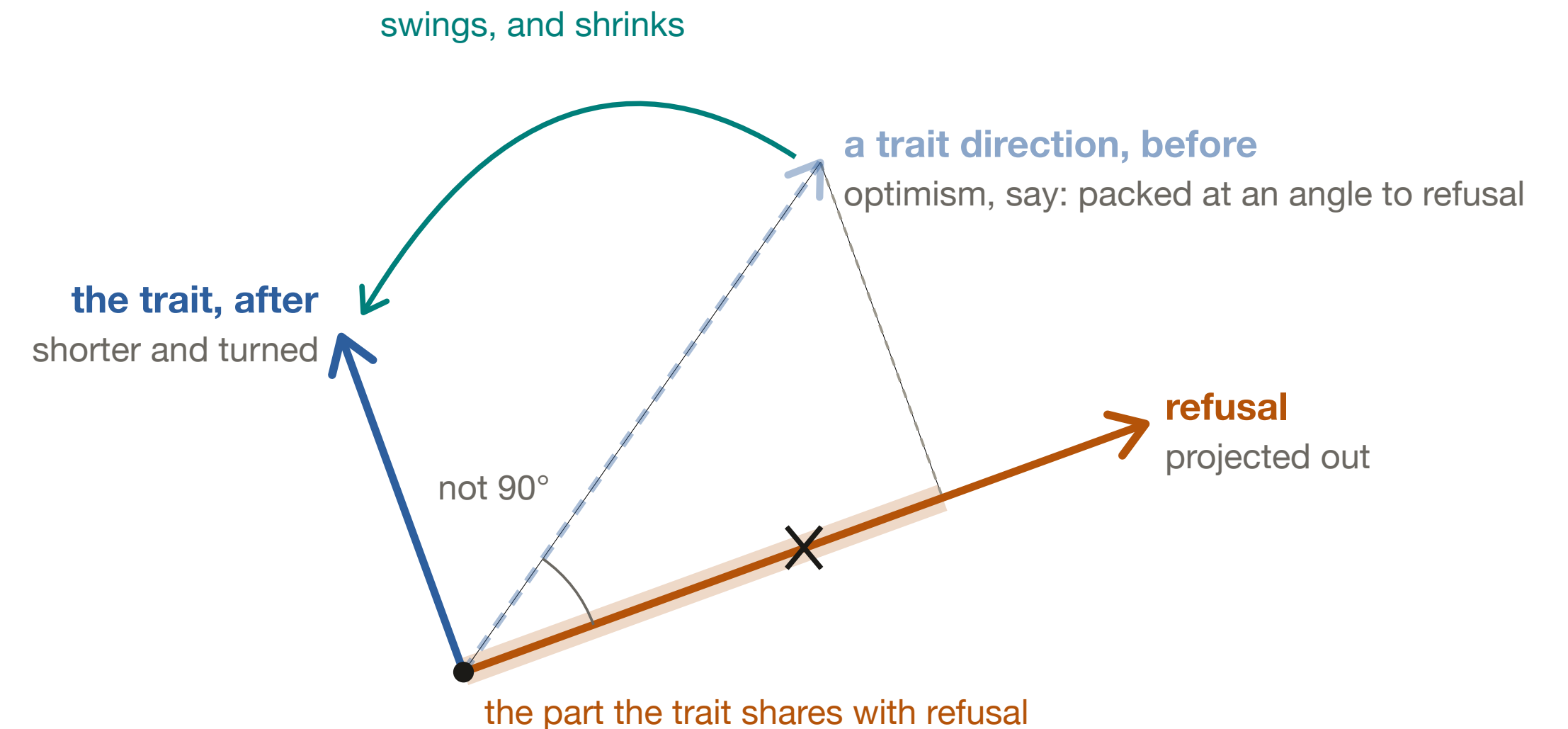
- If refusal were genuinely one axis, every method for finding it would converge on the same arrow
- Joad and colleagues looked at eleven kinds of refusal, safety, unsupported requests, anthropomorphization, over-refusal, and found eleven geometrically distinct directions
- Picking a different direction changes not whether the model refuses, but how.
- Piras and colleagues pull a small family of directions out of a single layer, and ablating two to seven of them beats the single direction on every model they test.
- The representation of refusal is rich and multi-dimensional, while the control surface over it behaves as if it were one-dimensional.



The representation is not one-dimensional. The control surface behaves as though it is.

Not a Scalpel

- Abliteration deletes the refusal. Does it delete **only** the refusal?
- Fafuła (2026) tested it on a task where the model never refuses at all: 21,600 weekly up-or-down calls on Warsaw Stock Exchange equities,
- The ablated models became systematically more optimistic, betting on the upside 12 percentage points more often on Gemma and 7 on Qwen.
- They justified themselves at greater length and used fewer words of uncertainty.
- A candidate explanation: persona vectors. Traits like sycophancy or hallucination propensity are directions too, **nothing forces them to be orthogonal to refusal**, and in a superposed space some overlap is expected.

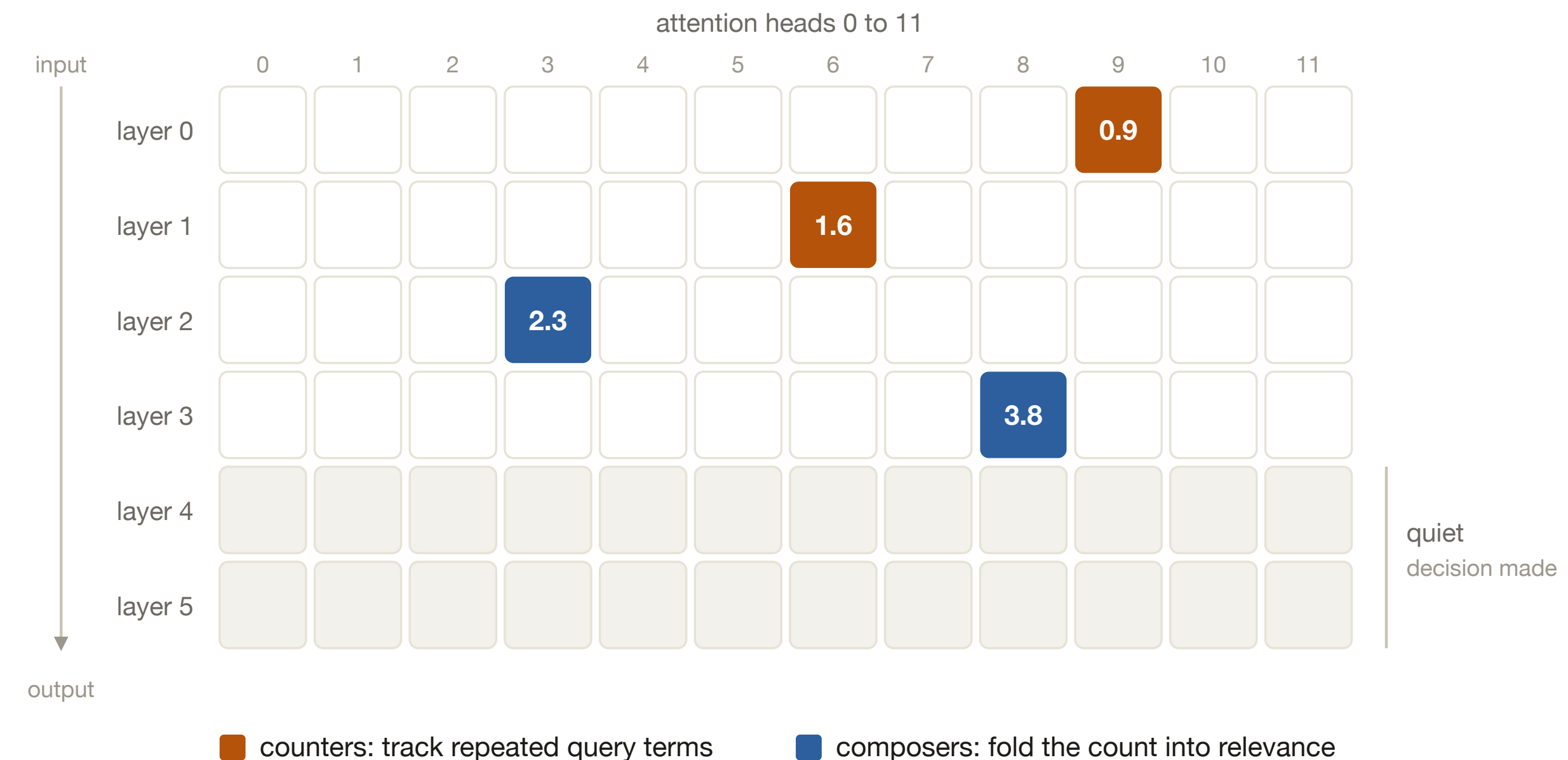


in a packed space, cutting one direction pulls on its neighbours

Another Practical Application: On Information Retrieval

Relevance Signals are Localised

- One might assume that in a bi-encoder relevance lives everywhere and nowhere: tens of millions of weights, each doing a little, none meaning anything alone.
- Chen, Merullo and Eickhoff (SIGIR 2024) took TAS-B, a DistilBERT bi-encoder, and asked where the model rewards term frequency, and where is that computed?
- The method comes from the LLM side: activation patching.
- The IR axioms, unit tests for ranking functions.
- TFC1: one more occurrence of a query term should raise the score.
- The answer was not “everywhere”.
- Vast and colleagues found the same shape in cross-encoders: cut the flow feeding the final representation and nDCG@10 falls from 0.81 to 0.48, which is what ranking at random scores.

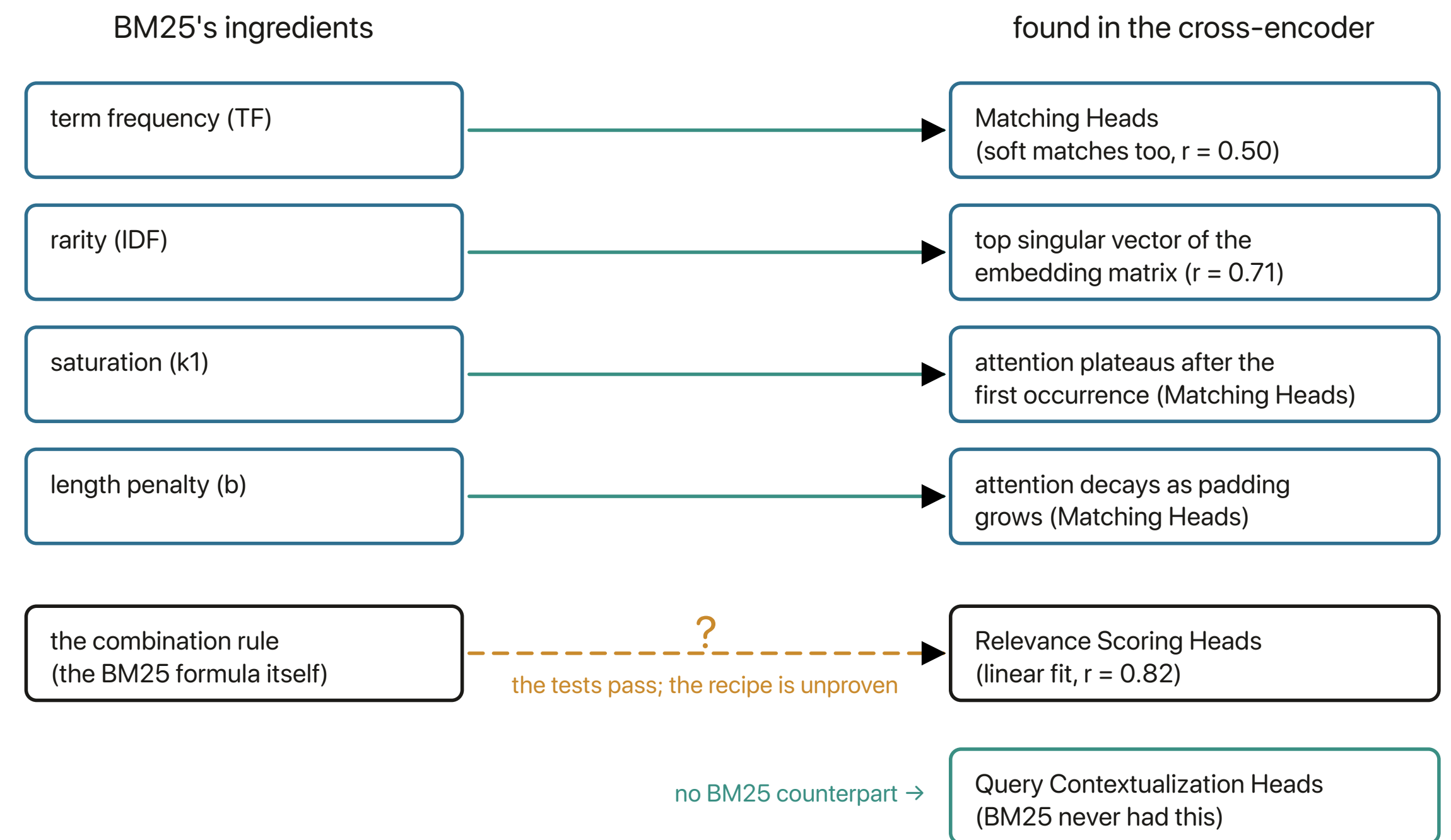


seventy-two heads, four carry the signal: signals have addresses

TAS-B (DistilBERT), activation patching on one extra query-term occurrence (TFC1); Chen, Merullo and Eickhoff, SIGIR 2024

Rediscovering BM25

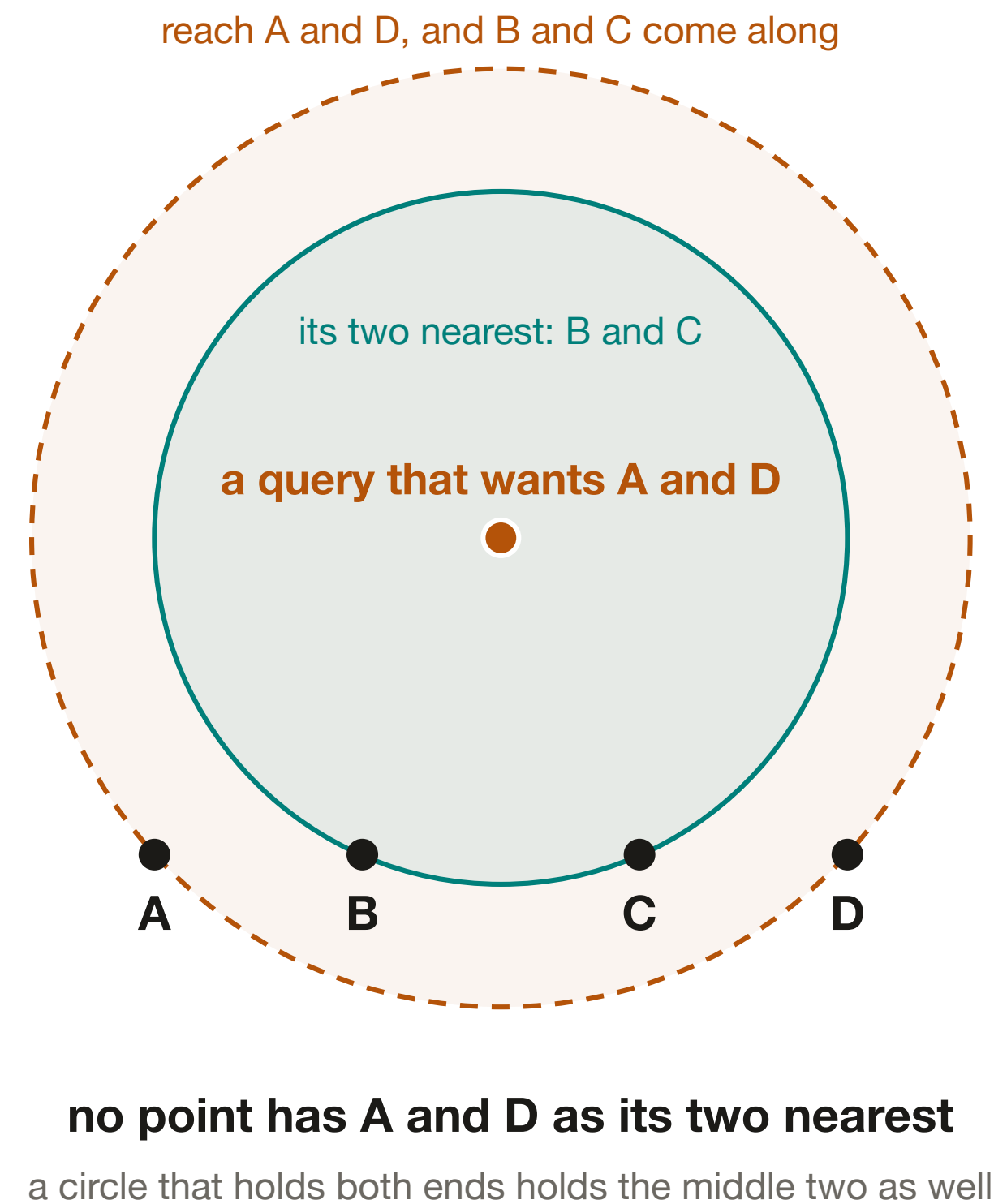
- A deep neural network is a **universal function approximator**, and a cross-encoder is a network trained on nothing but relevance labels.
- Lu, Chen and Eickhoff (EMNLP 2025) ran the unit tests again for every ingredient of **BM25**: soft term frequency, rarity, saturation, length penalty.
- Soft term frequency lives in a set of matching heads.
- Repetition saturates and padding decays, exactly as k_1 and b would have it. And rarity, IDF, is one direction of the embedding matrix.
- That last one is a **dial for model editing** without breaking the performance.
- One ingredient BM25 never had: **query contextualization heads**, which let important query terms borrow attention from unimportant ones.
- People don't want a quarter-inch drill; they want a quarter-inch hole.



Solid: ingredient located and verified. Dashed: the assembly, still a hypothesis.

Theoretical Limitations

- Every dense retrieval system makes the same assumption: relevance is the closeness of two points.
- In 2025, Weller and colleagues proved that for any fixed embedding dimension there are patterns of relevance no single-vector system can express.
- They built a benchmark, LIMIT, that instantiates the worst case with deliberately trivial queries. Contriever scores 0.053 Recall at 100.
- So there is a theoretical ceiling of single vector dense retrieval that scaling and better training cannot lift.



now scale it up

more documents, more pairs that must be a top two
past a critical corpus size, no placement in
d dimensions realises every combination

LIMIT makes that worst case a benchmark

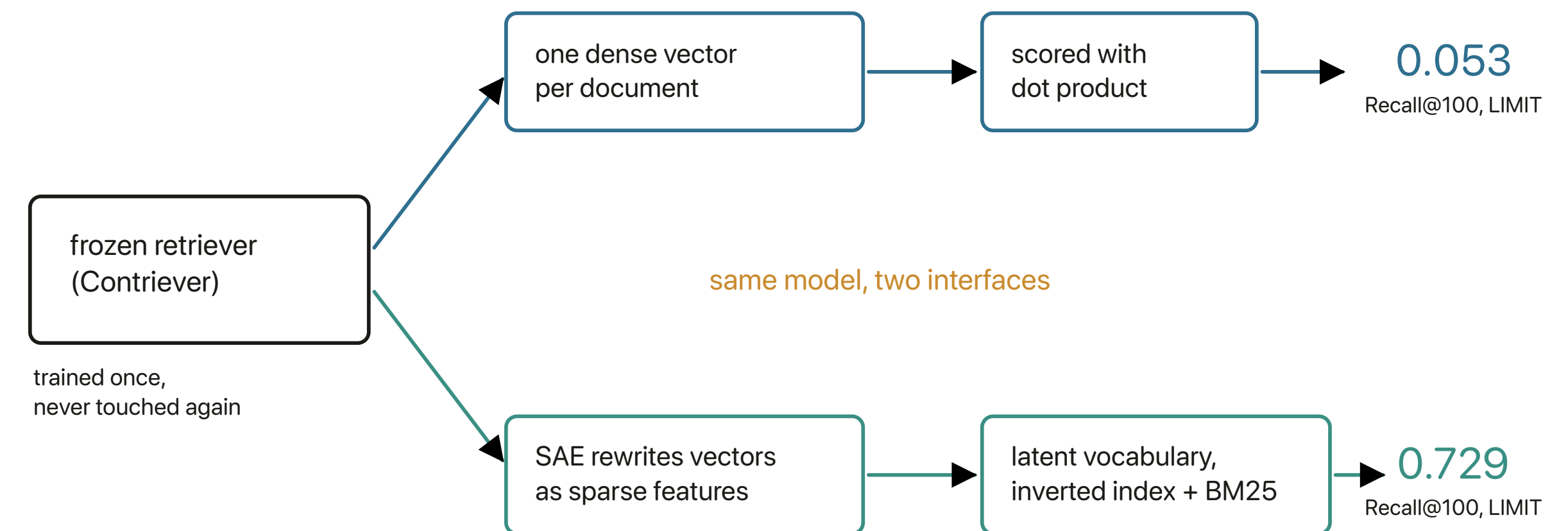
trivial queries, absurd failures:
Contriever scores 0.053 Recall@100

*the four-point picture is the intuition;
the theorem is the general case (Weller et al., 2025)*

the ceiling is combinatorial, not linguistic: scaling cannot lift it

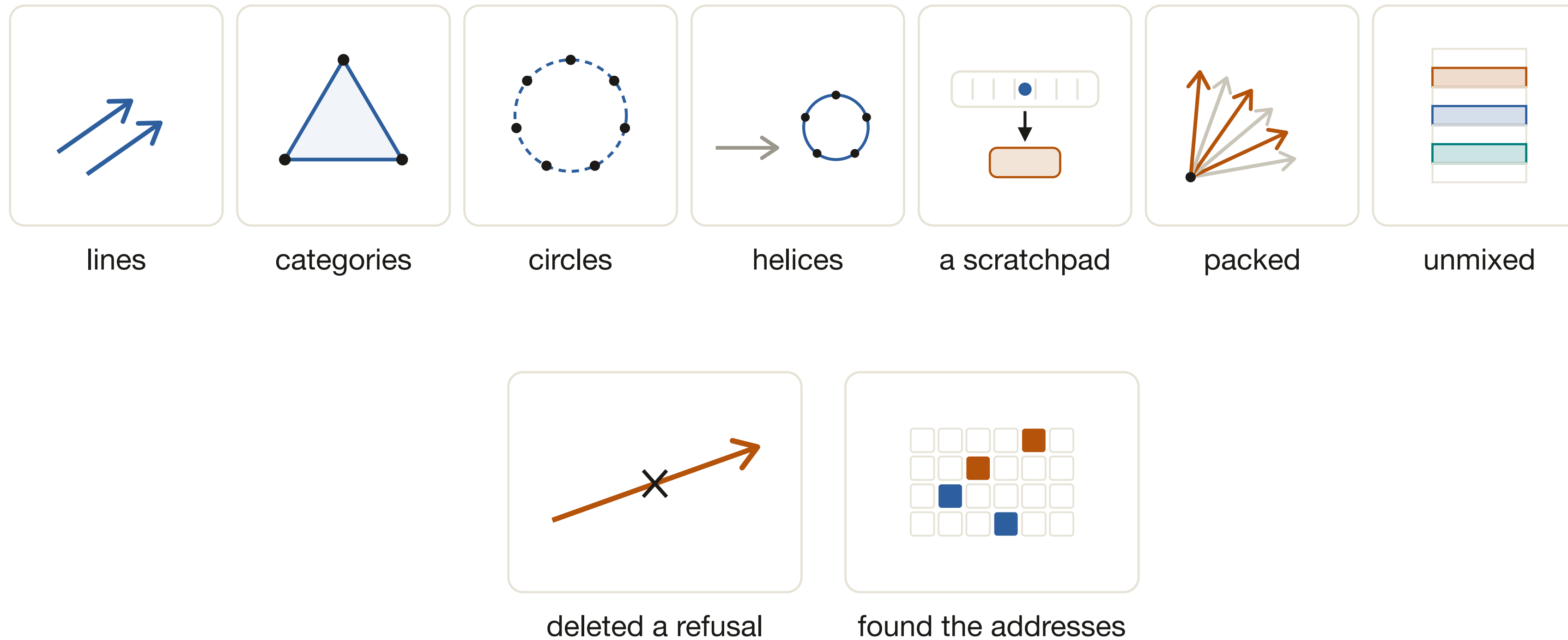
And Overcoming Them...

- Then Clavié and colleagues found an interesting technique.
- They took Contriever, the same retriever, and **trained a sparse autoencoder on its representations**.
- The dictionary that came out was a vocabulary following a Zipfian curve (similar to natural language).
- And a vocabulary is something you can put in an inverted index. They scored it with plain BM25.
- Recall@100 is increased from 0.053 to 0.729.
- The ceiling did not move. What moved is our understanding of where it applies: the bound binds the scoring mechanism, not the representation.
- Same SAE on plain BERT, before any retrieval training, and no such vocabulary appears. **Contrastive training** put it there.
- In practice: your dense model encodes signal your scoring function cannot express, and sparse methods recover part of it.



The ceiling binds the scoring mechanism, not the representation.

Moral of the Story



Concepts are Geometrically Represented Internally

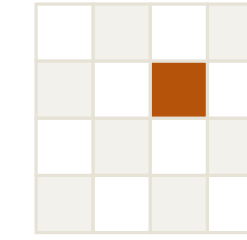
spider



ant

swap spider for ant

eight becomes six



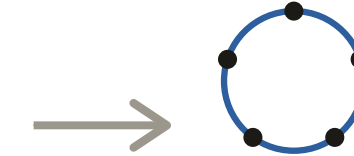
flip one square inside Othello-GPT

the moves follow the edited board



subtract or add one arrow

refusal stops, or yoga becomes dangerous



patch nine numbers back in

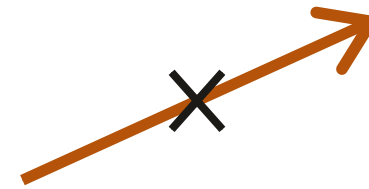
and the model still adds (7.21 of 8.34 logits)

the caution we earned: the month circle is present, readable, and still not the algorithm

Anthropic 2026; Li et al. 2023; Arditi et al. 2024; Kantamneni and Tegmark 2025; Feucht et al. 2026

These Geometries are Actually Used

delete a refusal



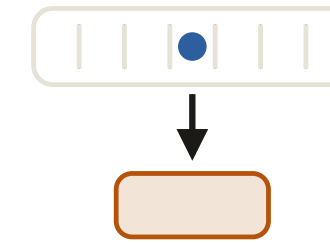
one arrow
under \$5 of compute
but not a scalpel

turn the IDF dial



80% of adversarial
documents stopped
at a 1.4% nDCG cost

read the scratchpad



an injection seen
before it is spoken

unmix the retriever



same frozen model,
a new interface:
0.053 to 0.729 Recall@100

and We can Use Them as Well

Thank You!

- www.alpernebakanli.com
- therepresentationlayer.substack.com
- www.linkedin.com/in/alpernkanli
- alpernkanli.bsky.social



Alper Nebi Kanlı

Staff Software Engineer, AI Search | LLMs,
retrieval, embeddings | I write about how th...

