The Emergence of Convergence in Different Levels of Biology and AI

Brian Cheung

cheungb@mit.edu

Minyoung Huh (Jacob)

Tongzhou Wang

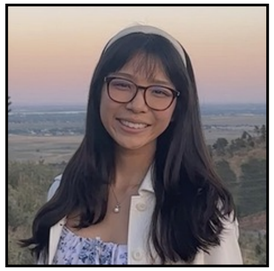
Phillip Isola

REDWOOD CENTER

for Theoretical Neuroscience

Erin Grant

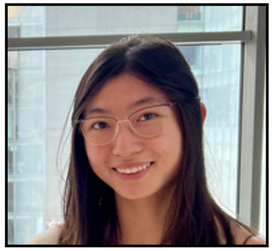
Sophie Wang



Yena Han

Eric Weiss

Helen Yang



Kushagra Tiwary

Aaron Young

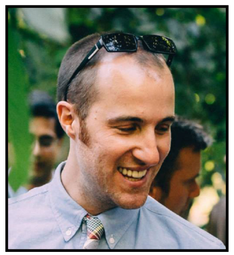
Pulkit Agrawal

Bruno Olshausen

Boris Katz

Tomaso Poggio

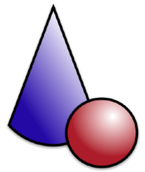
Jascha Sohl-Dickstein



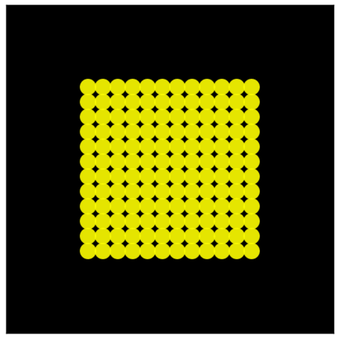
CENTER FOR
Brains
Minds+
Machines

Google DeepMind

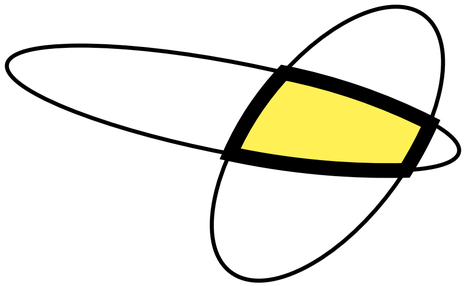
Observing convergence across many forms of intelligence



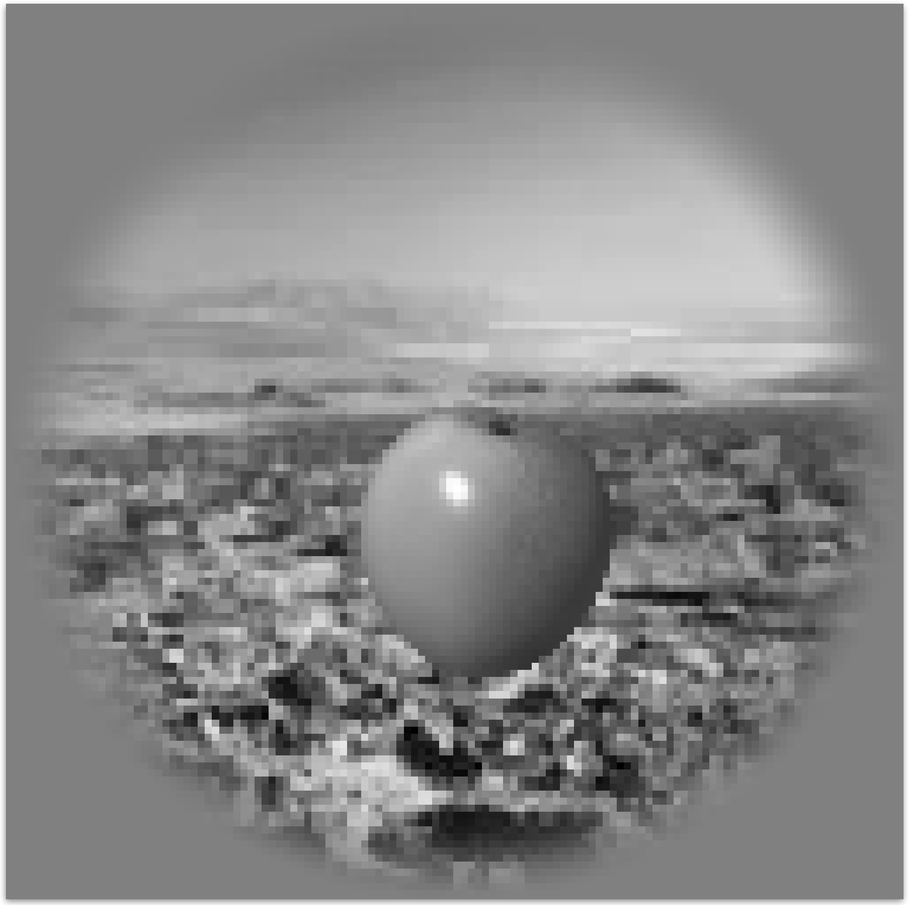
Controlling convergence to explain biology

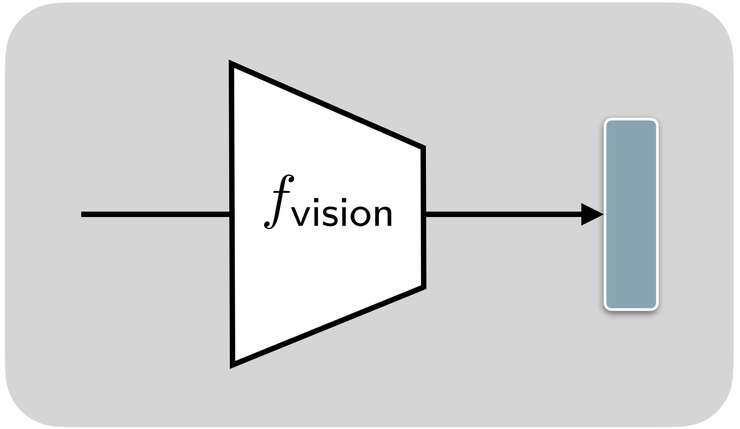


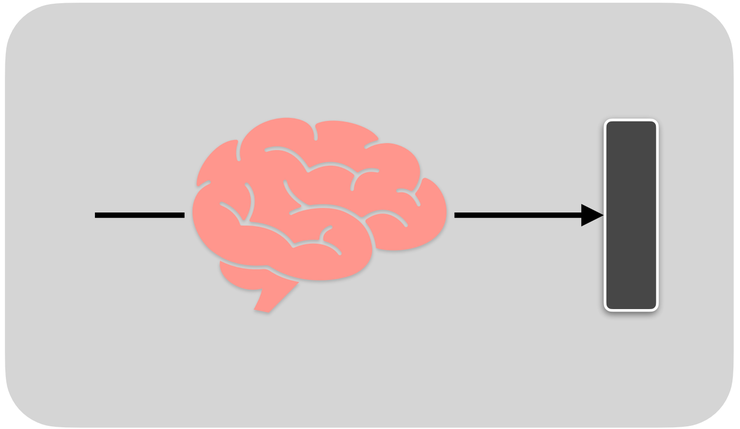
Future Work: Optimizing for convergence



Representational similarities between Brains and Machines



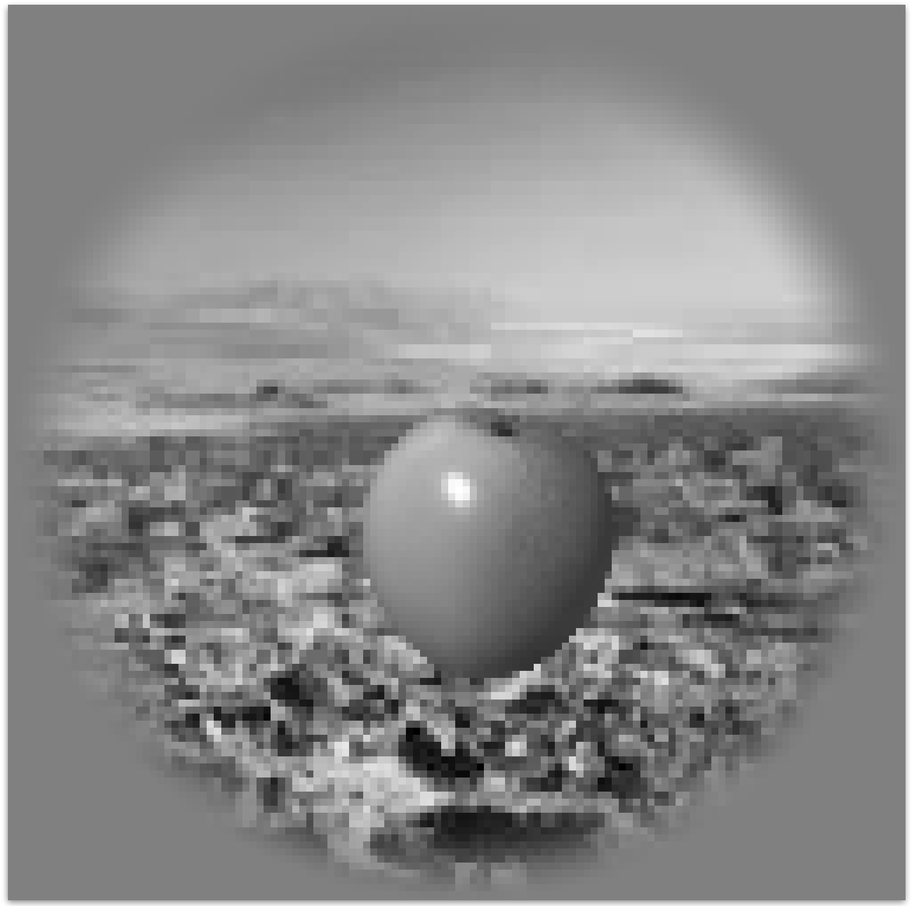


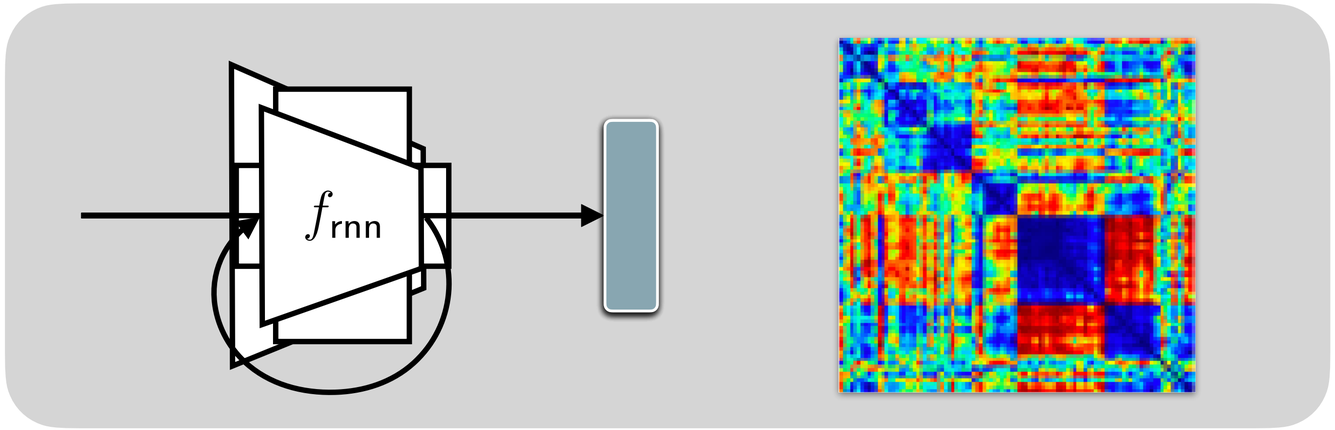


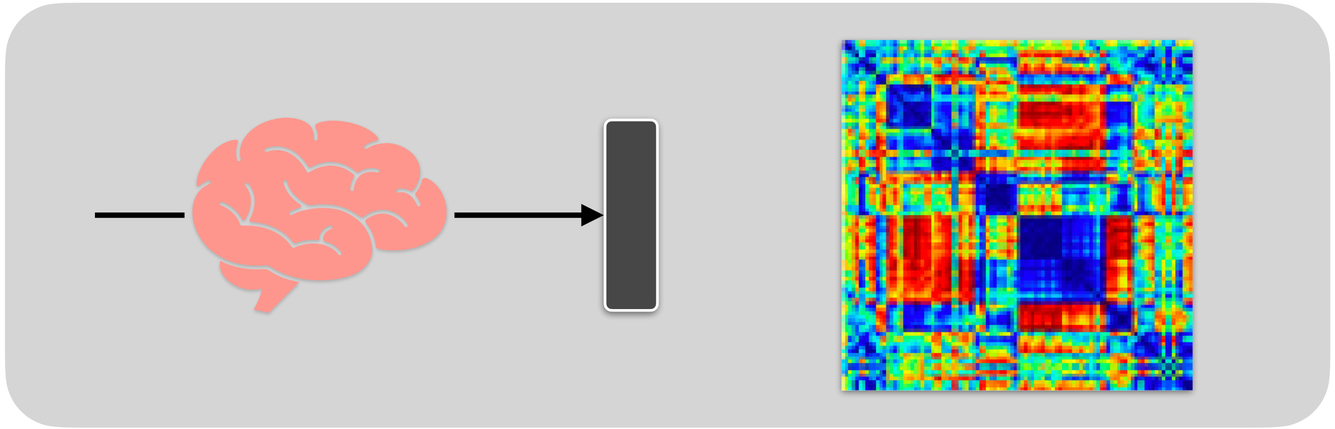
[Han, Poggio, **Cheung** 2023]

[Yamins, Hong, Cadieu, Solomon, Seibert, DiCarlo 2014]

Representational similarities between Brains and Machines



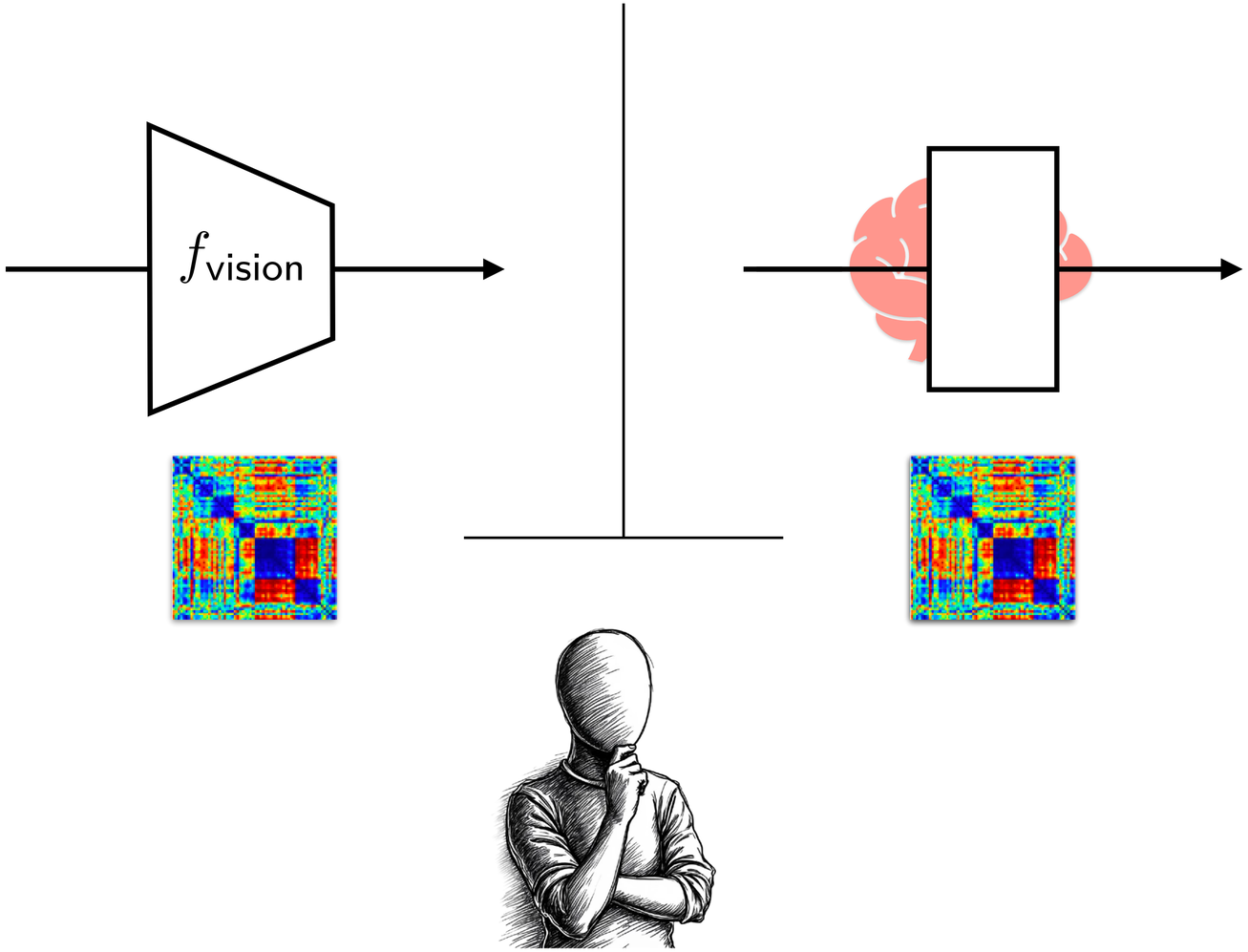




[Han, Poggio, **Cheung** 2023]

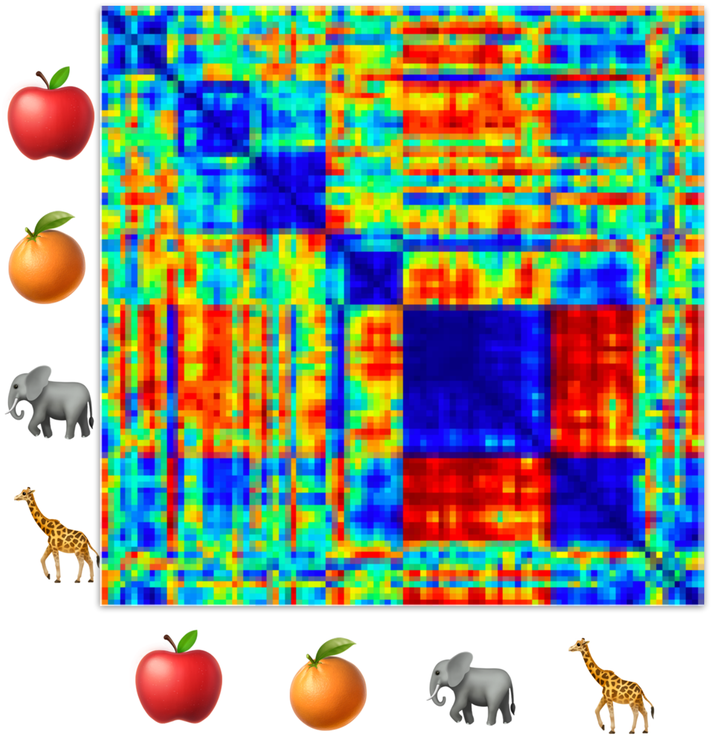
[Yamins, Hong, Cadieu, Solomon, Seibert, DiCarlo 2014]

Representational Turing Test



Characterizing representations using kernels

$\mathbf{K}_{\text{vision}}$



Restrict our attention to **vector embeddings**

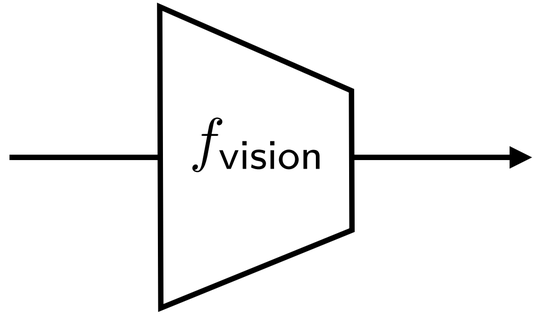
$$f : \mathcal{X} \to \mathbb{R}^n$$

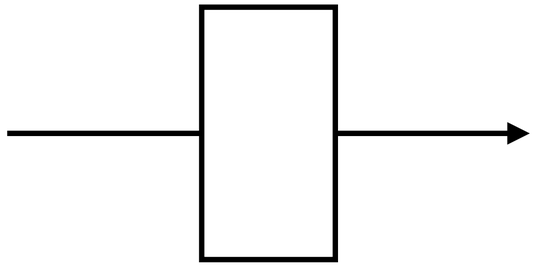
Characterize a representation in terms of its **kernel**

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

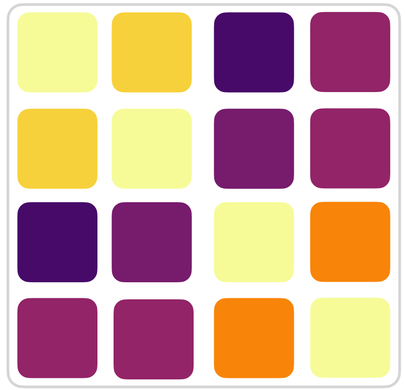
$$K(x_i, x_j) = \langle\ f(🍎), f(🍊)\ \rangle$$

Do models measure similarity in **similar** ways?

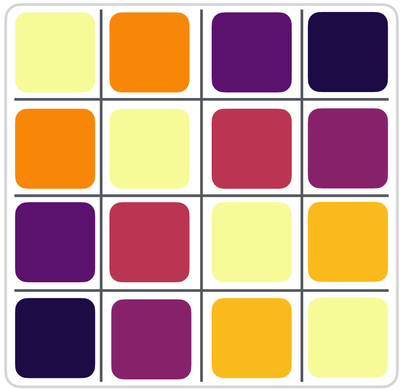




sim(

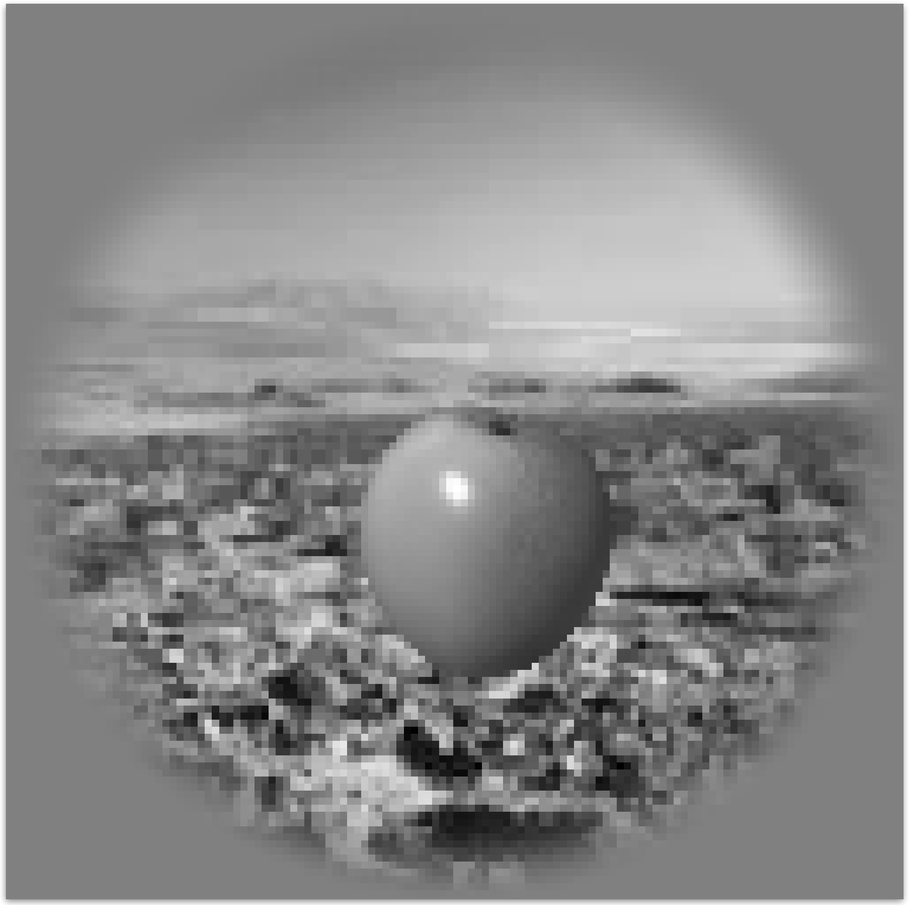


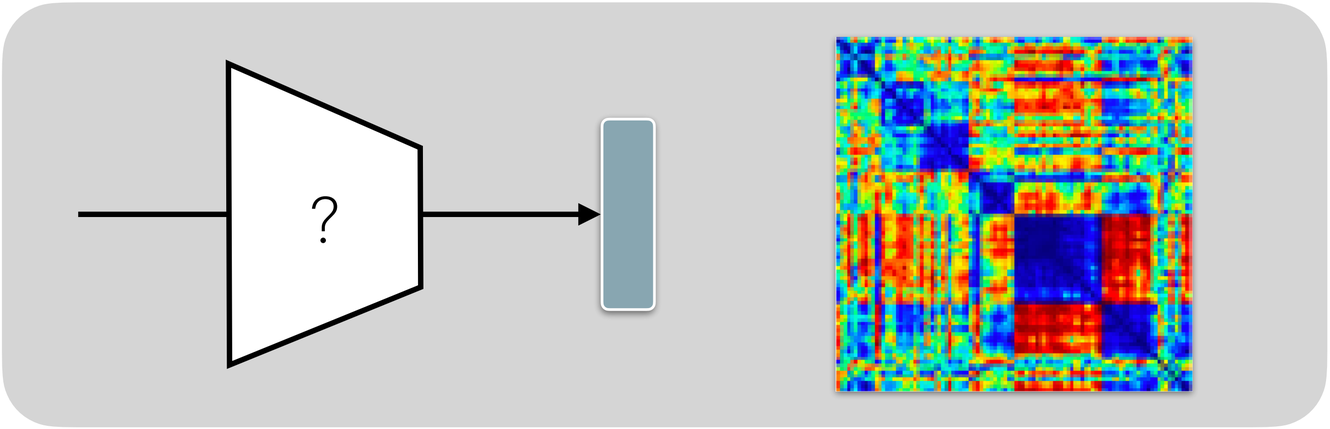
,

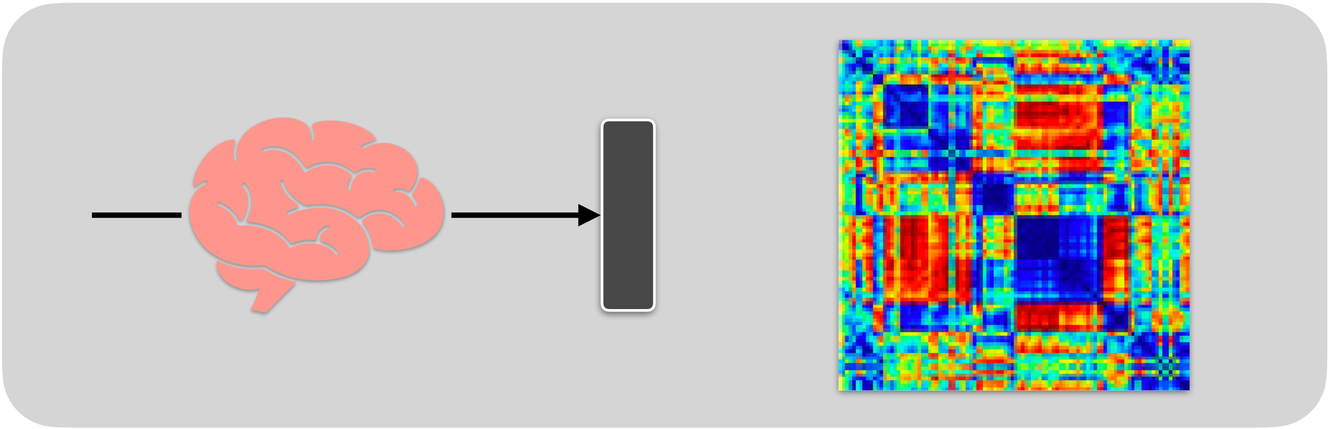


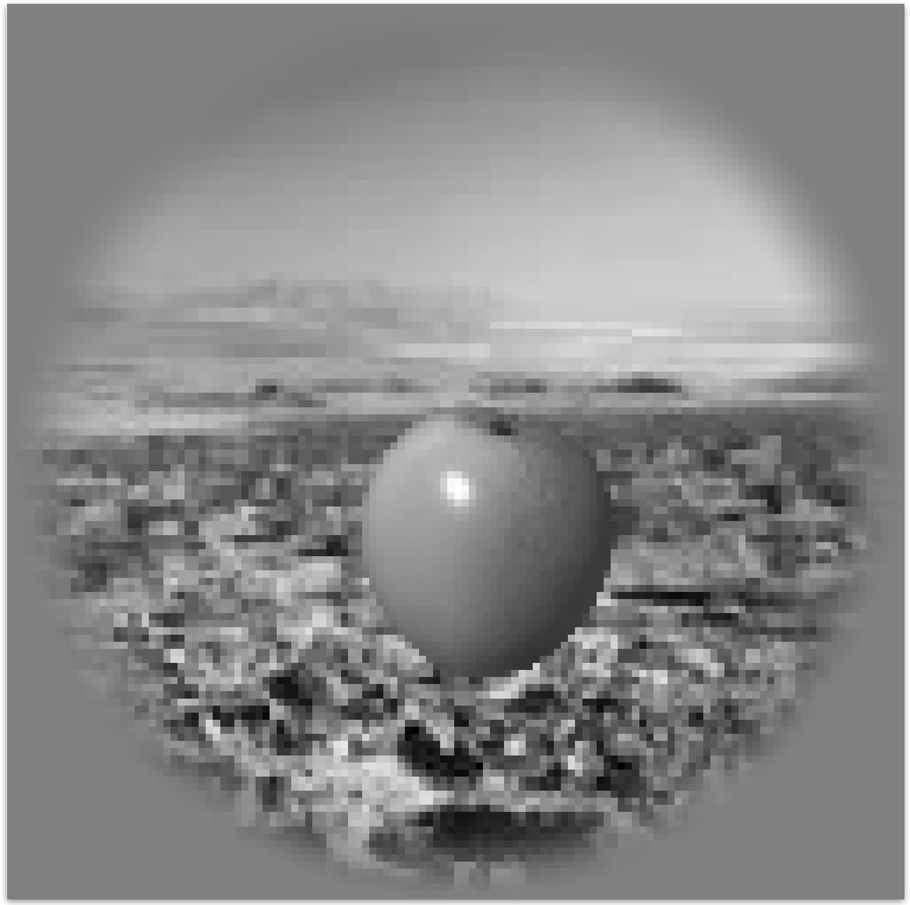
)

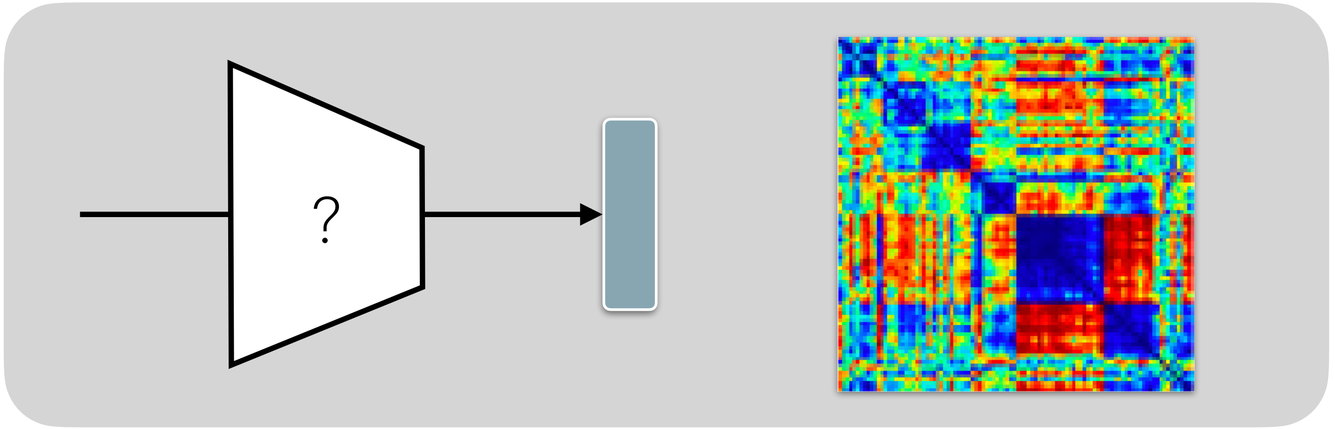
Representational Turing Test

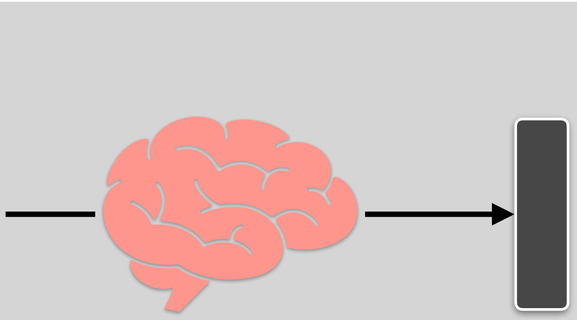


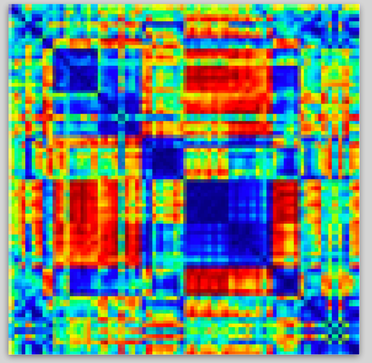


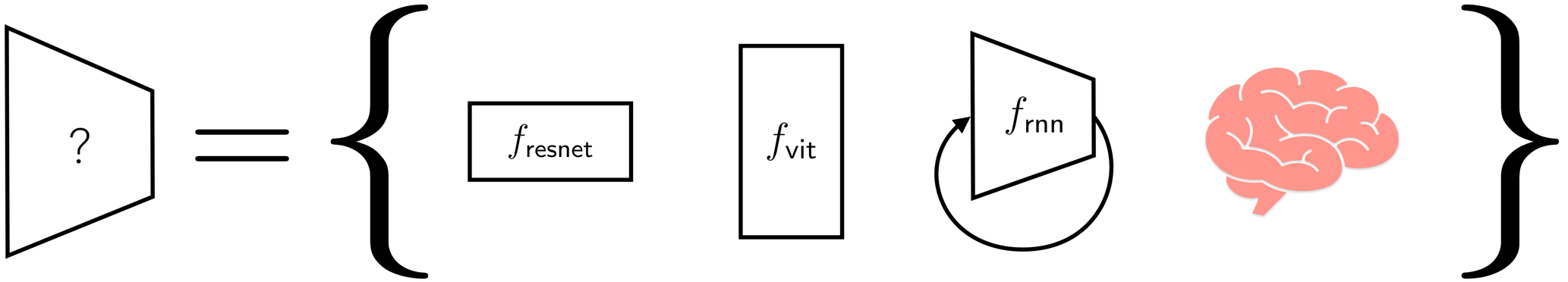


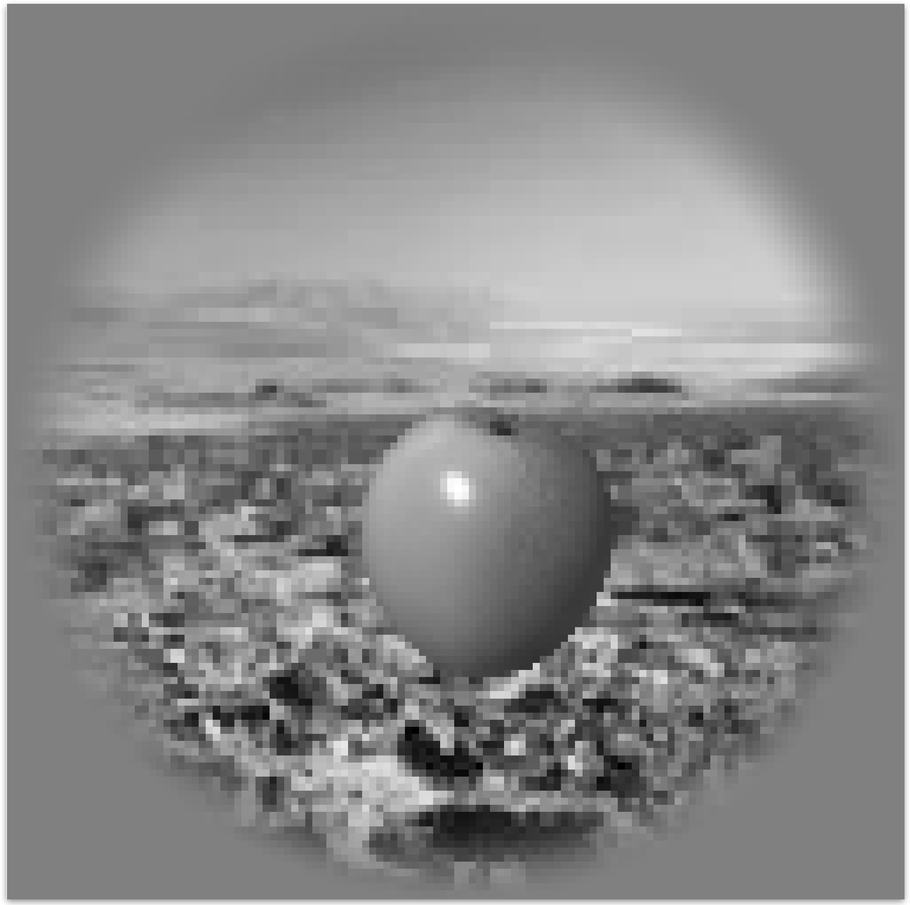


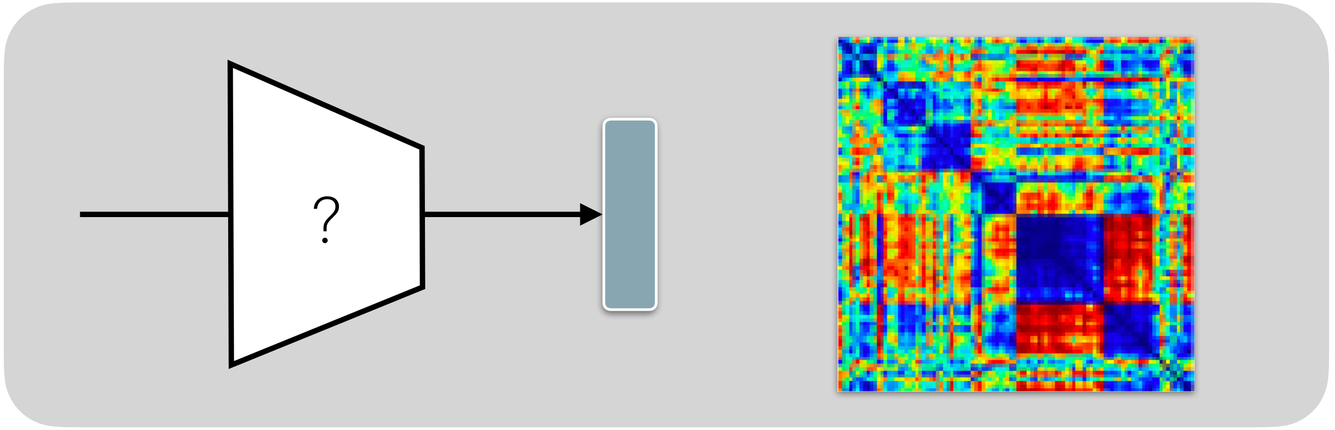

?

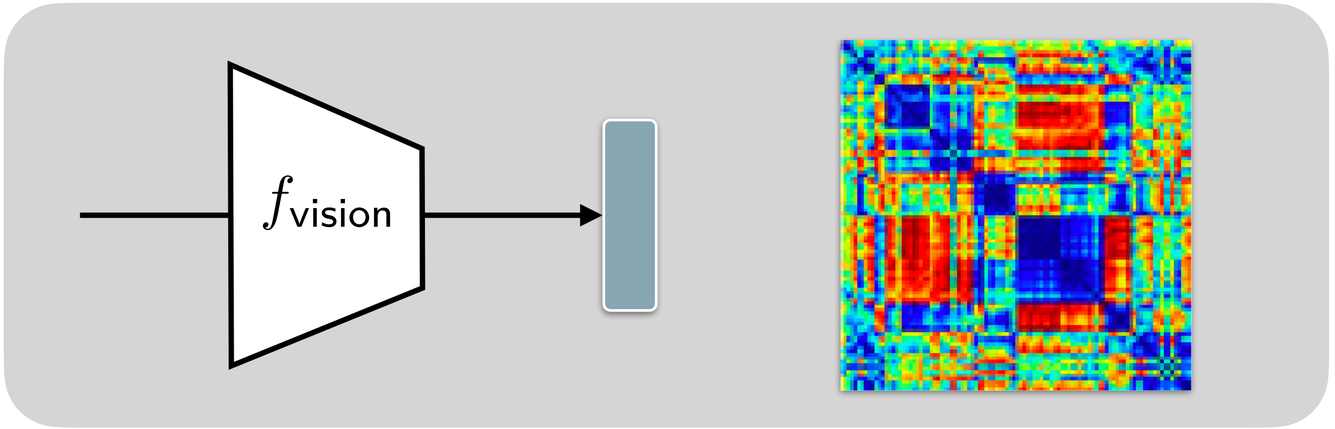


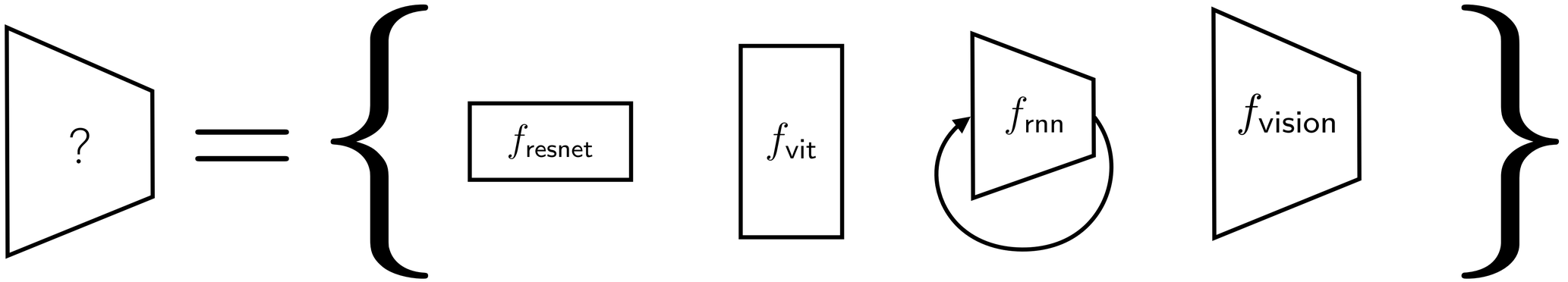



?
f_{resnet}
f_{vit}
f_{rnn}



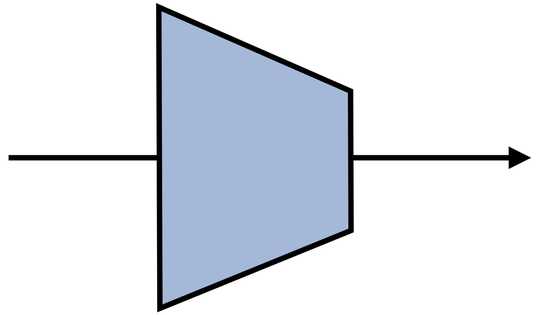
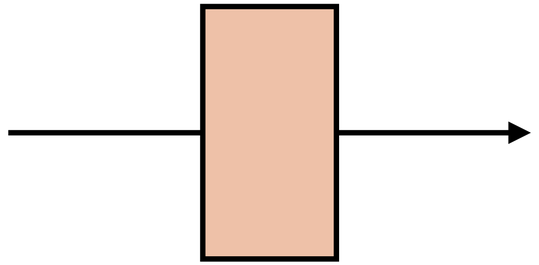

?


f_{vision}

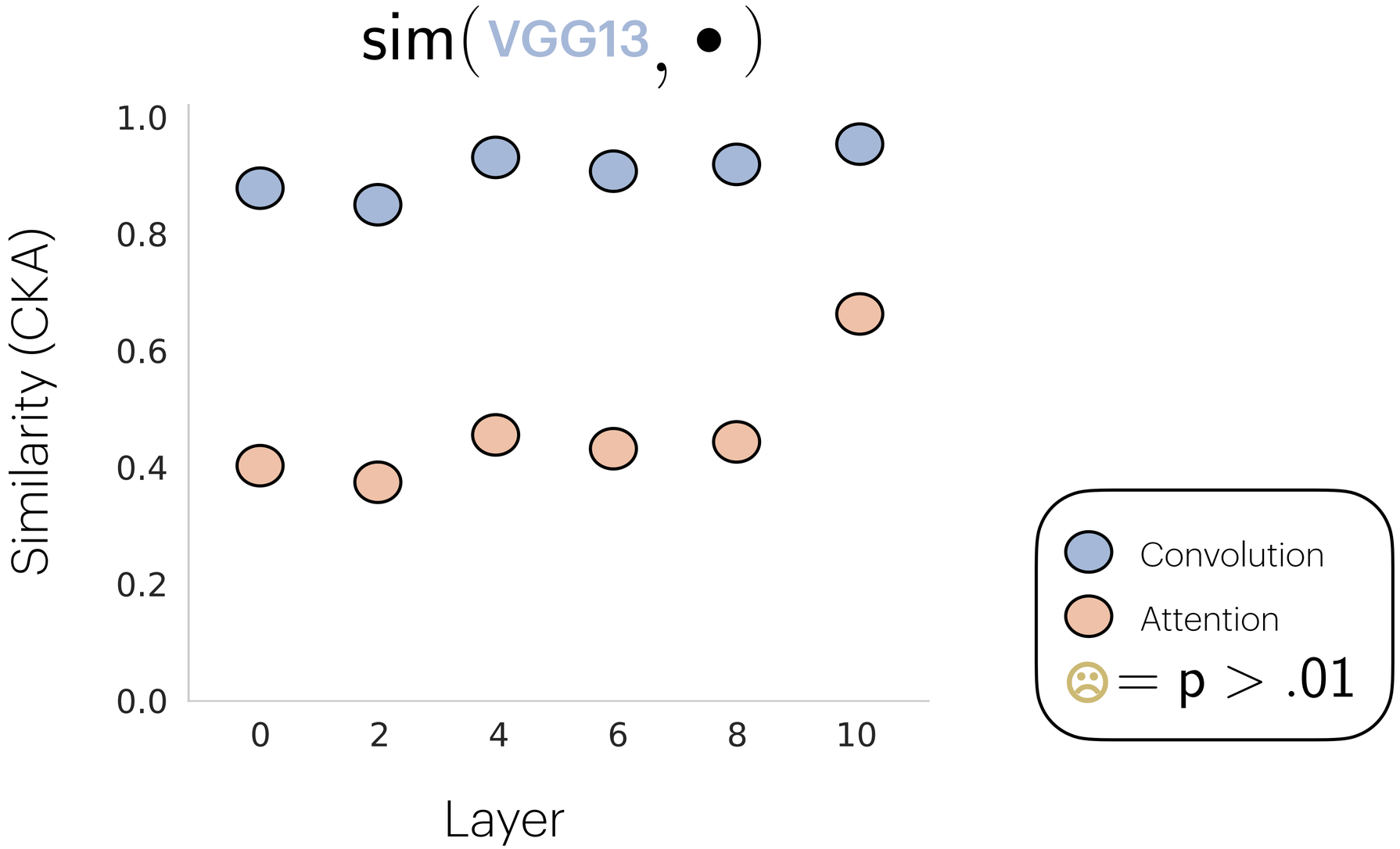

?
f_{resnet}
f_{vit}
f_{rnn}
f_{vision}

Convolution

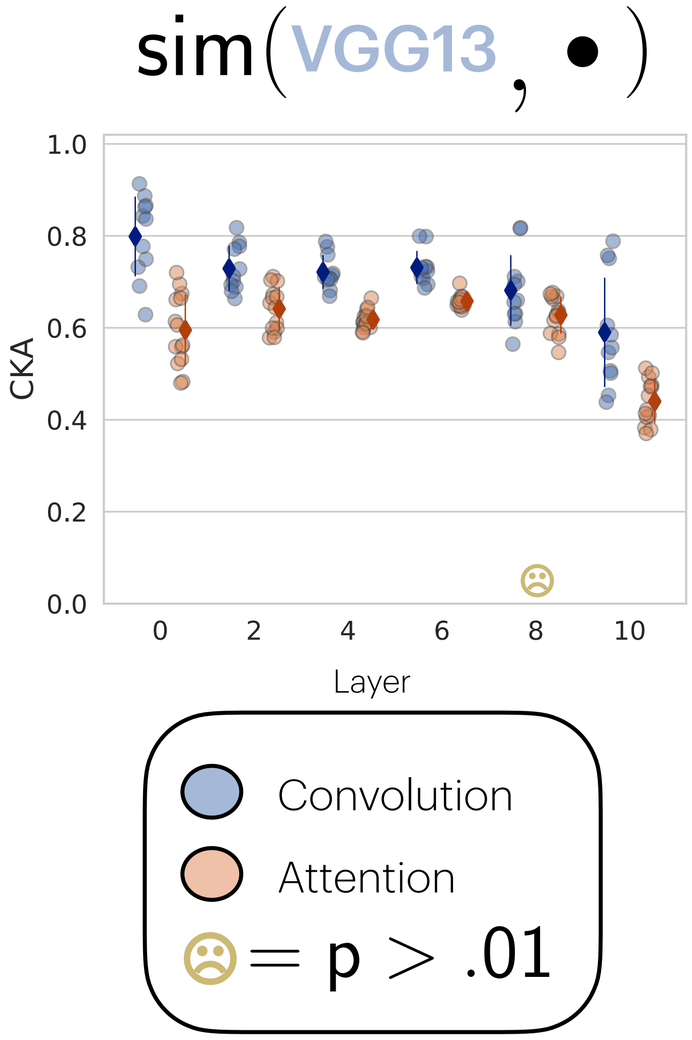
Attention

sim( , )

Easy to tell Convolution apart from Attention



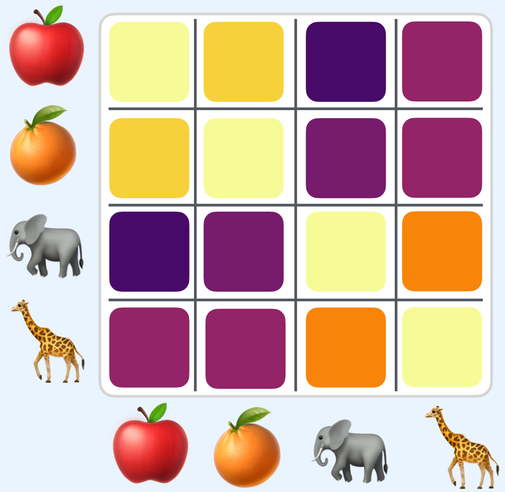
Easy to tell Convolution apart from Attention?



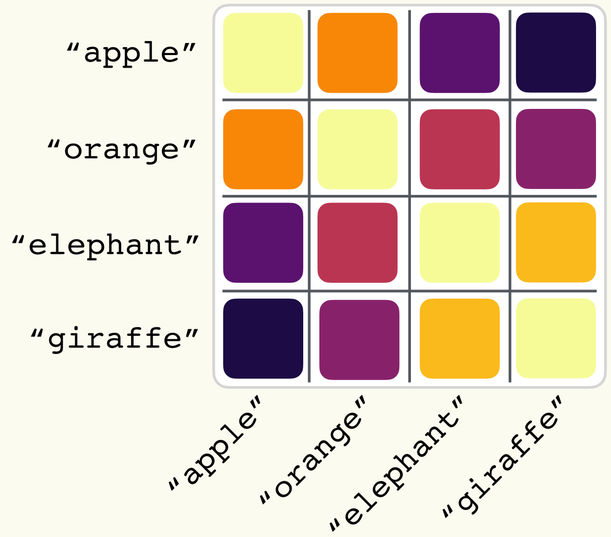
Do **different** models represent the world in **different** ways?
Or are they somehow **all becoming alike**?

Do models measure similarity in **similar** ways?

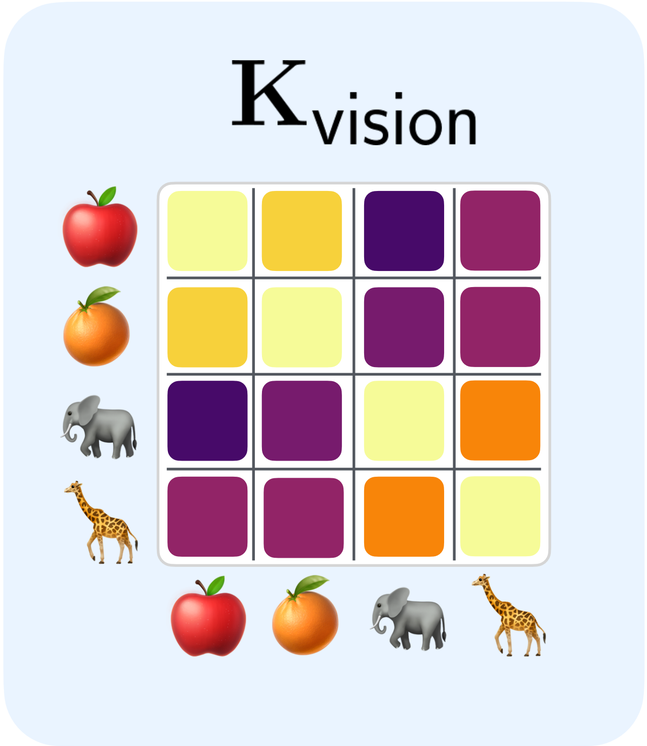
$\mathbf{K}_{\text{vision}}$

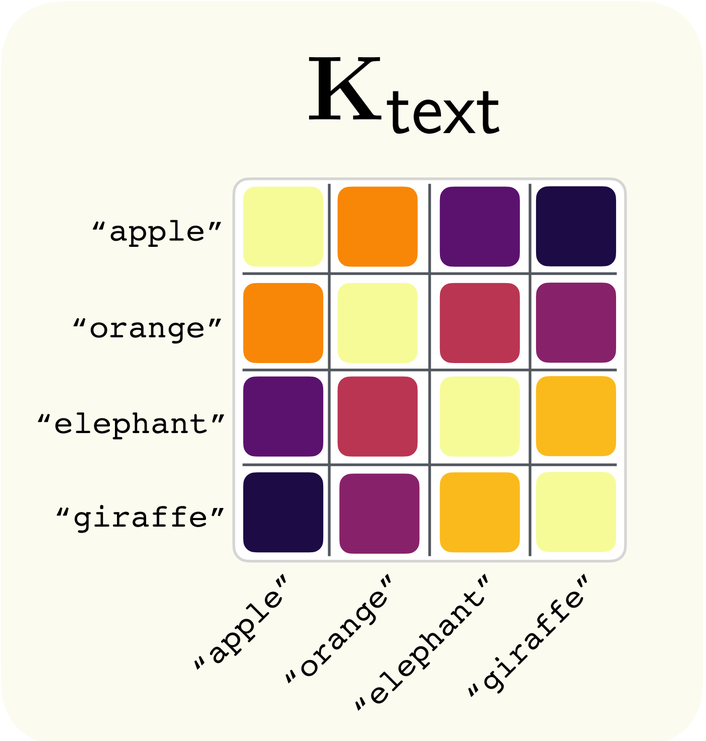


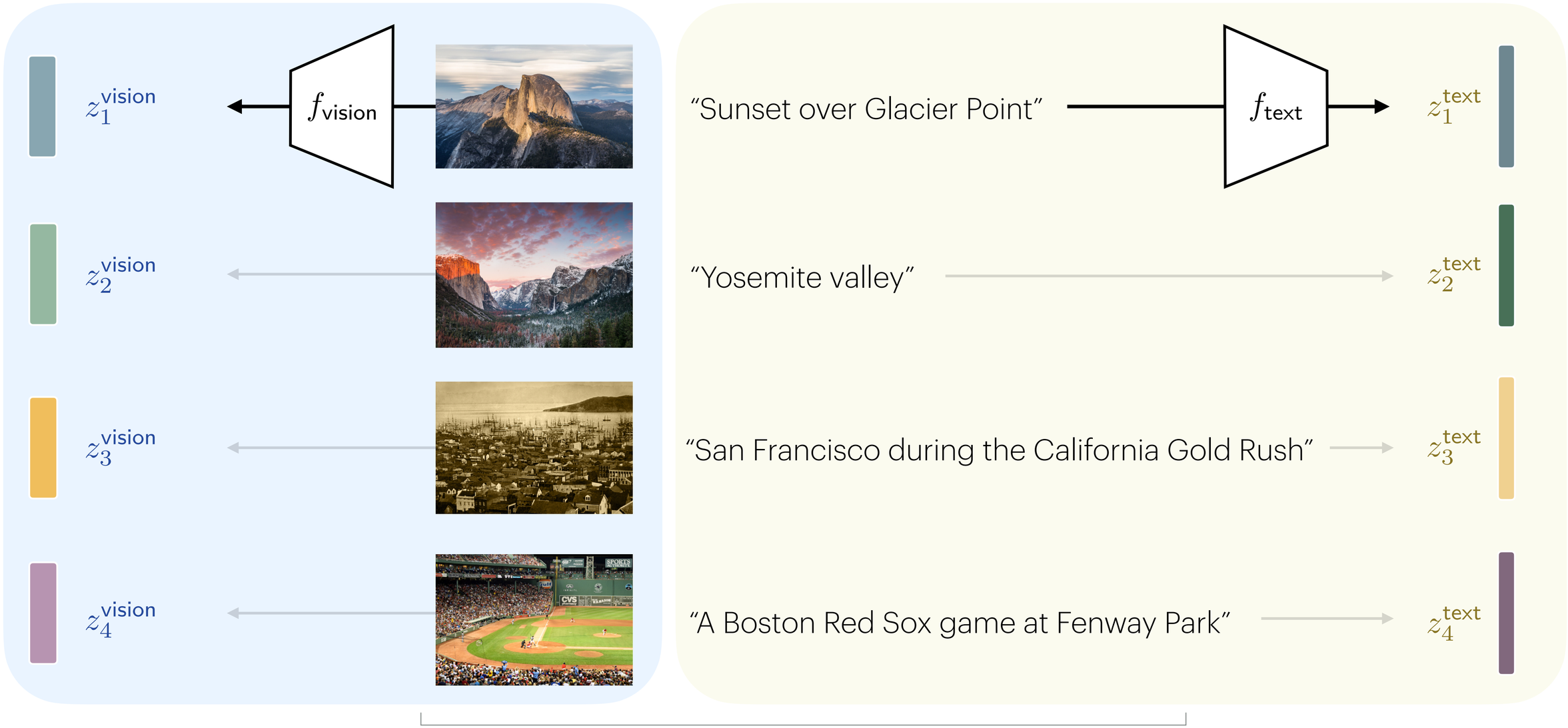
$\mathbf{K}_{\text{text}}$



$$\text{sim}\Big(\mathbf{K}_{\text{vision}}, \mathbf{K}_{\text{text}}\Big)$$



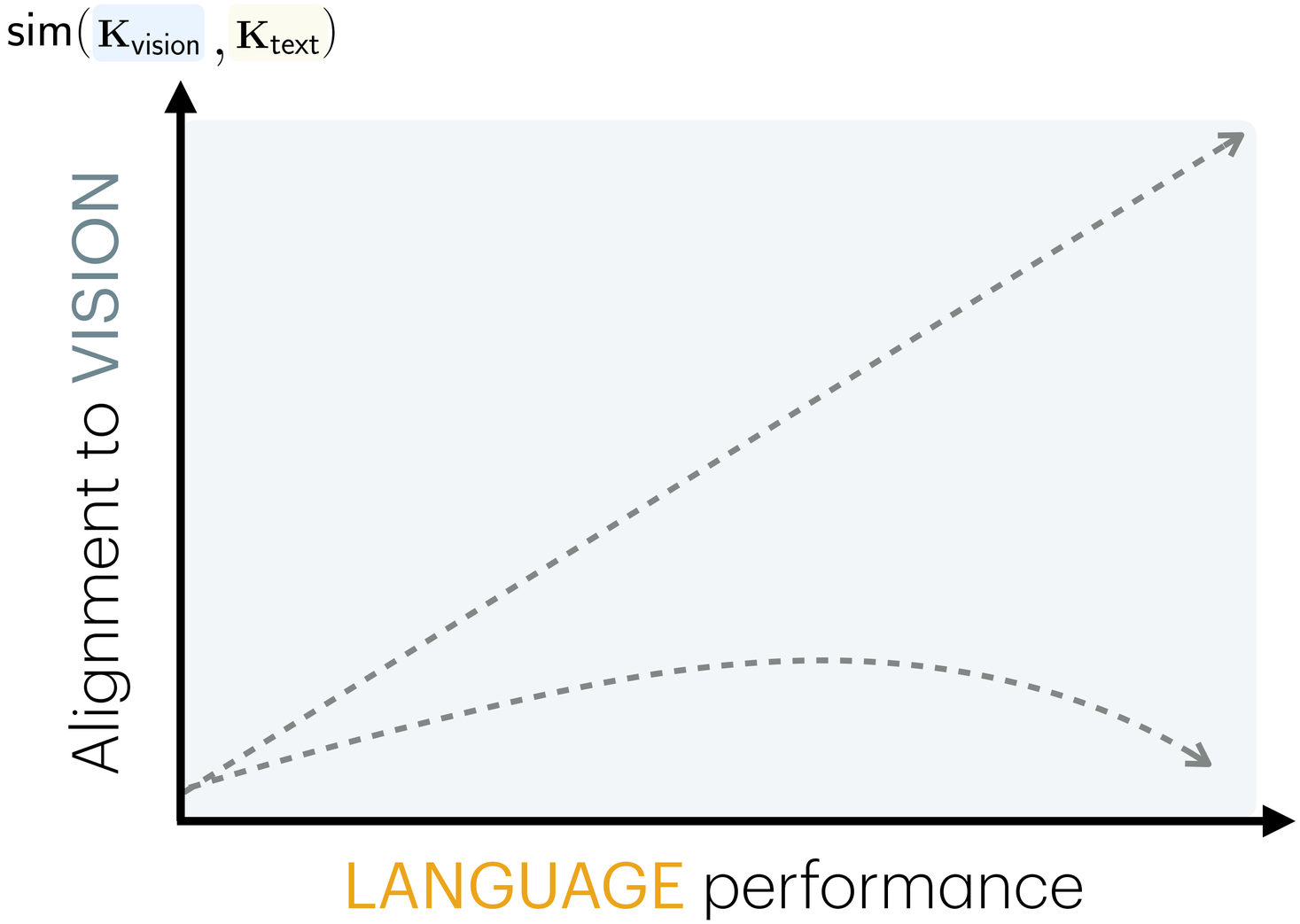




Wikipedia Image Text Dataset

[Srinivasan, Raman, Chen, Bendersky, Najork 2021]

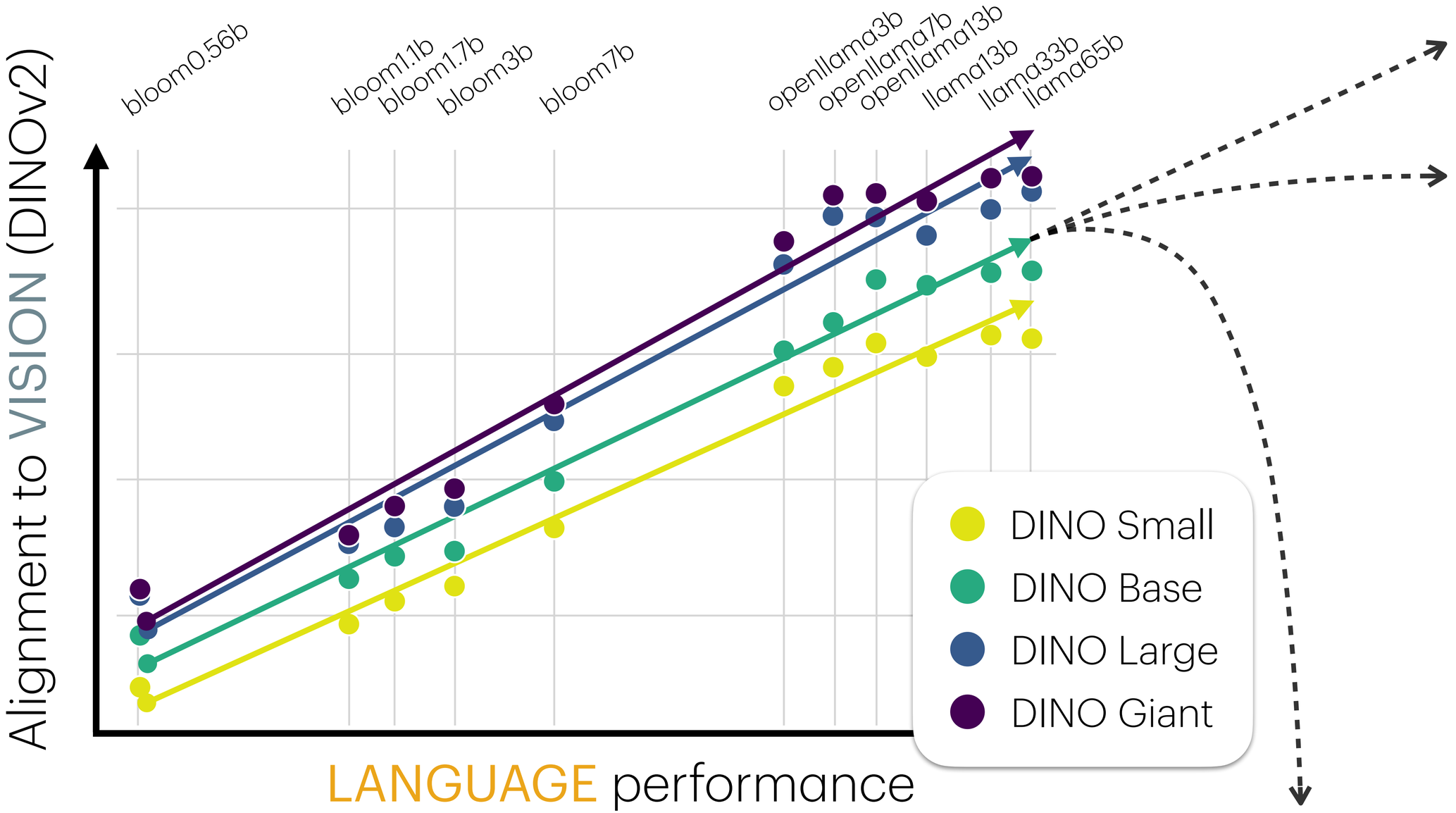
Do strong language models align better with vision models?



Hypothesis B:
Better language models are better vision models.

Hypothesis A:
Better language models overfit to language.

Strong models converge in representation



There's a large space of work that has shown similar results supporting our hypothesis

Published as a conference paper at ICLR 2023

RELATIVE REPRESENTATIONS ENABLE ZERO-SHOT LATENT SPACE COMMUNICATION

Luca Moschella[1,*] Valentino Maiorca[1,*]

Marco Fumero

[1]Sa

Neural netwo
high-dimensi
data points in
and other arc
weights initia
in the training
reuse. Nevert
ing choices, t
change. In th
a fixed set of
can enforce th
neural archite
practice, inva
space commu
between dive
of our approa
graphs), tasks
GCNs, transf

NCE

OPEN ACCESS

Check for updates

ometry u

La Jolla, CA; receiv

Explanatory models in neuroscience:
Constraint-based intelligibility

Rosa Cao and Daniel Yamins*

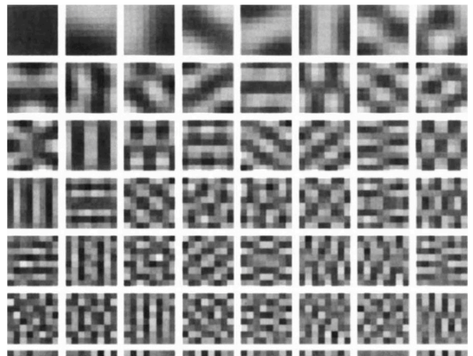
Abstract

LETTERS TO NATURE

Emergence of simple-cell receptive field properties by learning a sparse code for natural images

Bruno A. Olshausen* & David J. Field

Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853, USA



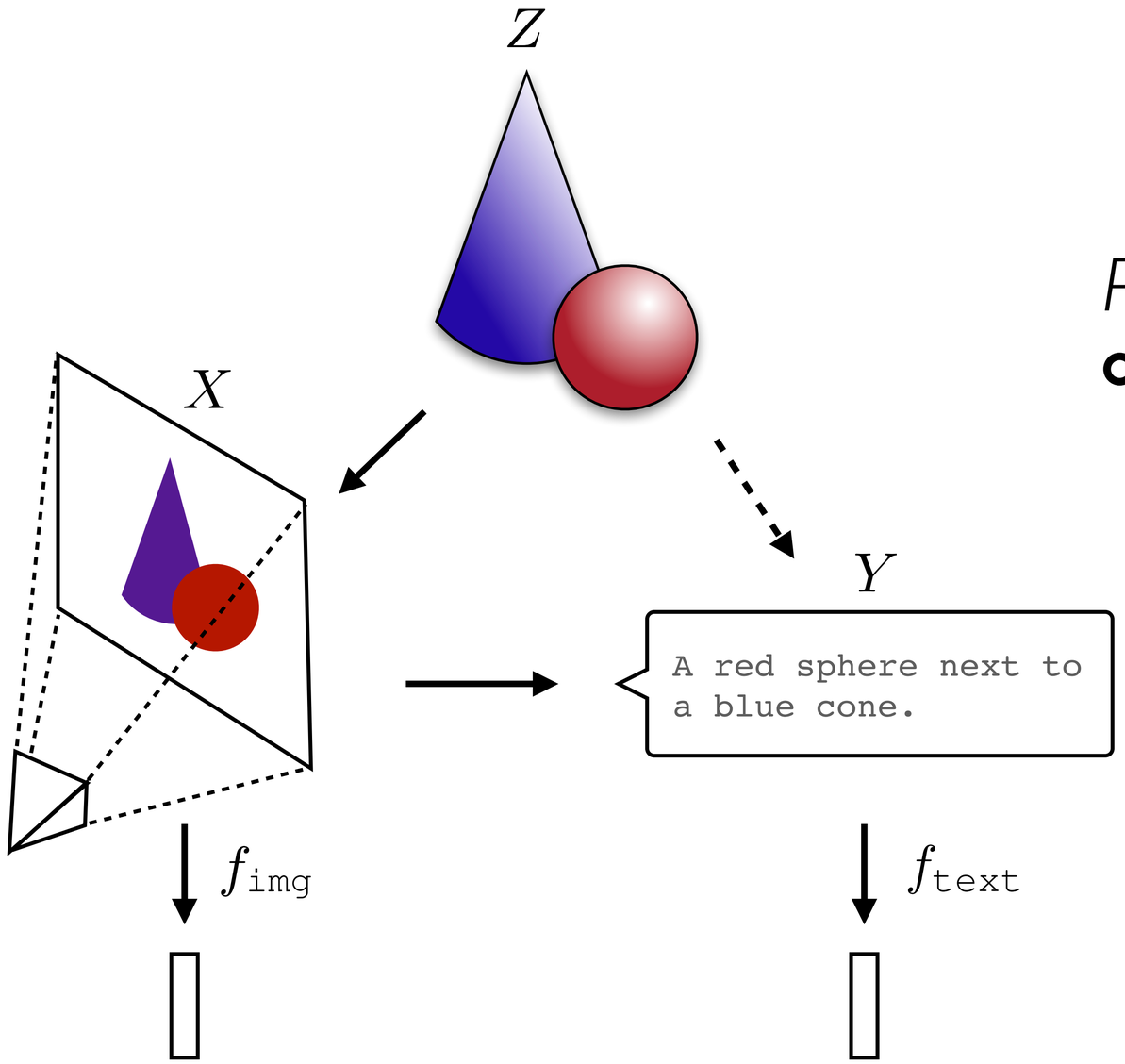
THE receptive fields of simple cells in mammalian primary visual cortex can be characterized as being spatially localized,

21

Why is this happening?

- ~~It's all about the data!~~
- ~~It's all about transformers!~~
- *It's all about the world*

[Plato 375 BC]



Platonic representation hypothesis

Different neural networks are converging toward the **same** way of representing the world (*same* kernel).

[Huh*, **Cheung***, Wang*, Isola* 2024]

How many words is a picture worth?

5 words

"Illuminated escalators in indoor mall."

10 words

"Illuminated escalators in a lush indoor mall with greenery and polished tiles."

20 words

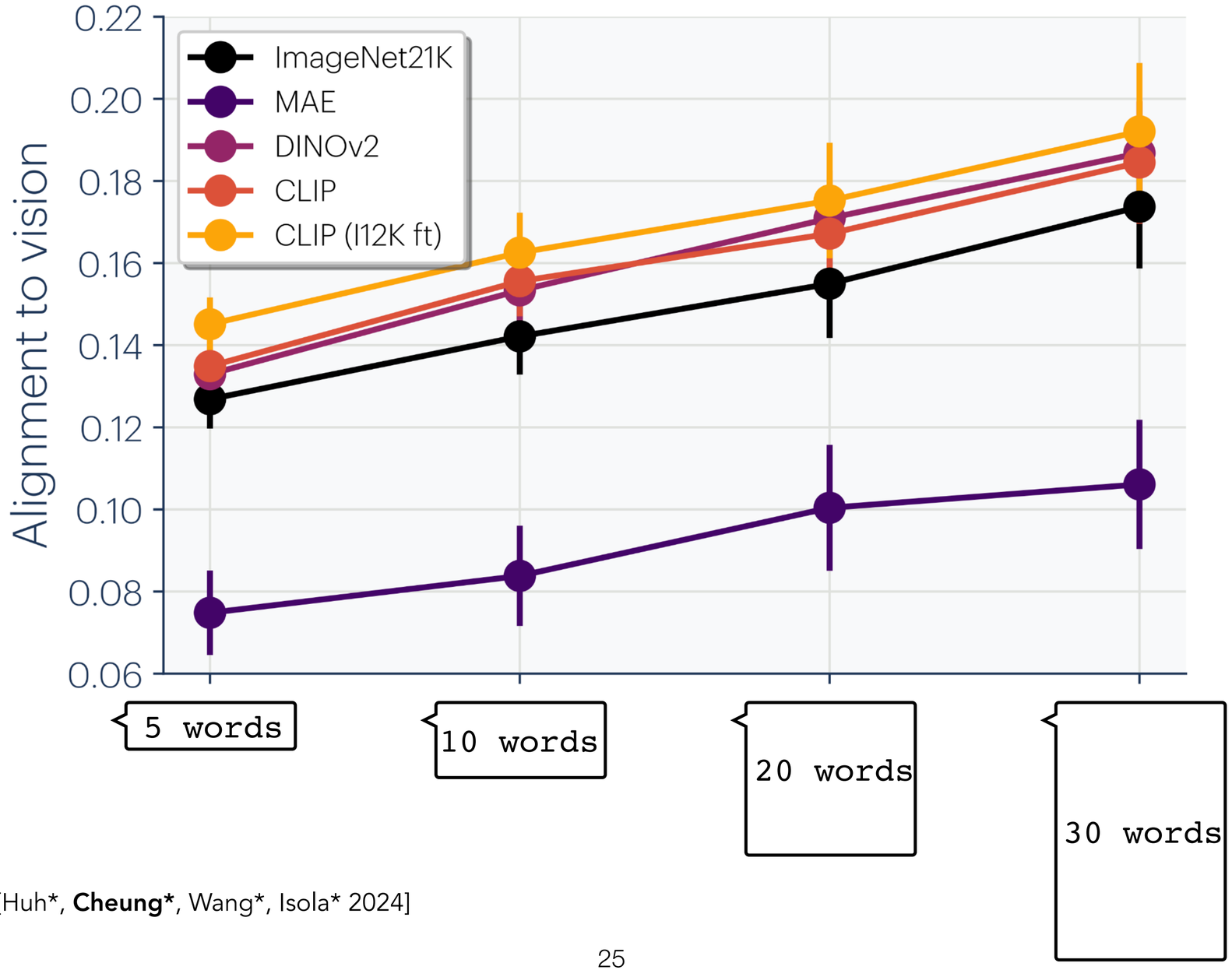
"Illuminated escalators in a vibrant indoor mall with lush greenery, polished tiles, and tropical plants, surrounded by stone columns and recessed lighting."

30 words

"An indoor shopping mall with three illuminated escalators, surrounded by lush greenery and polished colored tiles, featuring a man ascending one escalator and various shops and plants on the upper level."

DCI Dataset [Urbanek*, Bordes* et al. 2023]

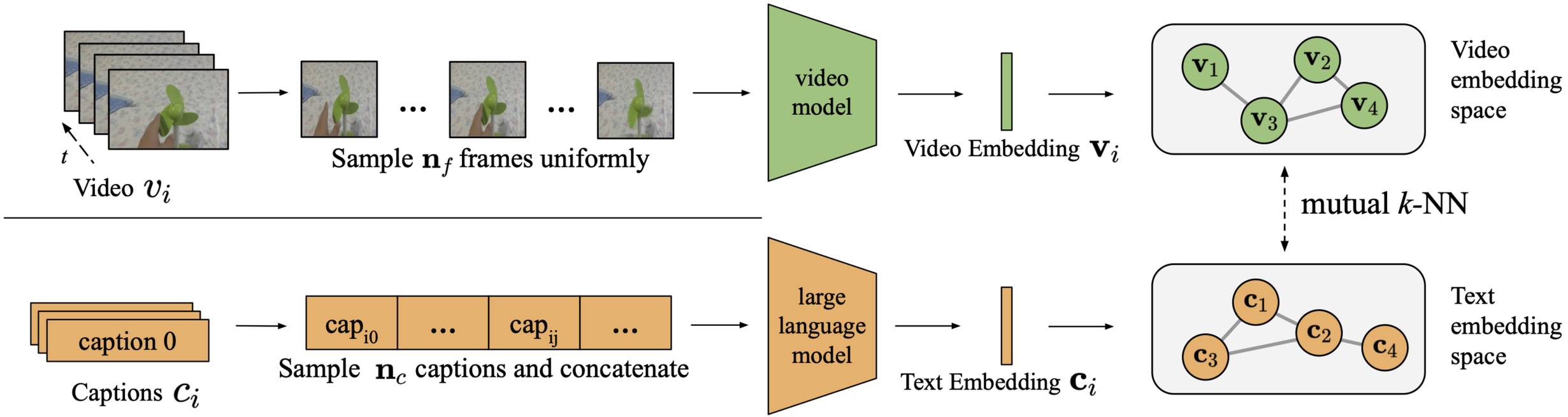
How many words is a picture worth?



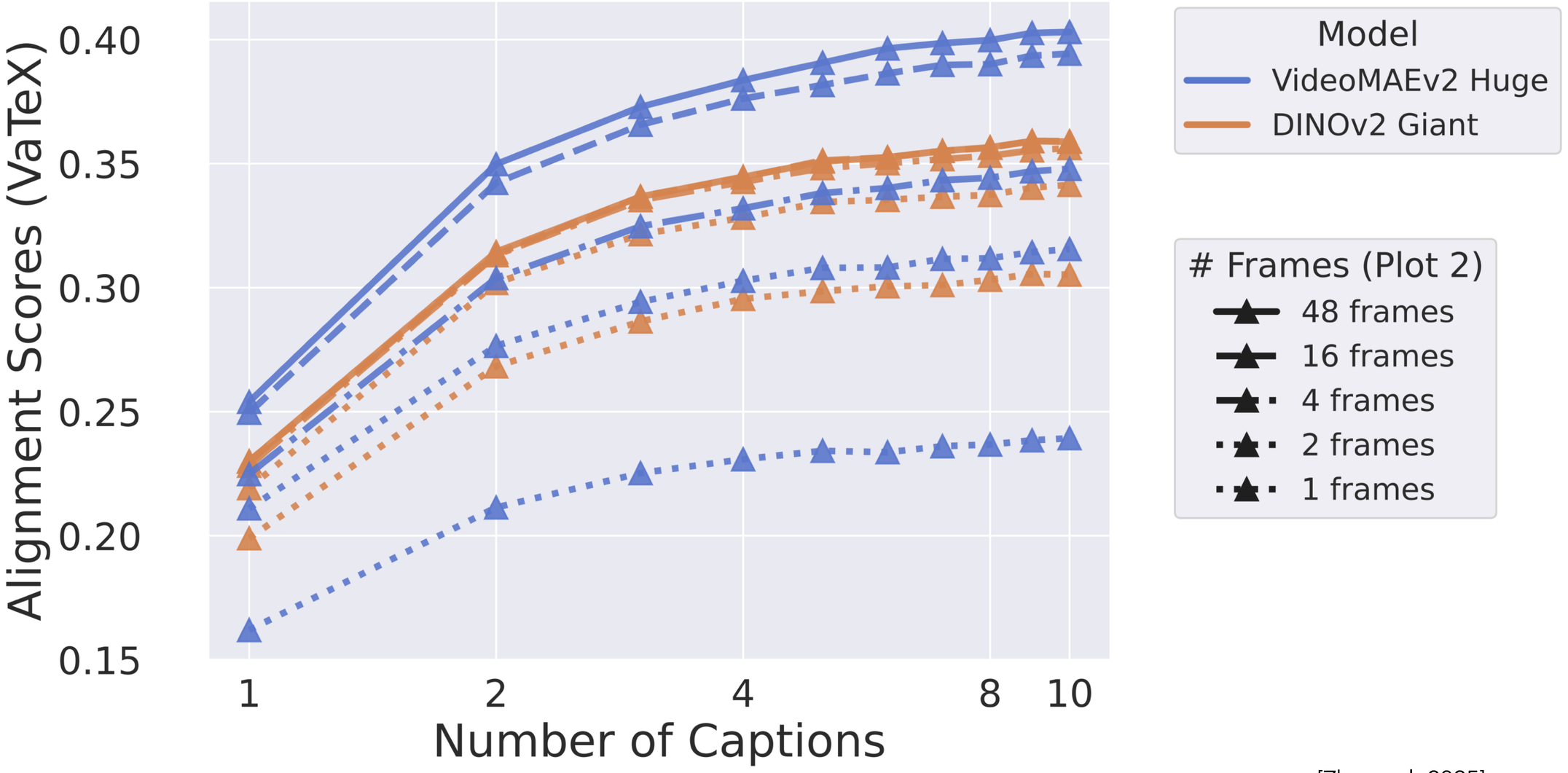
[Huh*, **Cheung***, Wang*, Isola* 2024]

DYNAMIC REFLECTIONS: PROBING VIDEO REPRESENTATIONS WITH TEXT ALIGNMENT

Tyler Zhu[†*] **Tengda Han**[‡] **Leonidas Guibas**[‡] **Viorica Pătrăucean**[‡] **Maks Ovsjanikov**[‡]
Princeton University[†] Google DeepMind[‡]

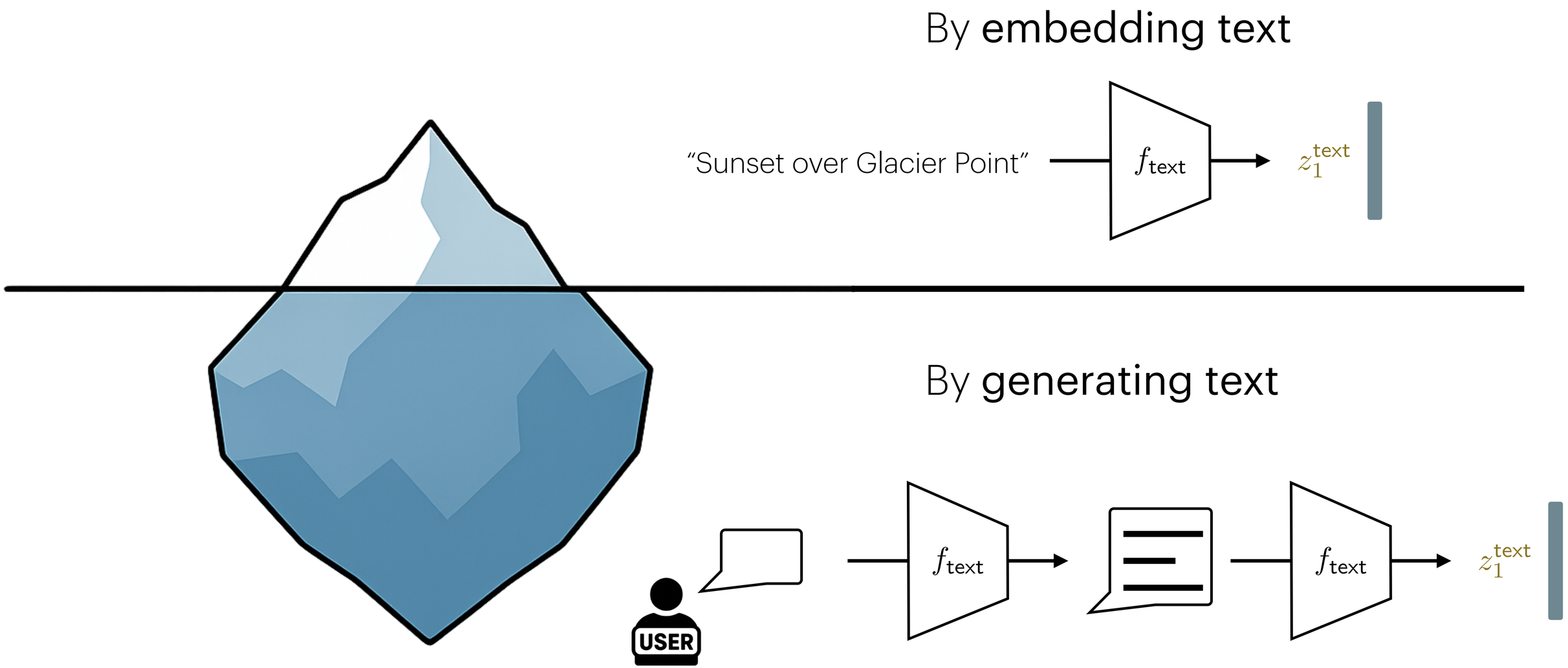


How many captions is a video worth?

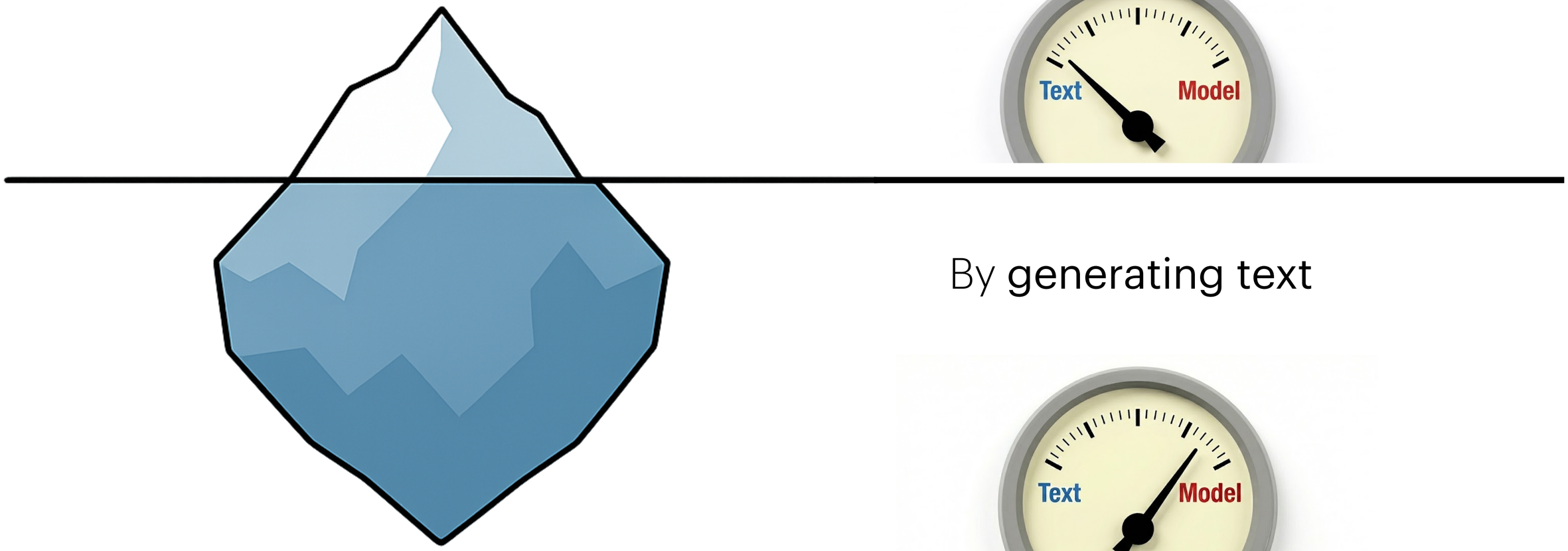


[Zhu et al. 2025]

Probing what language models know



Influence on model representation



By **embedding** text

By **generating** text

“You can just do things.”

—Sam Altman, CEO of OpenAI

You can just ask for things.

Chain-of-Thought Prompting

Zero-shot

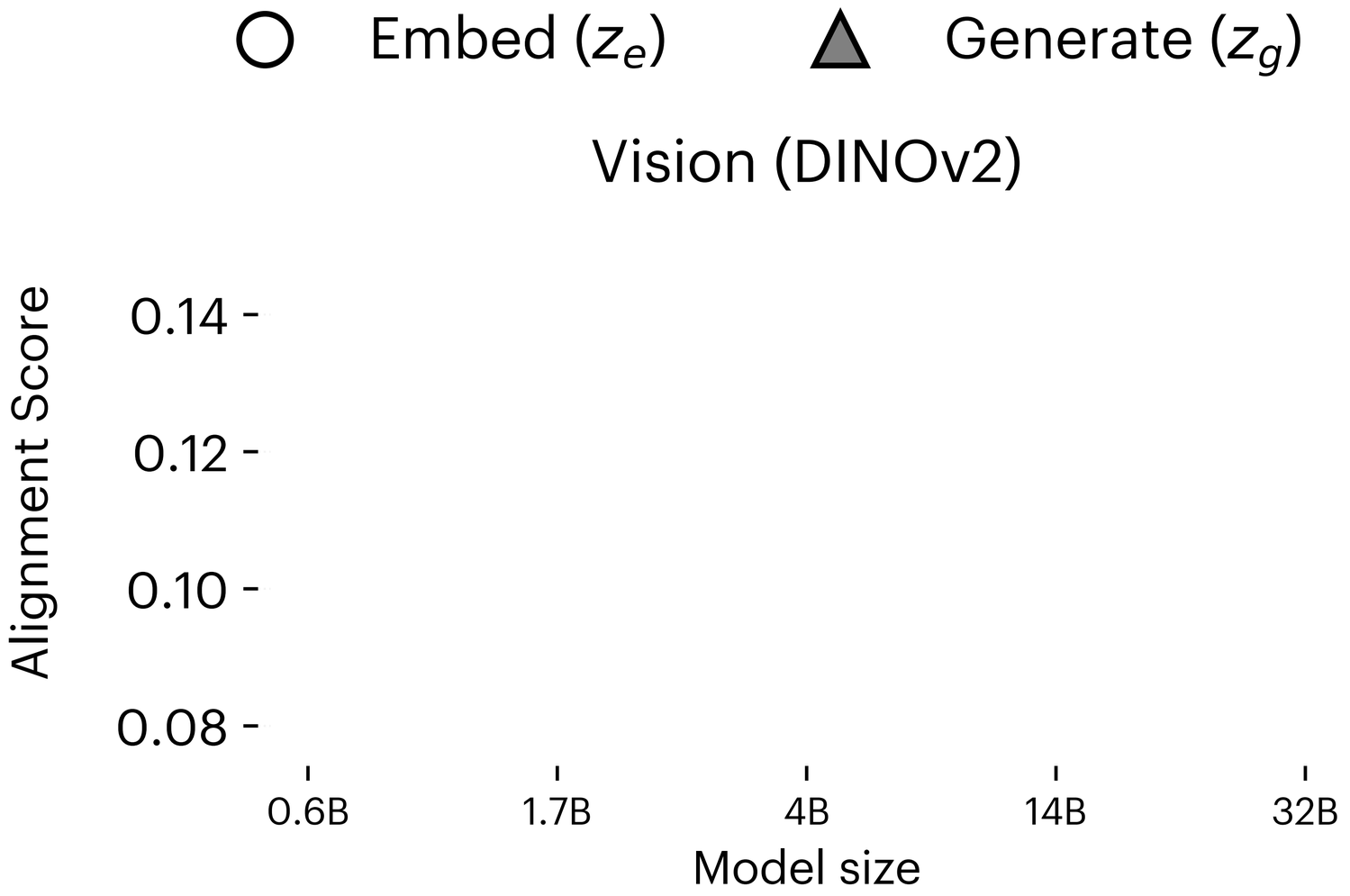
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

Zero-shot-CoT

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

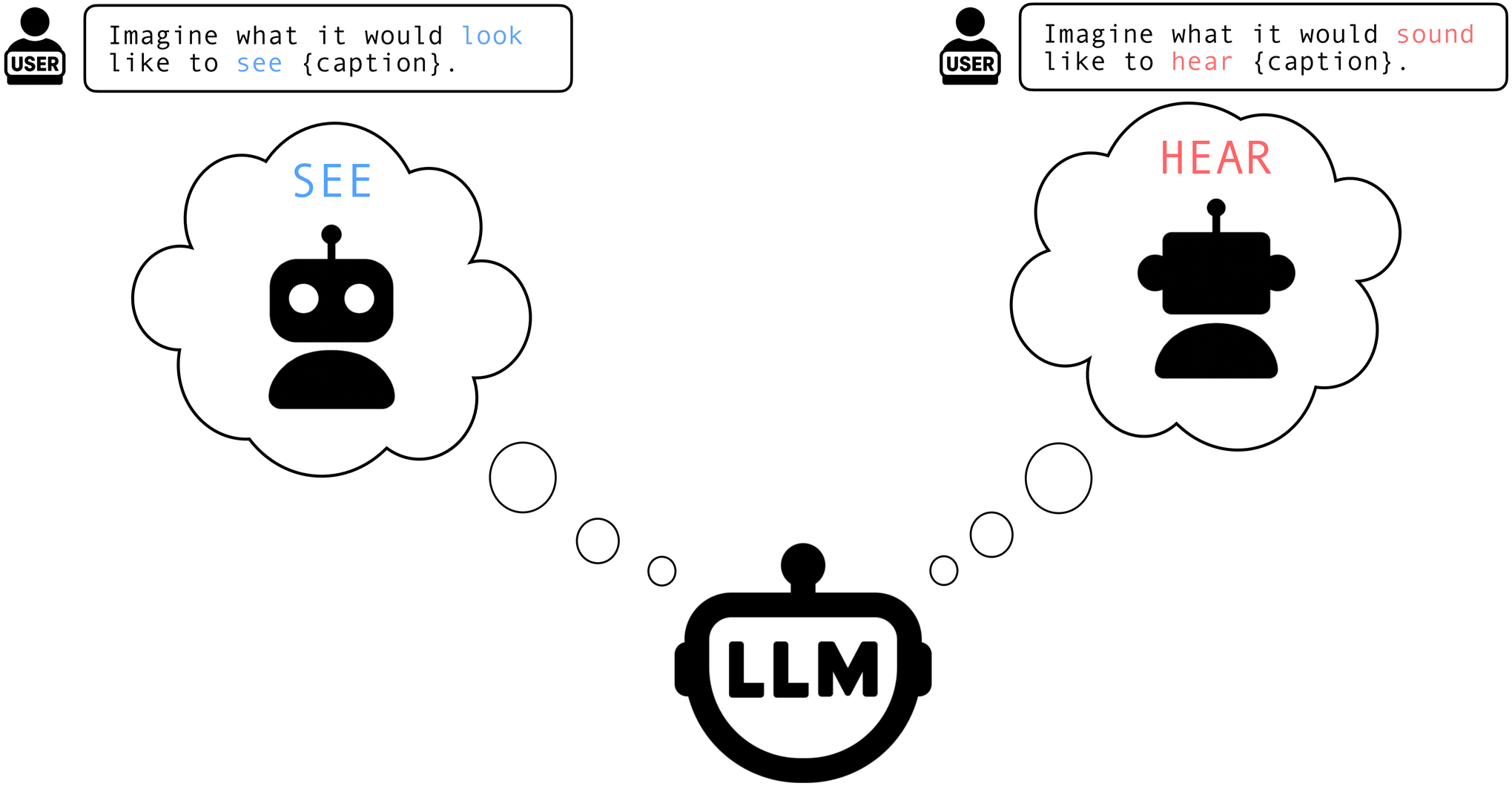
[Kojima et al. 2022]

Asking the model to `Imagine: {caption}.`



[Wang, Isola, **Cheung*** 2025]

What happens when you ask a language model to **imagine senses it never experienced**?

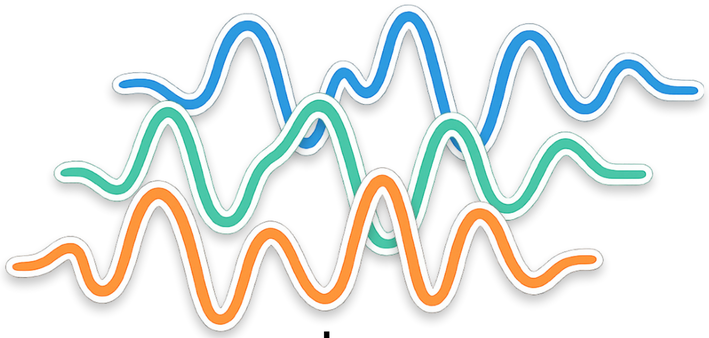


Sensory Prompting

Vision Model

"Awaiting a pitch: batter, catcher, and umpire in baseball"

Language Model



Audio Model

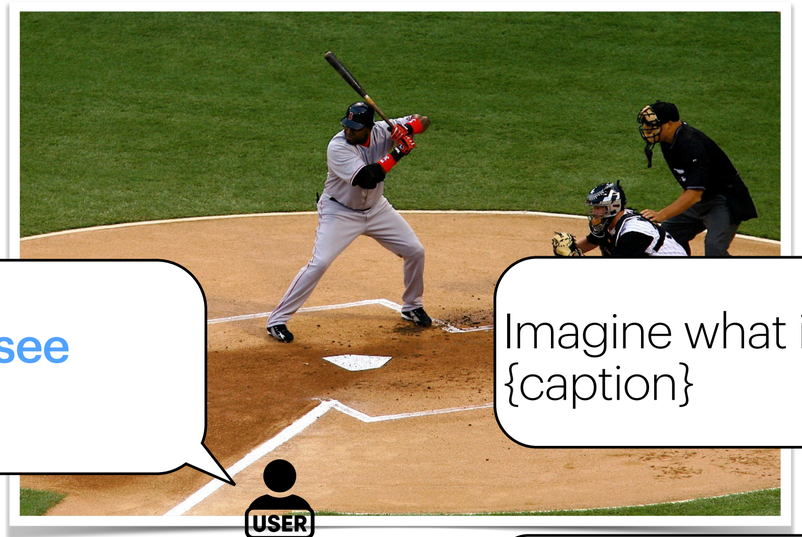
IMAGINE SEEING

IMAGINE

IMAGINE HEARING

Space of kernels

[Wang, Isola, **Cheung** 2025]



Imagine what it would **look** like to **see** {caption}

USER

...I need to visualize a baseball field. The batter stands at the home plate, right? They're holding the bat, probably in a ready position, feet shoulder-width apart. Their body language should show they're focused...

LLM

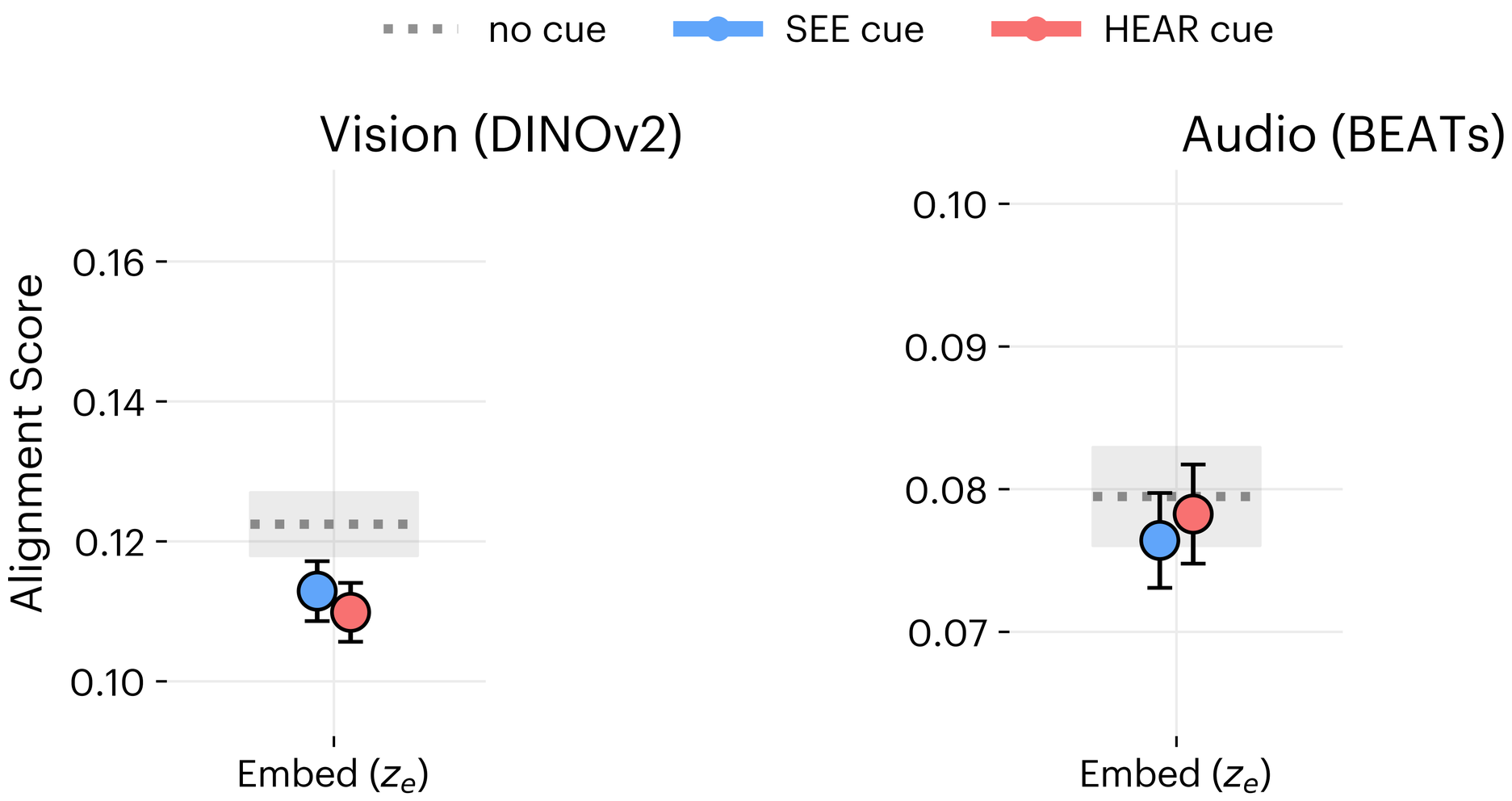
Imagine what it would **sound** like to **hear** {caption}

USER

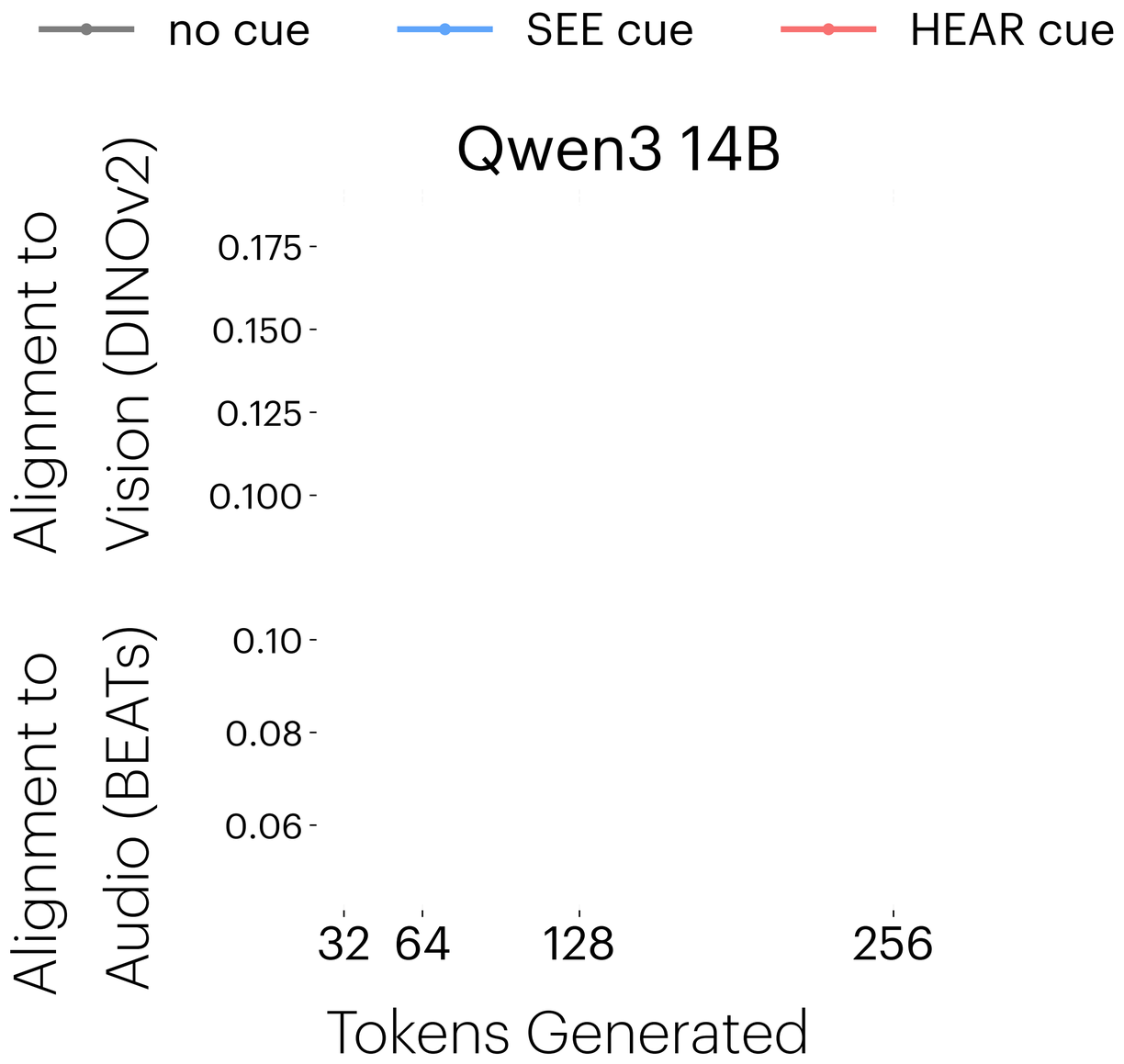
...There's the general ambient noise of the stadium: crowd murmurs, maybe some distant cheers or chatter. The sound of players moving, like cleats scraping on the dirt, gloves being adjusted...

LLM

Sensory Embedding vs Sensory Generation



How many **spoken** words is a picture worth?



Redirecting Sensory Cues

Seeing to Hearing

Please rewrite the following text to describe what the scene would sound like audibly, rather than how it looks: {text}

USER

The scene hums with anticipation, the only sounds the soft rustle of grass ~~underfoot and the occasional murmur of the crowd in the distance. A tense silence~~ hangs in the air...

LLM

Hearing to Seeing

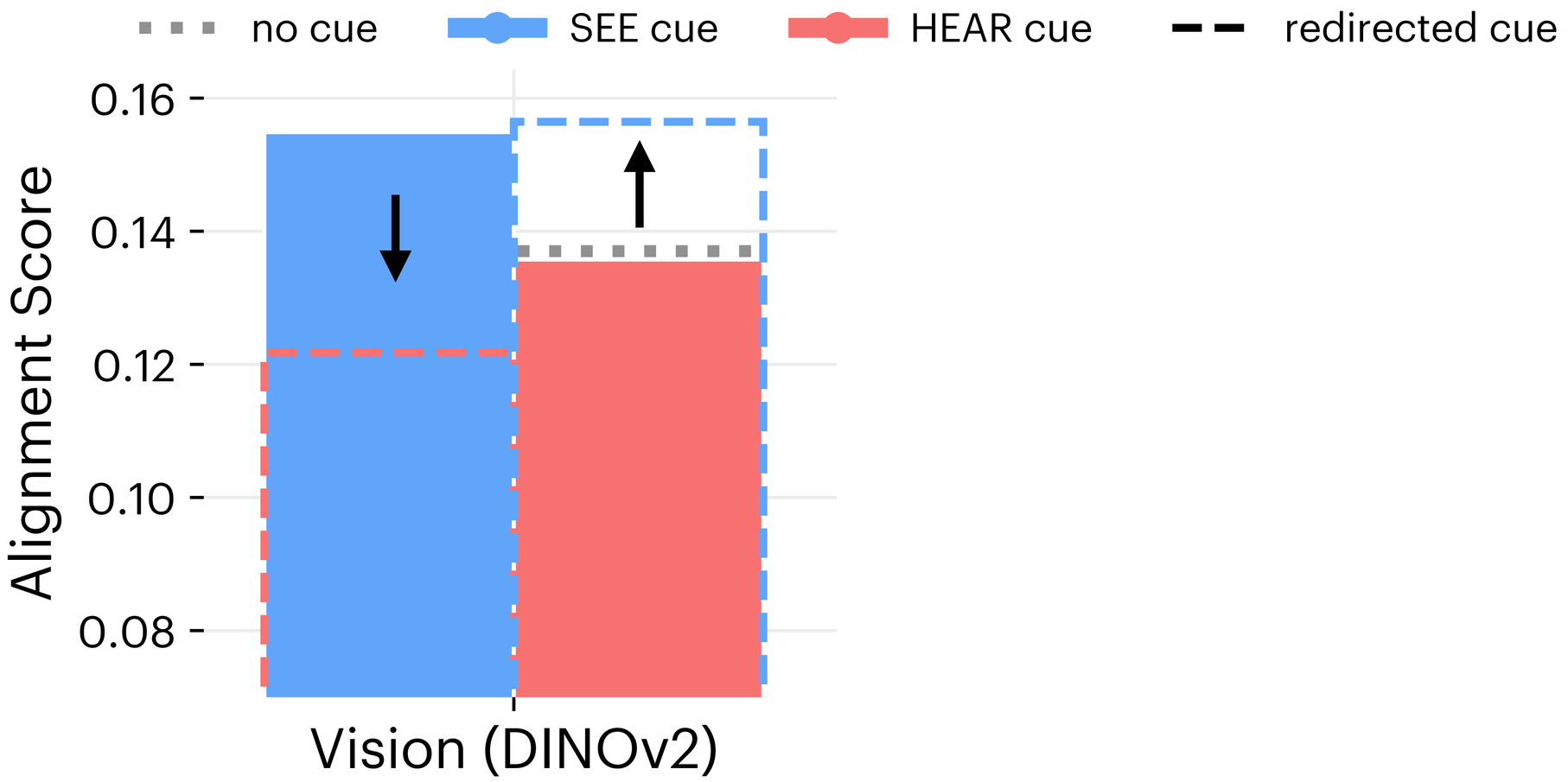
Please rewrite the following text to describe what the scene would look like visually, rather than how it sounds: {text}

USER

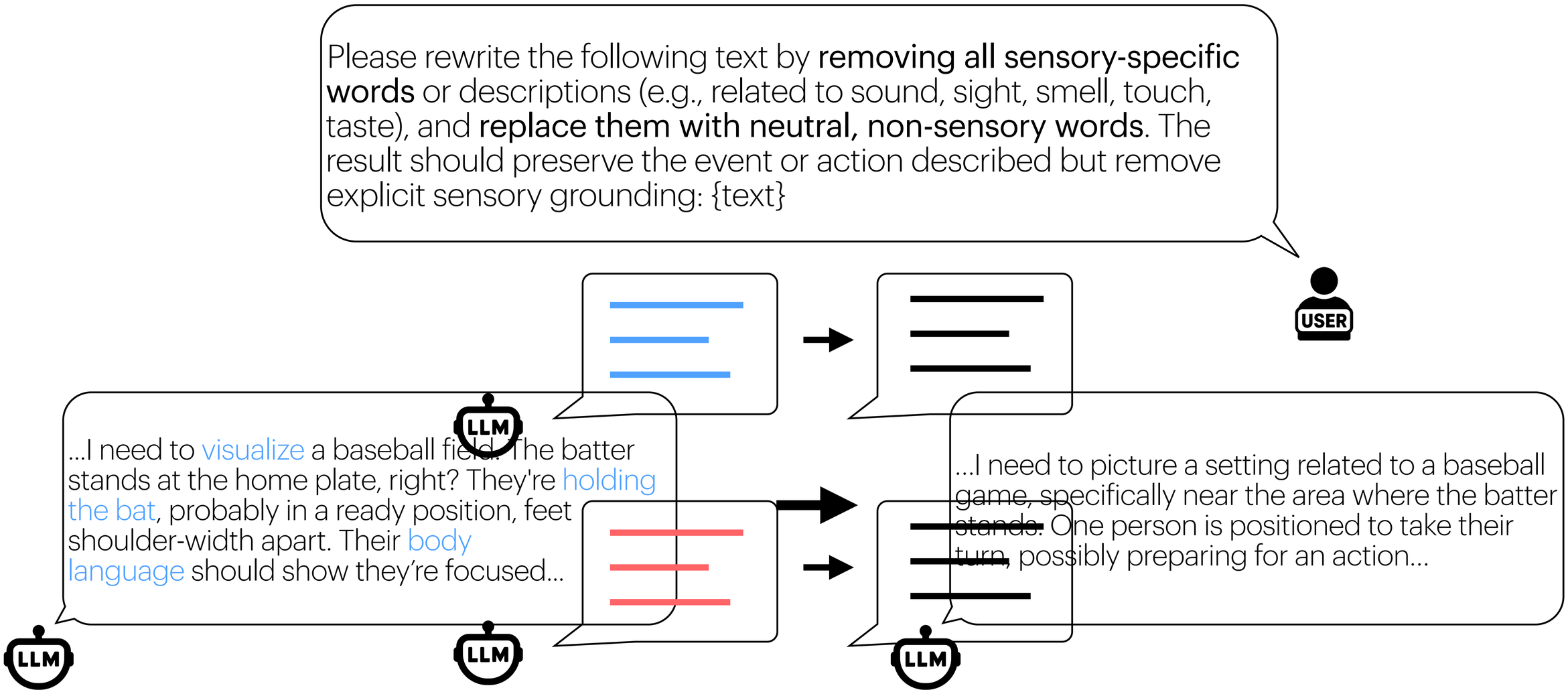
...The batter stands at the plate, gripping the bat tightly, his muscles coiled with anticipation. His eyes are locked onto the pitcher, scanning for the slightest movement...

LLM

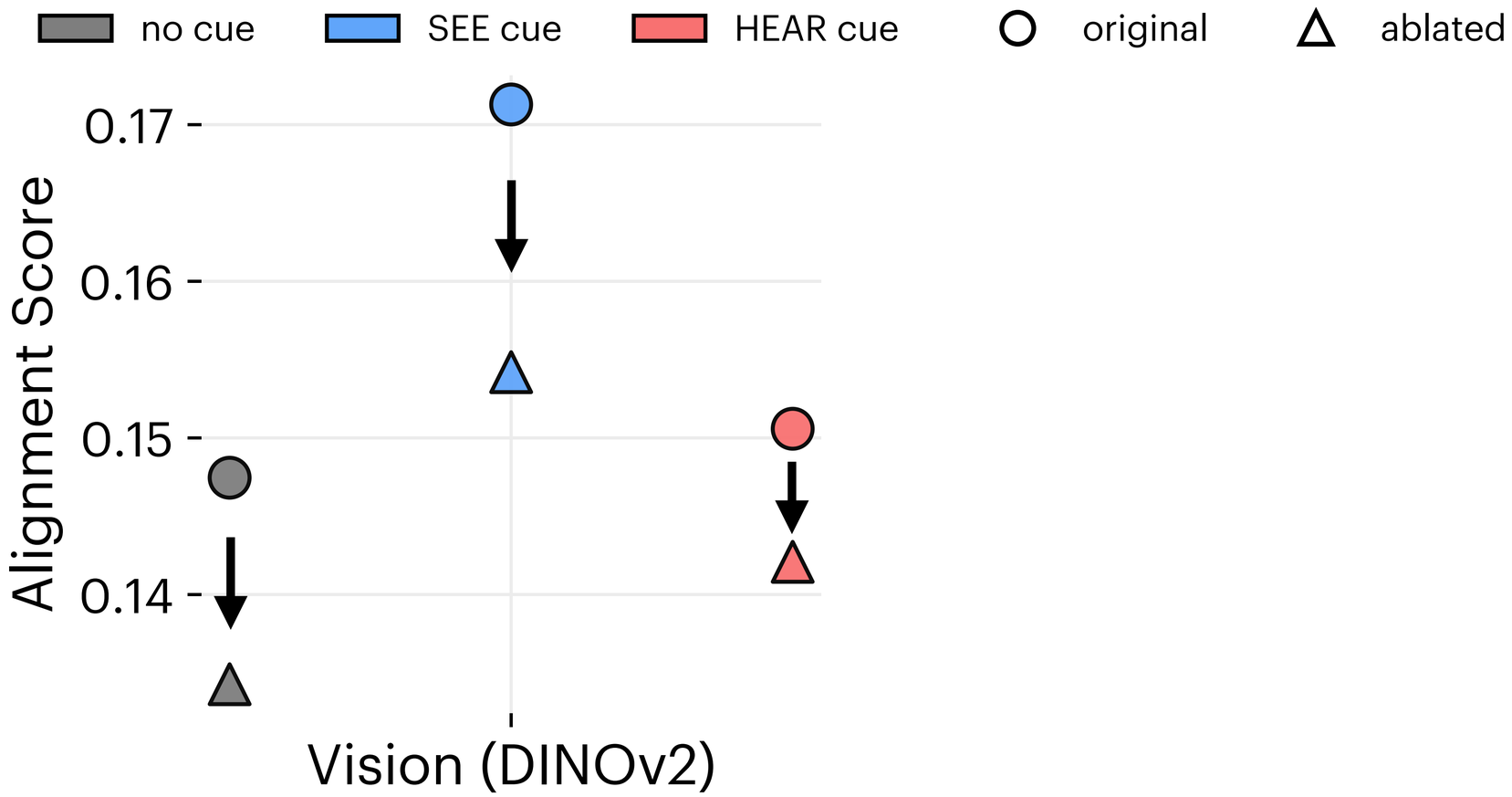
Redirecting sensory cues preserves modality-specific alignment



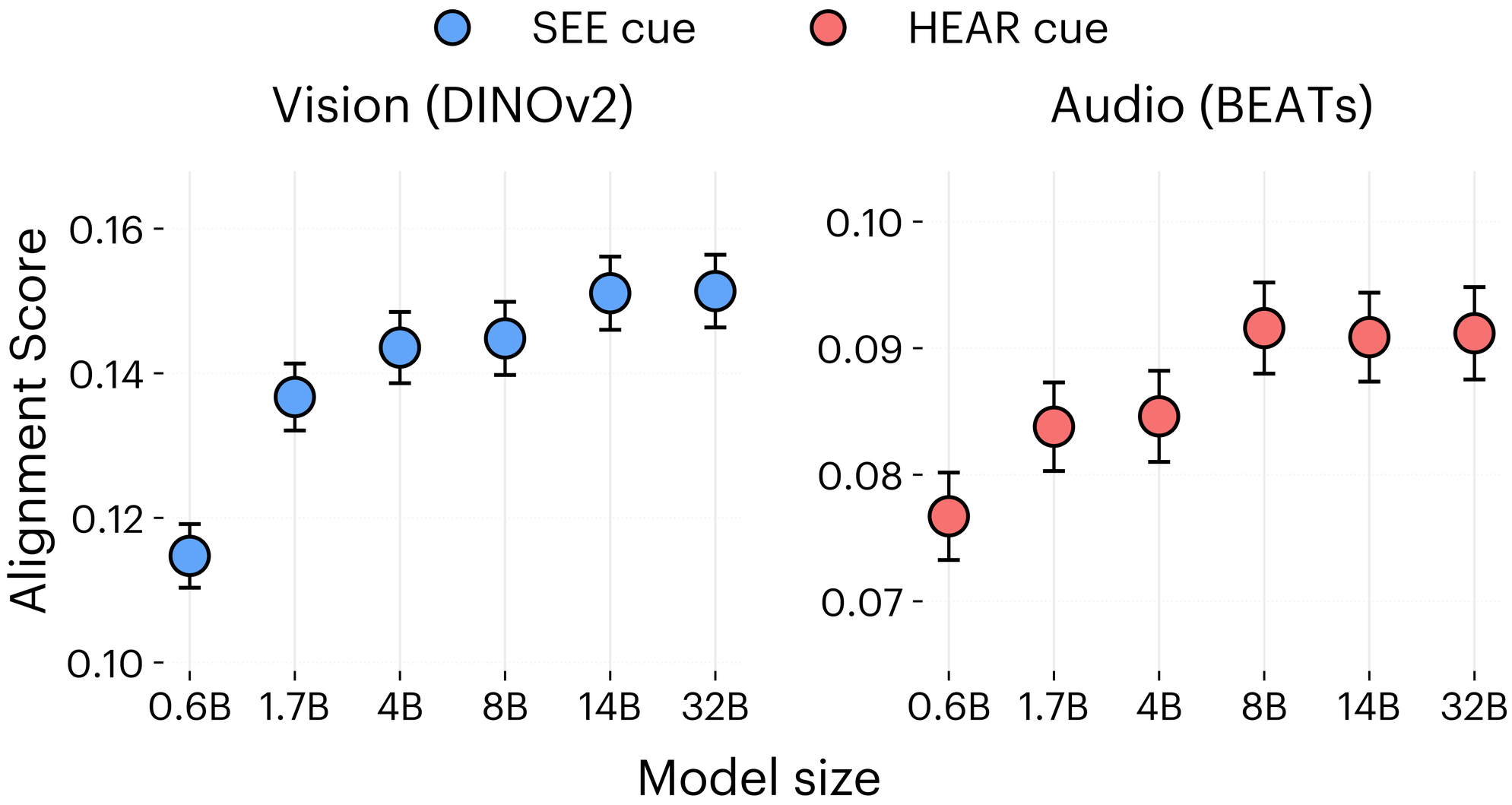
Removing sensory references



Use of sensory language strengthens modality-specific representations



Alignment is higher in larger models



What about senses humans don’t have?

X-ray Crystallography

https://msg.ucsf.edu/rigaku-ru-200

Cryo-EM

https://emcore.ucsf.edu/krios-g2-cryo-tem

What about senses humans don’t have?

LLM

Bio Sensory Prompting

SEQUENCE:
KKAVINGEQIRSISDLHQTLKKELALPEYYGENLDALWDCLTGWV
EYPLVLEWRQFEQSKQLTENGAESVLQVFREAKAEGCDITIILS

USER

Provide a thorough summary of {protein name}. Include its gene name, protein family, molecular weight, known structural domains, function in the cell, binding sites, any known interactions or pathways it participates in.

USER

ESM3

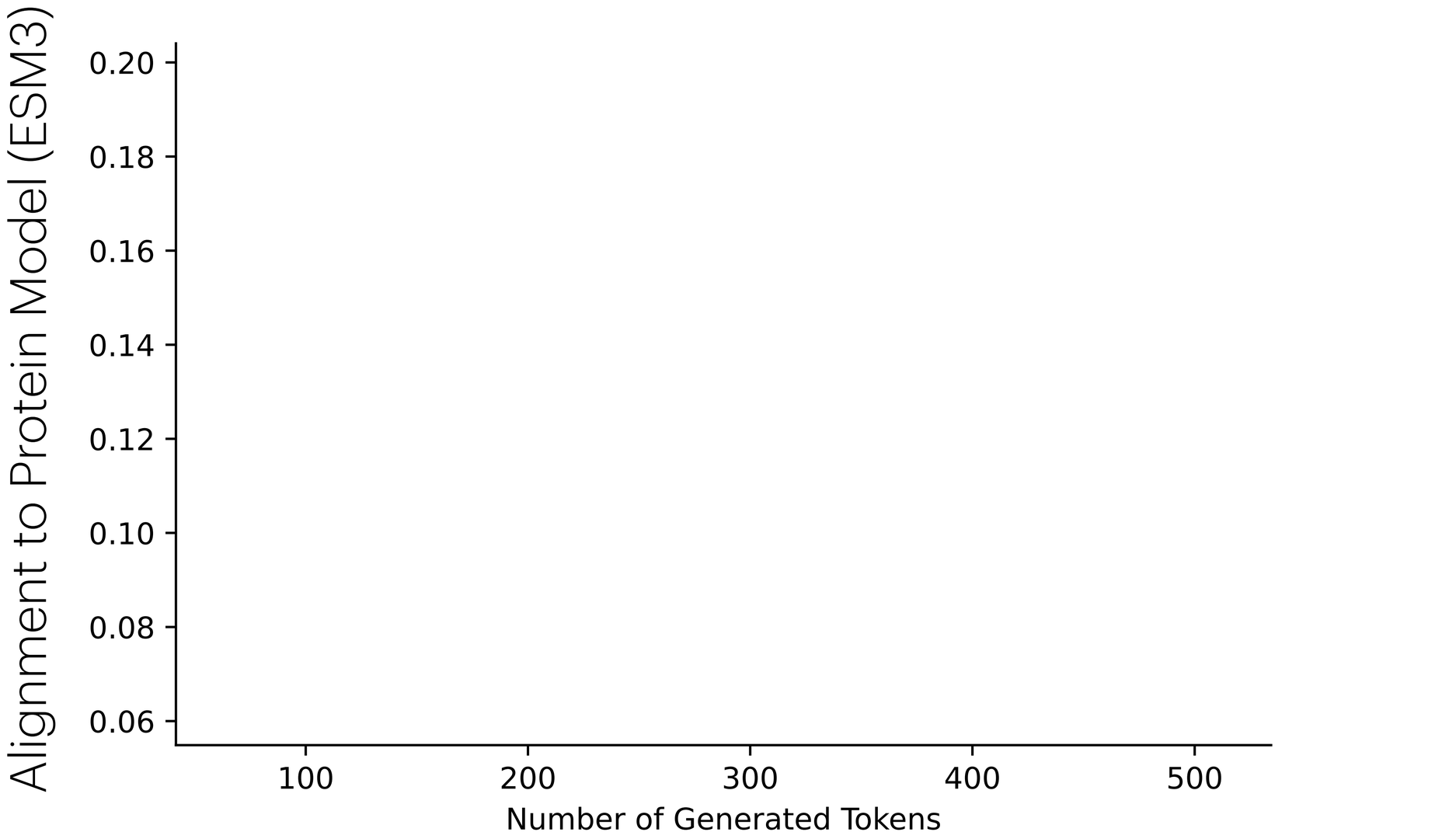
RIBONUCLEASE SA COM

Okay, I need to provide a thorough summary of the RIBONUCLEASE SA COMPLEX WITH BARSTAR. Let me start by recalling what I know about these proteins.

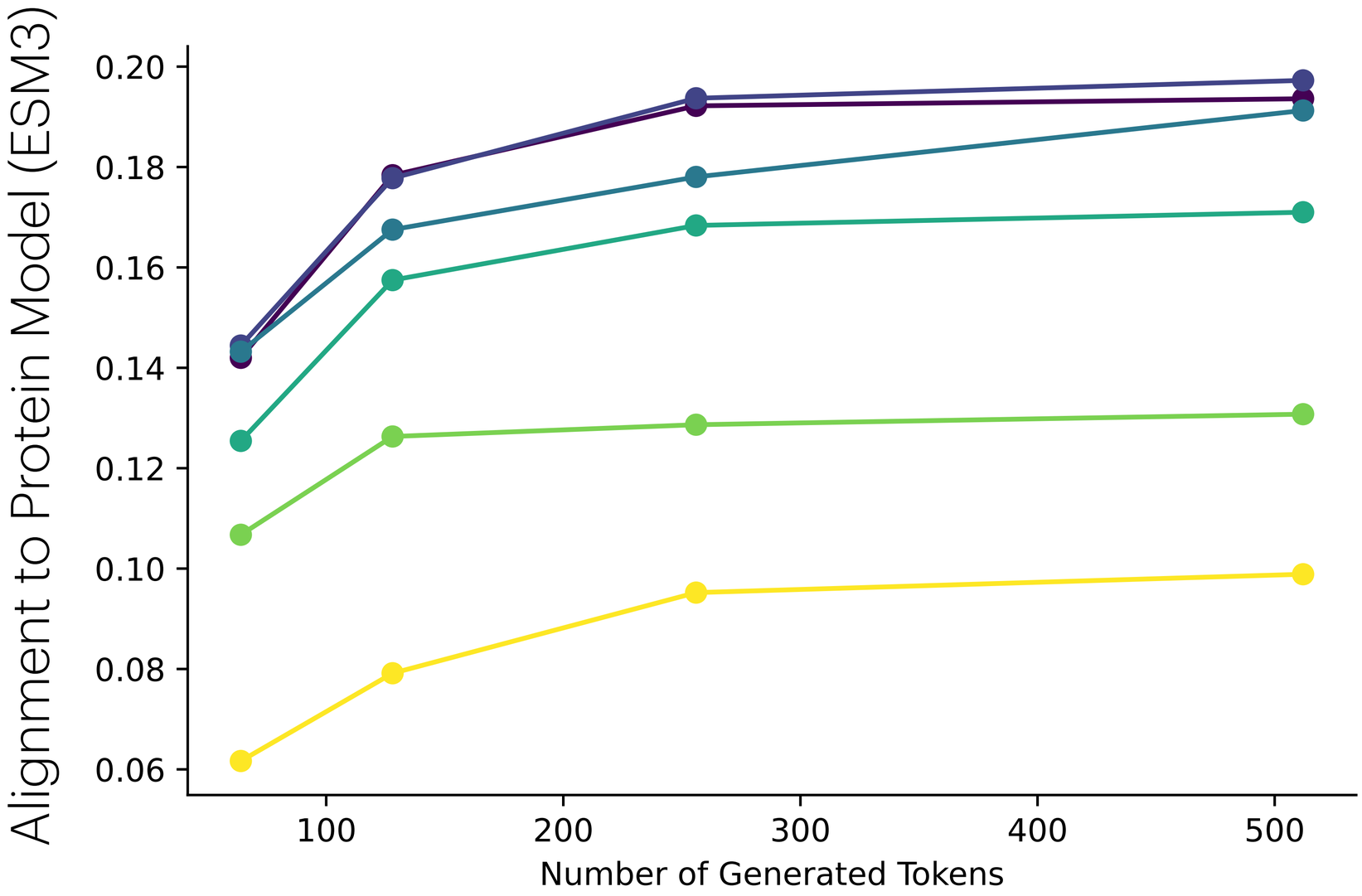
First, the gene names. I remember that RNase SA is from the bacterium Staphylococcus aureus. The gene might be named rnsa, but I should verify that. Barstar is the inhibitor of RNase SA, so its gene might be bar...

LLM

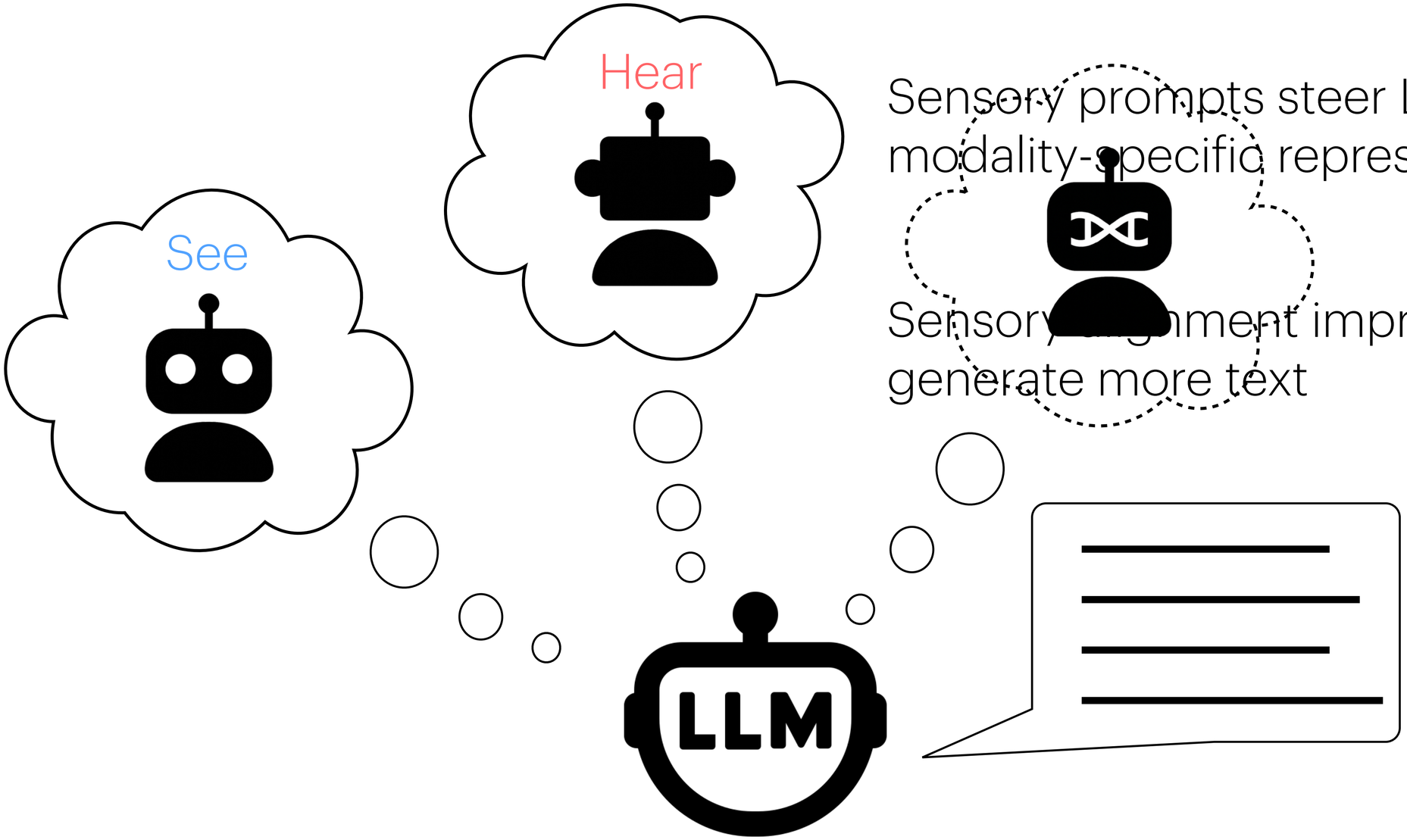
Bio Sensory Prompting



Bio Sensory Prompting



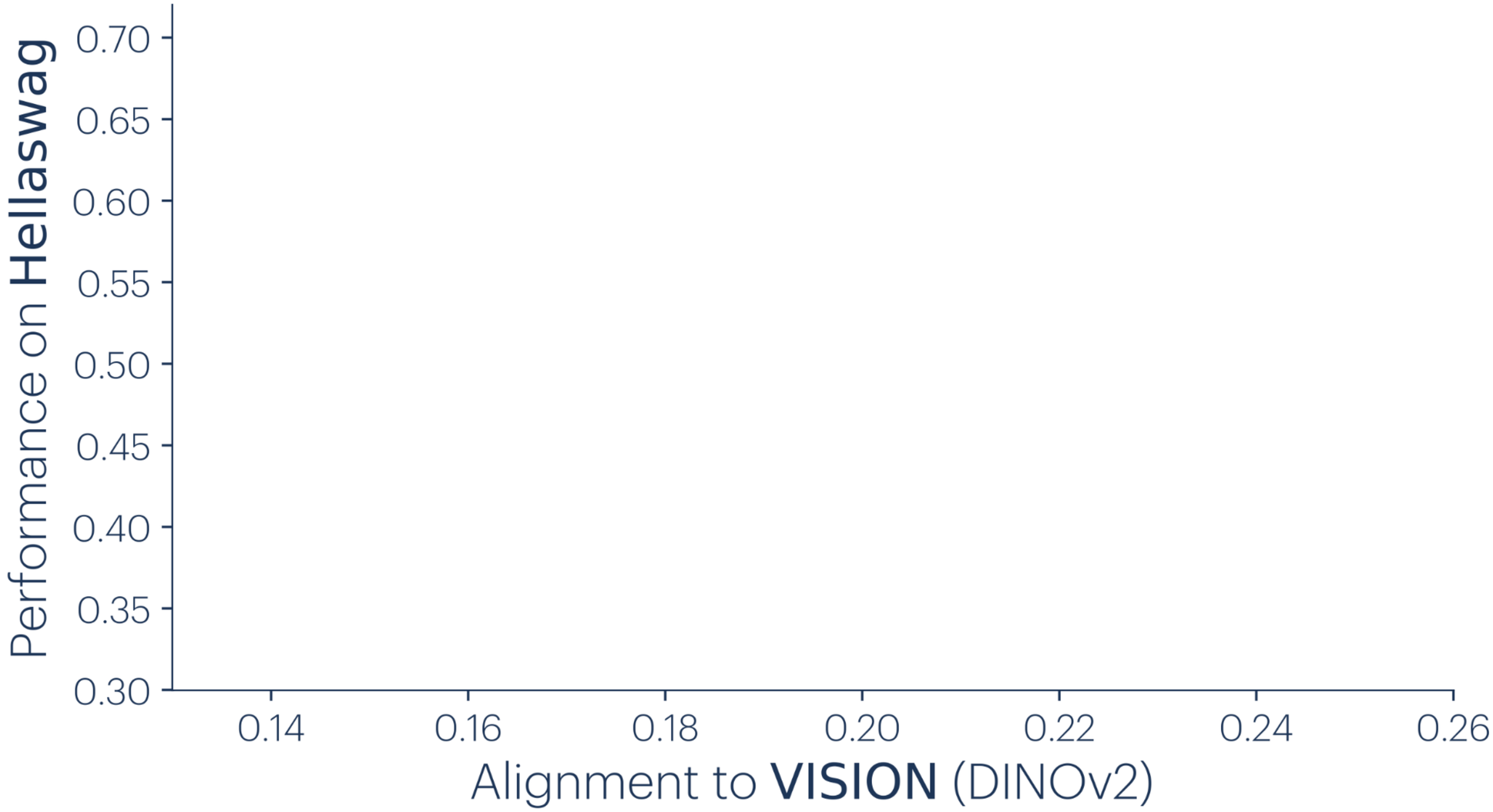
You can just ask for perception

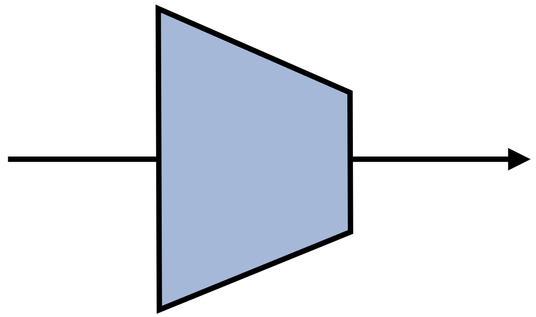
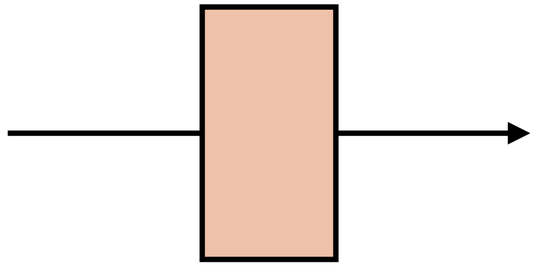


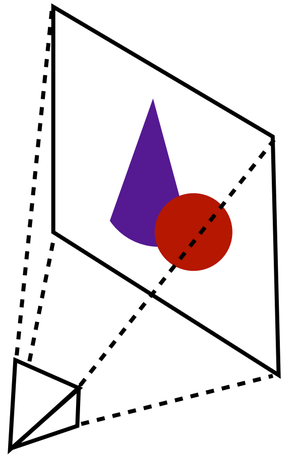
Sensory prompts steer LLMs toward modality-specific representations.

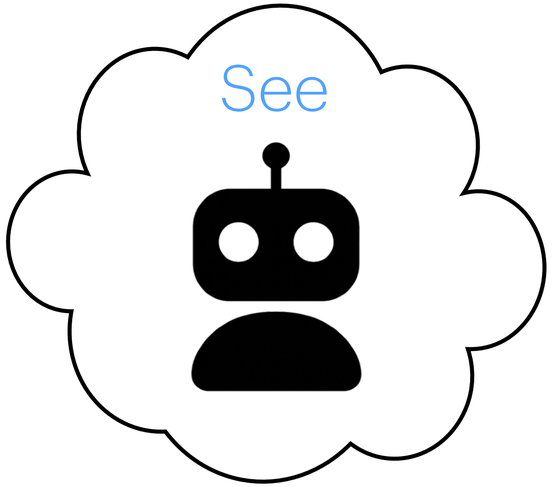
Sensory alignment improves as LLMs generate more text

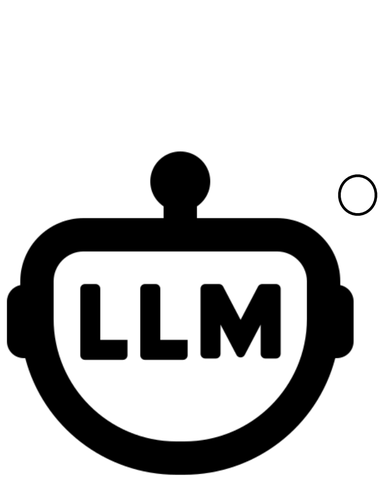
Representational alignment predicts downstream performance



sim( , )

sim( , A red sphere next to a blue cone.)

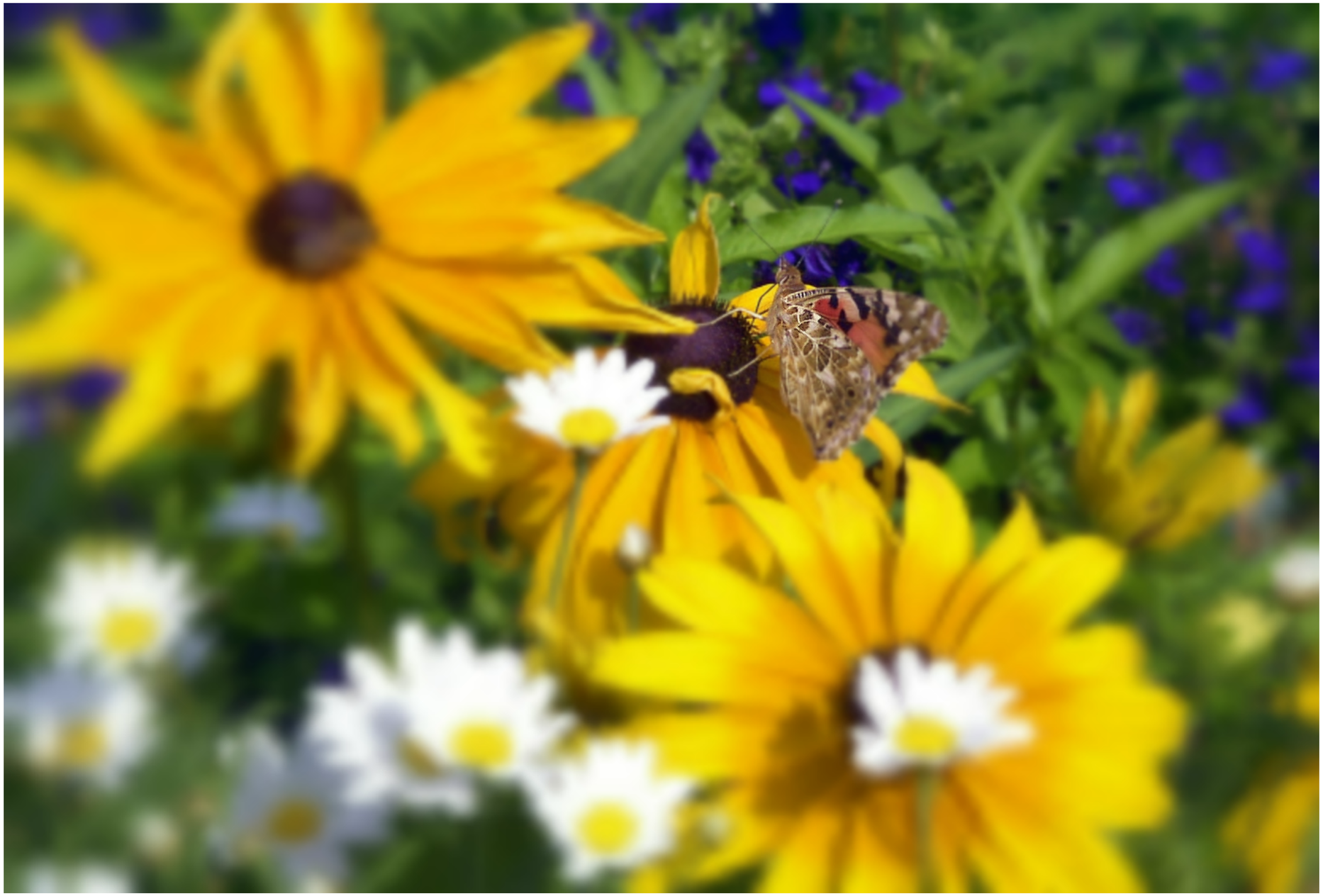
sim(



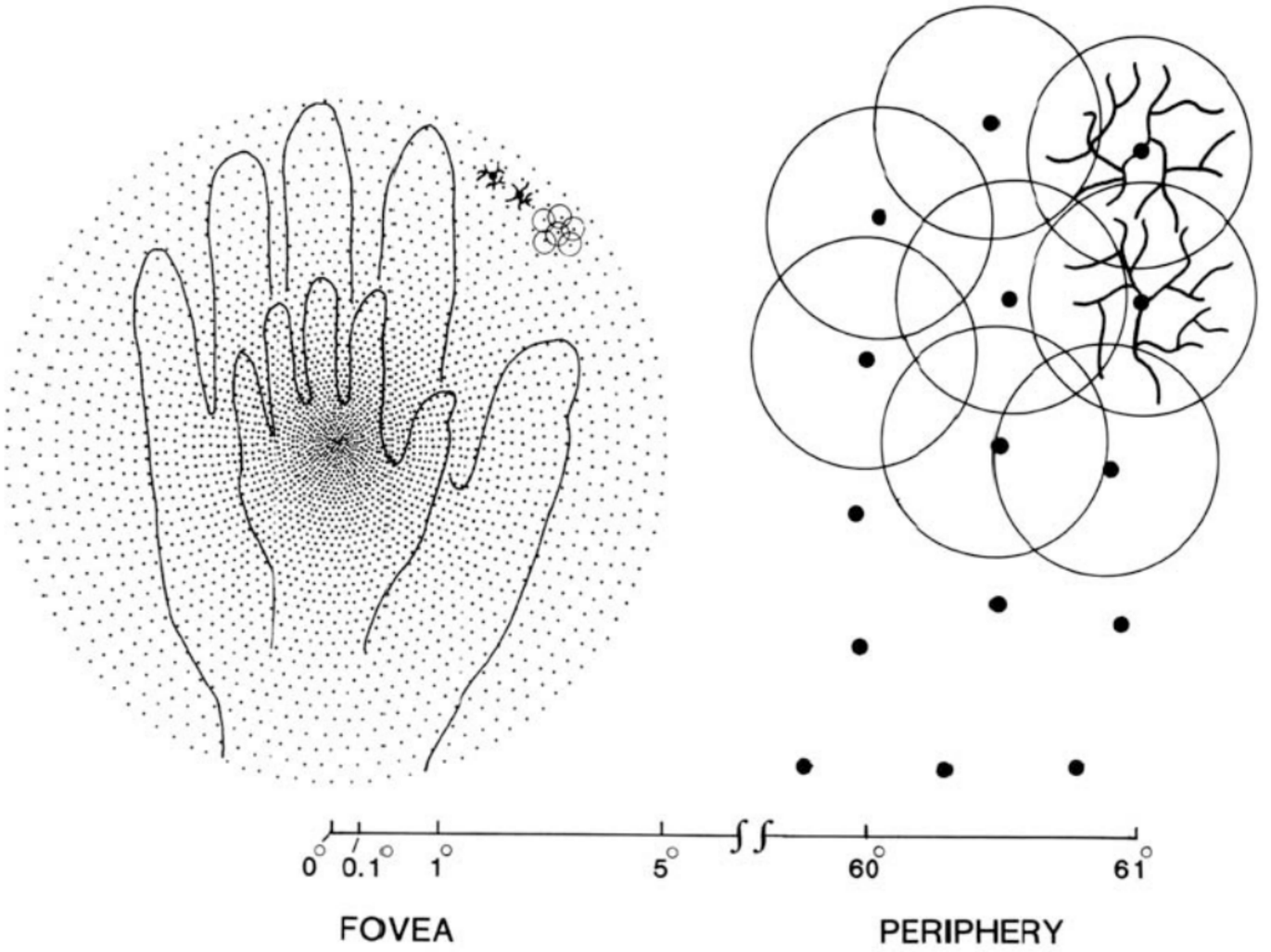
, Vision Model ↑ )

With the convergence of artificial and biological systems, what questions about biology can we answer?

Why do we have a fovea?

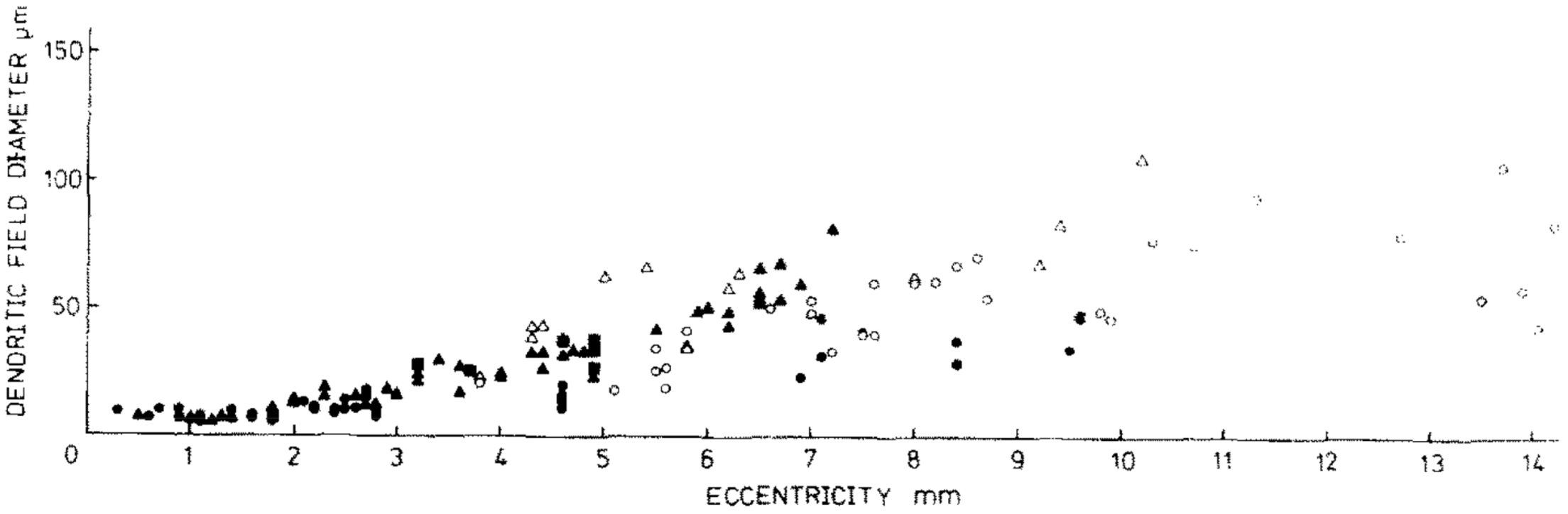


What is a fovea?



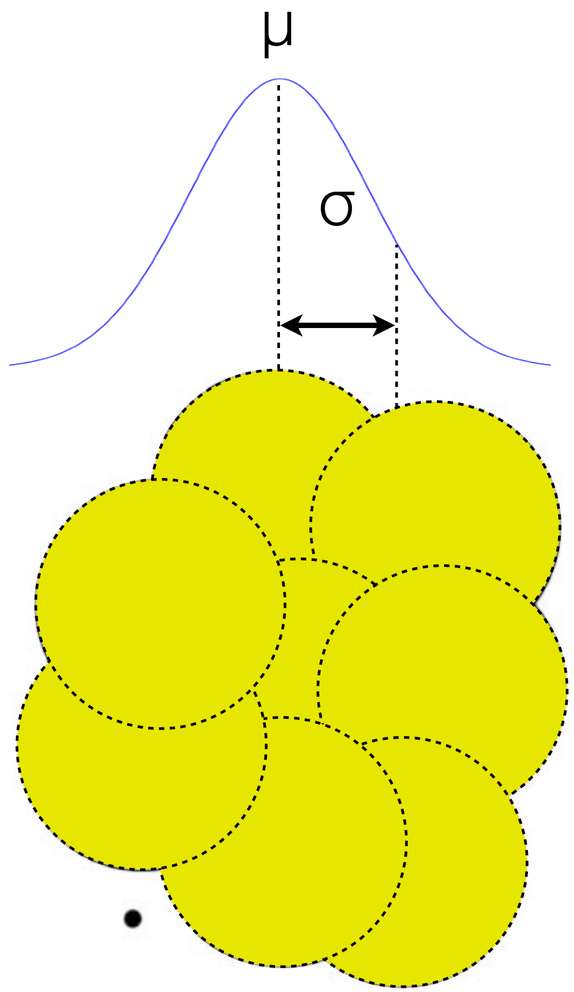
[Van Essen and Anderson 1995]

Scaling Law for Structure

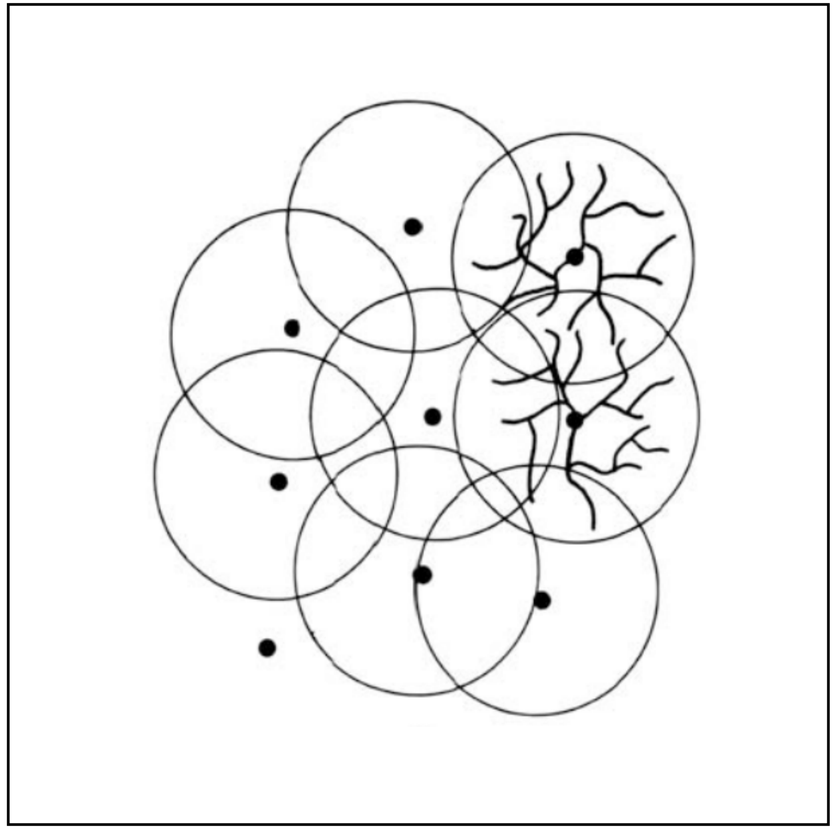


[Perry, Oehler, Cowey 1984]

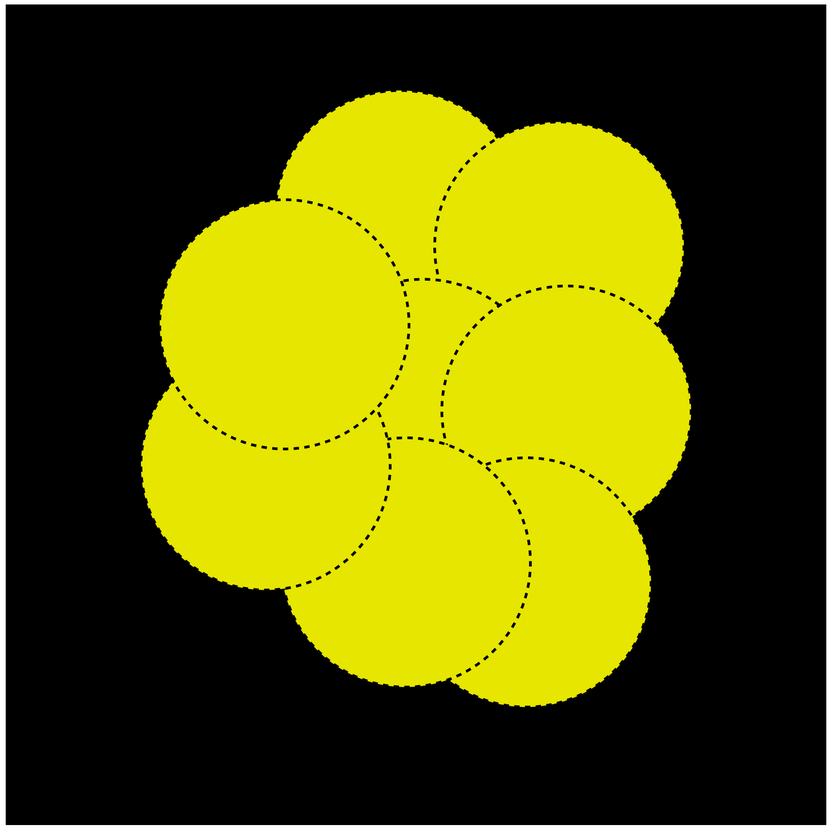
Model of a Retinal Ganglion Cell



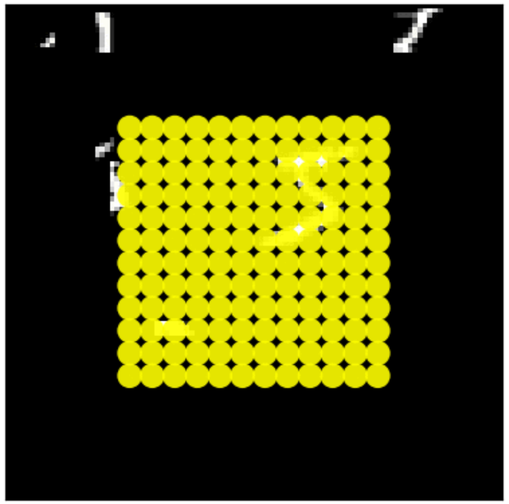
in-vivo



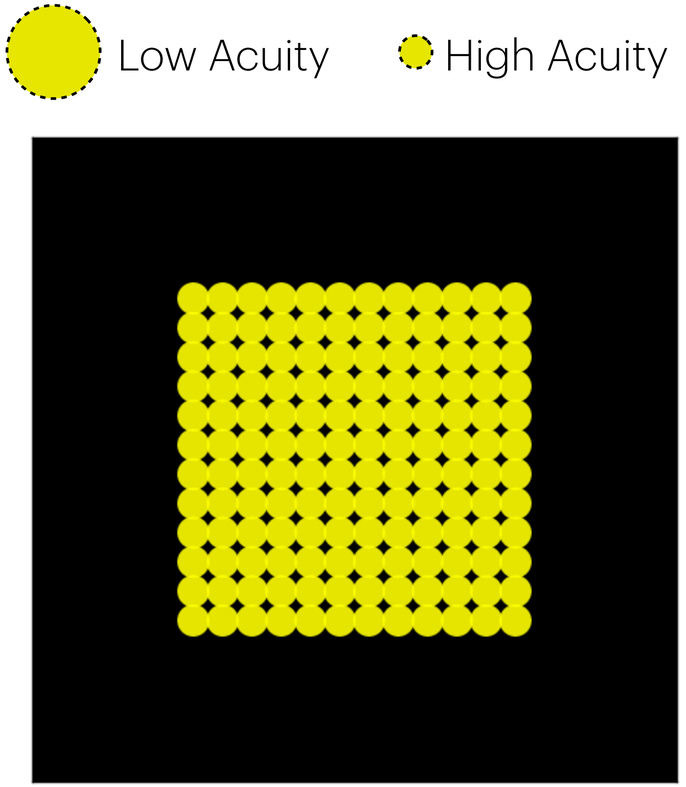
in-silico



Attention with Saccades

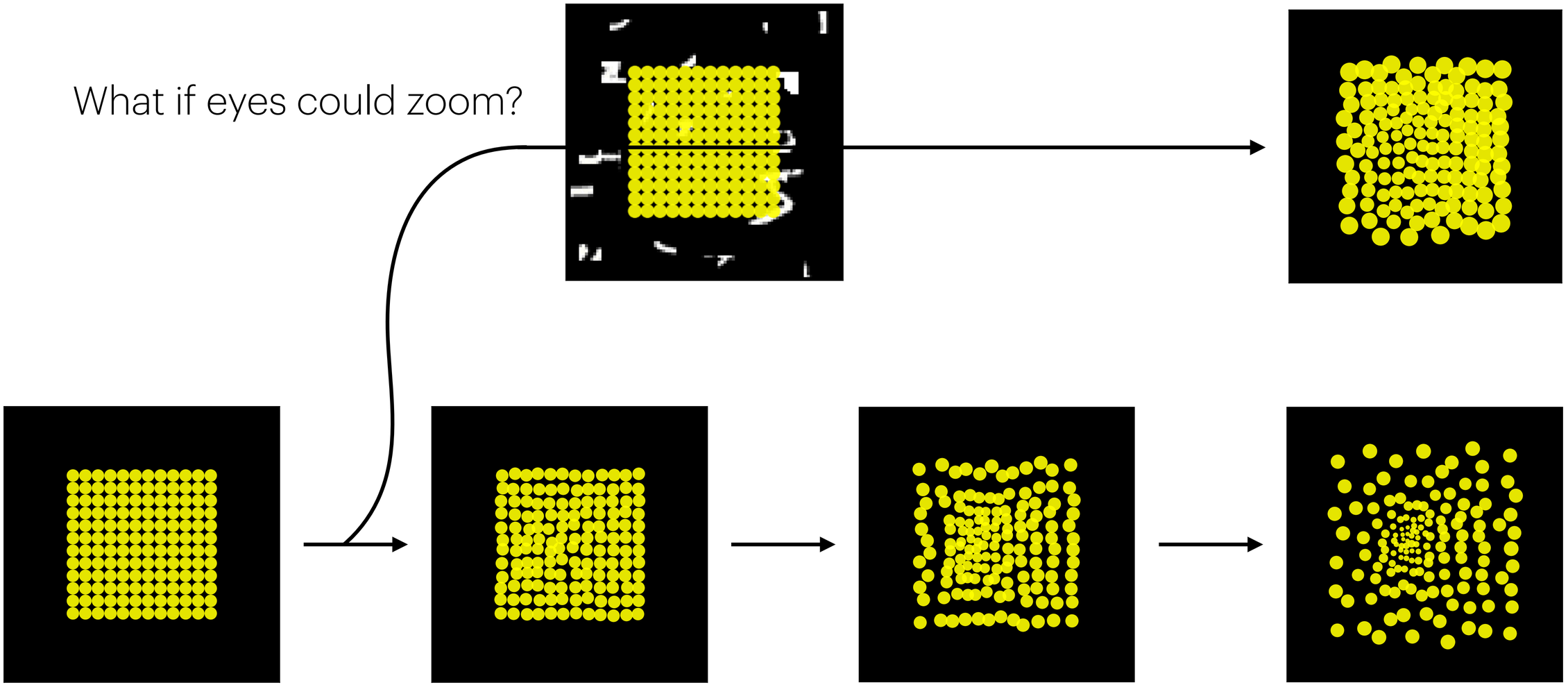


Evolution of a fovea



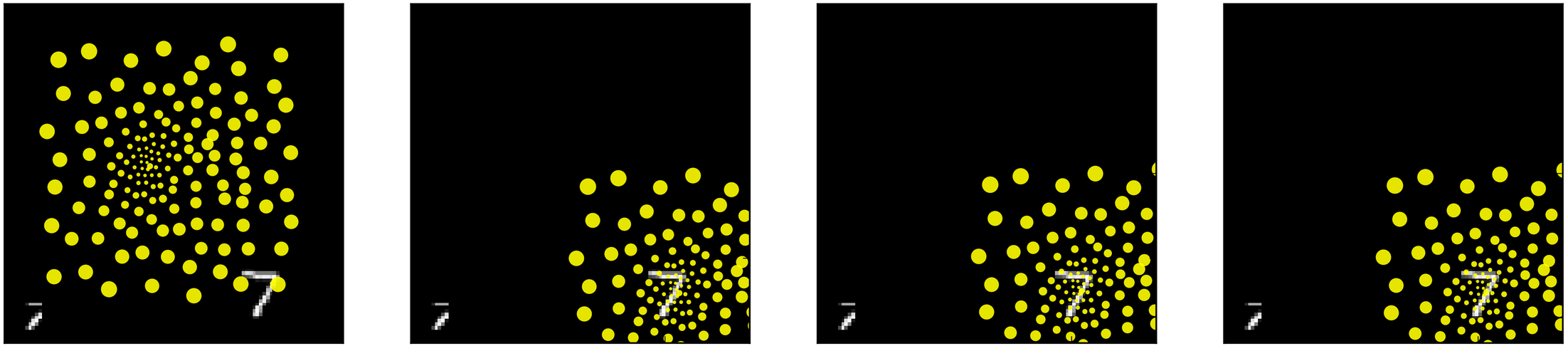
[**Cheung**, Weiss, Olshausen 2017]

What if eyes could zoom?



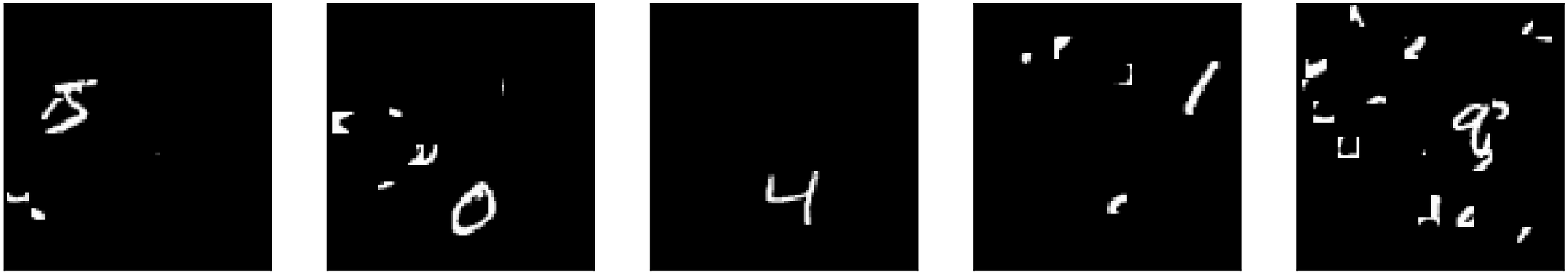
[**Cheung**, Weiss, Olshausen 2017]

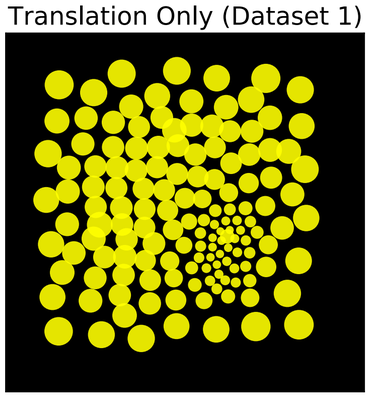
How to use a fovea

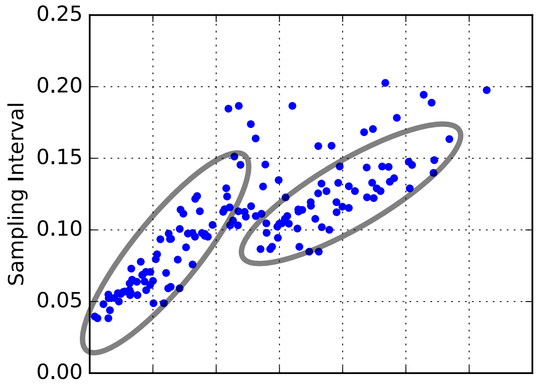


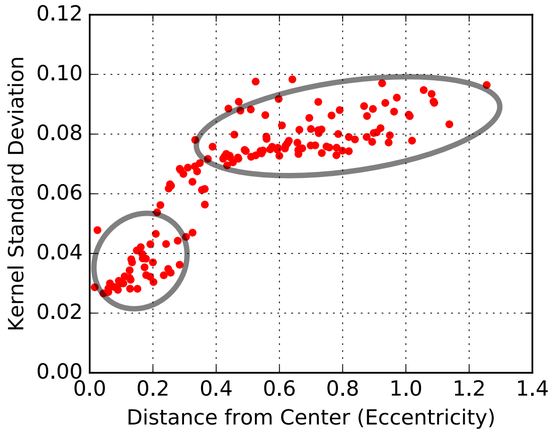
What if the data were different?

Dataset 1



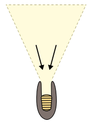

Translation Only (Dataset 1)


Sampling Interval
0.25
0.20
0.15
0.10
0.05
0.00

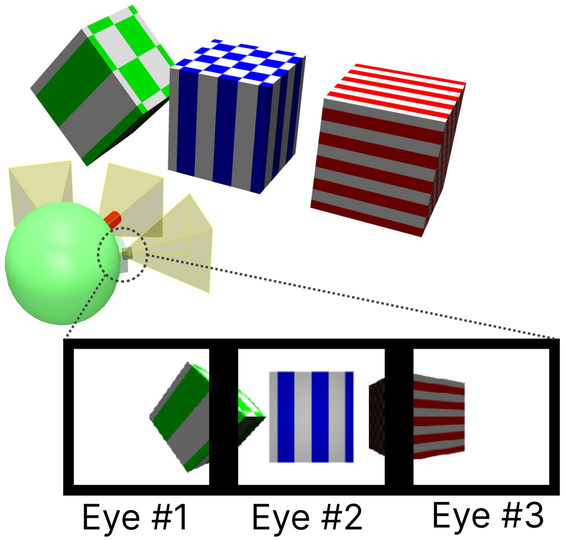

Kernel Standard Deviation
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Distance from Center (Eccentricity)

Building a **‘What if’ machine** for vision

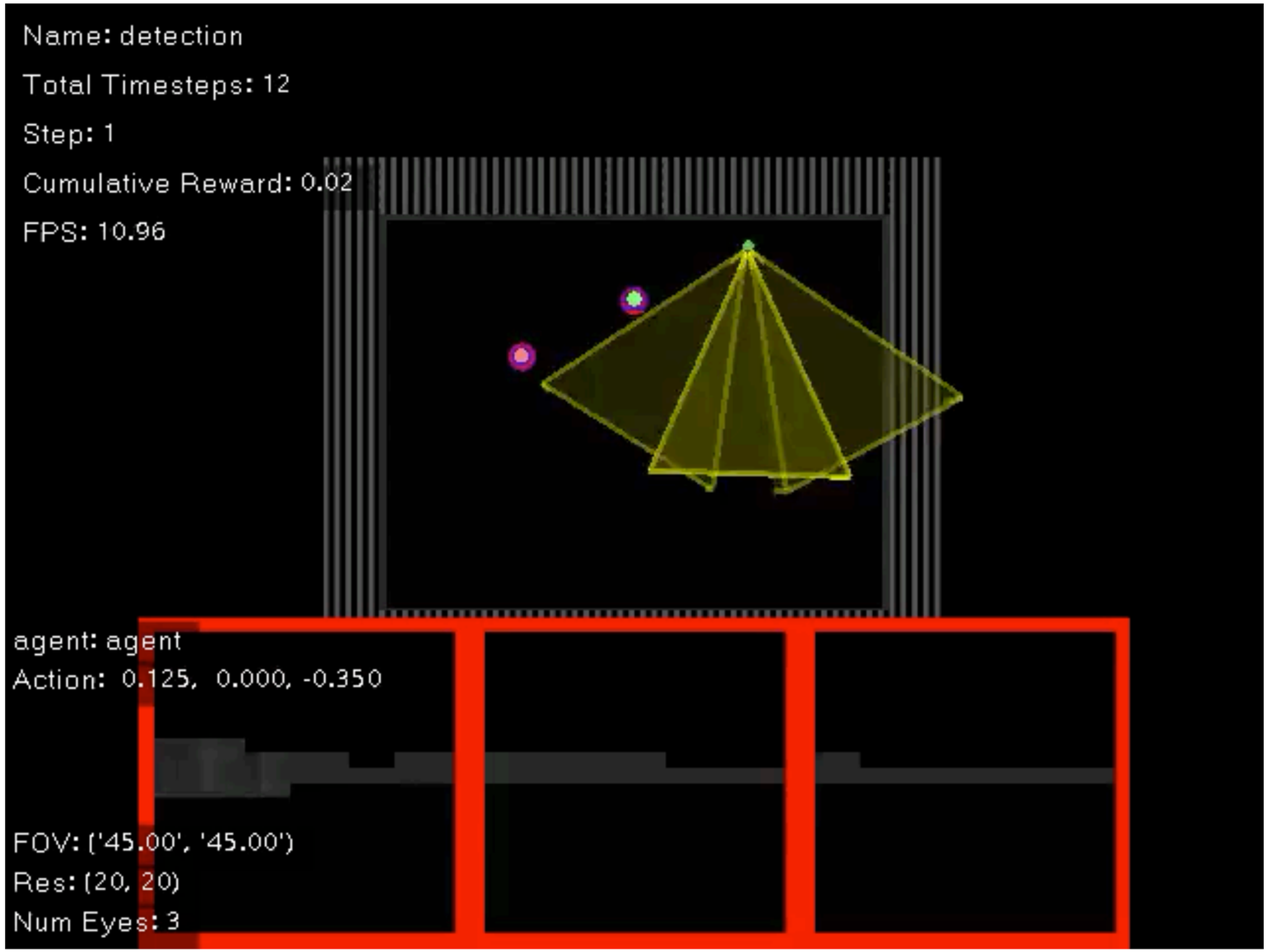
Photoreceptor



Evolution of an “AI-ball”



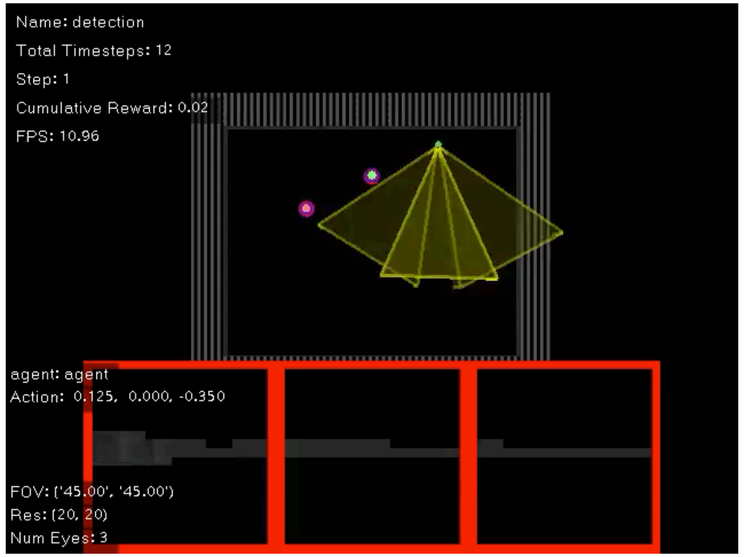
Reinforcement Learning Environments



Reinforcement Learning Environments

Navigation

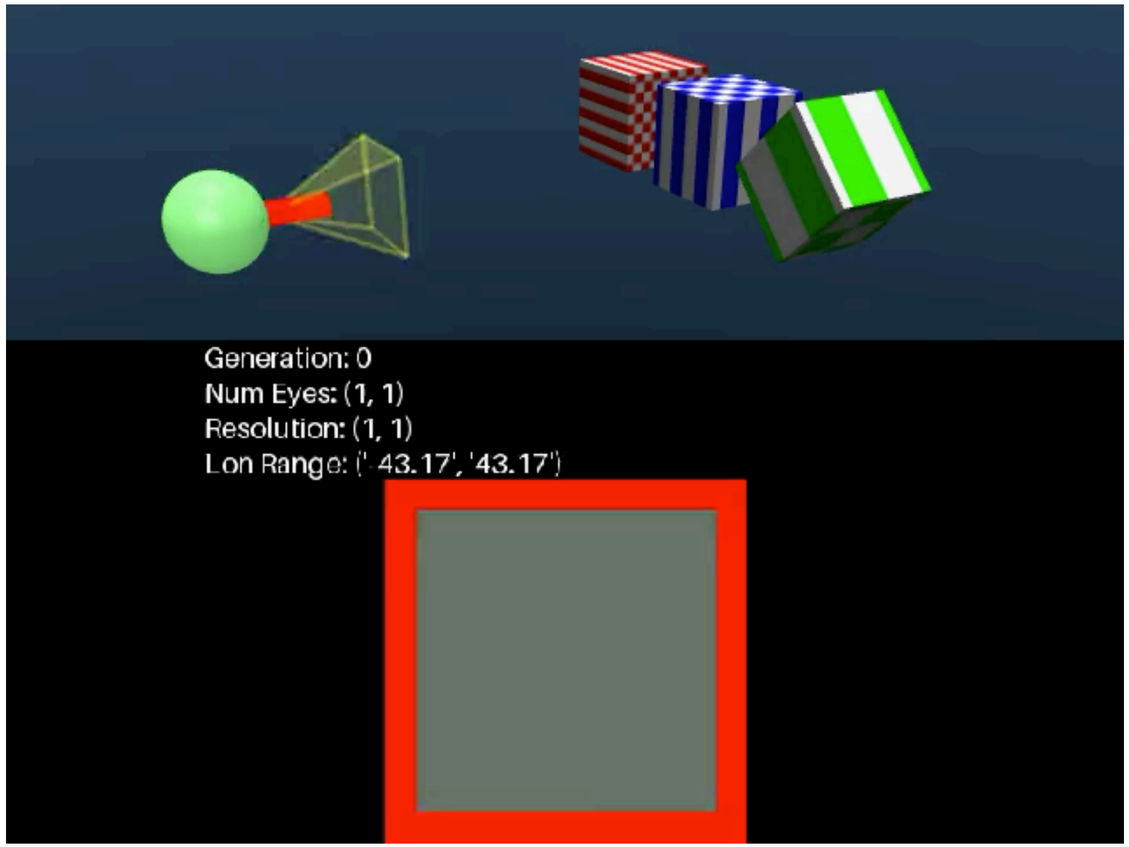
Detection



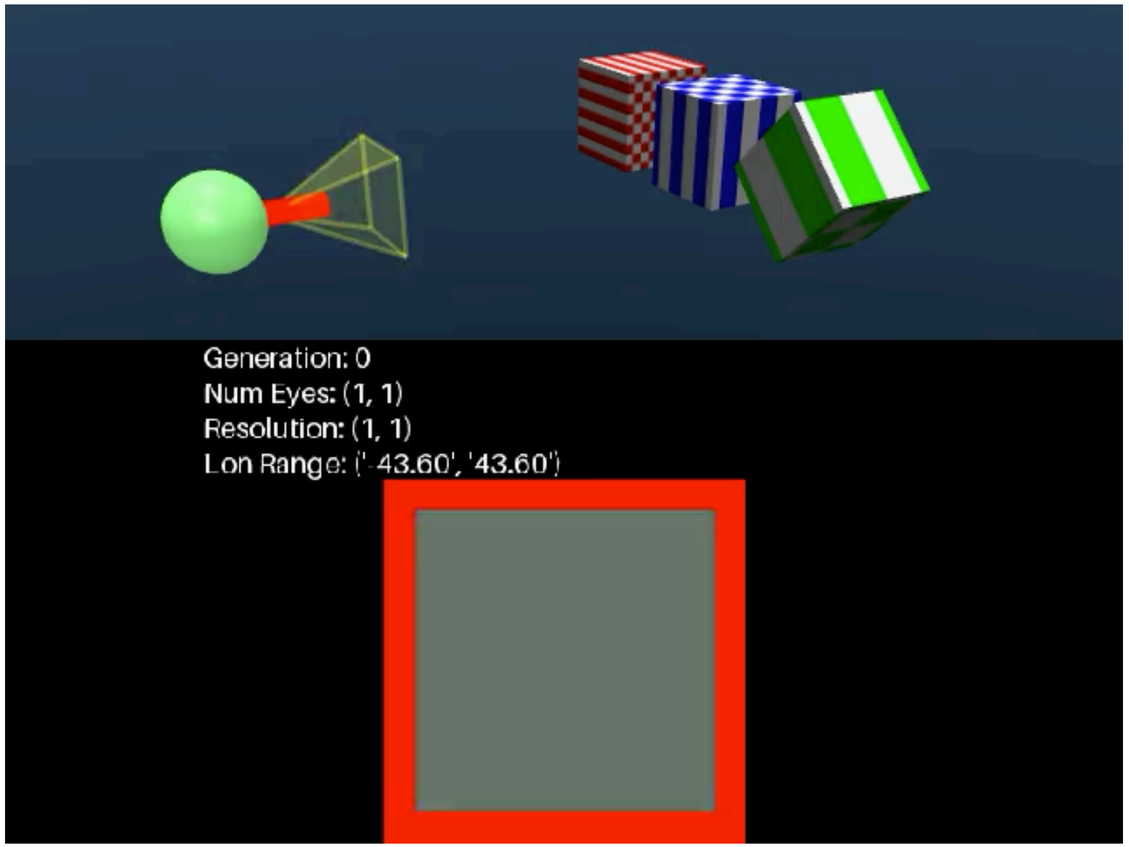
Tracking

What if the ~~data~~ goals were different?

Navigation

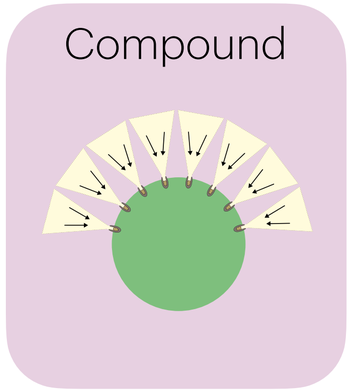


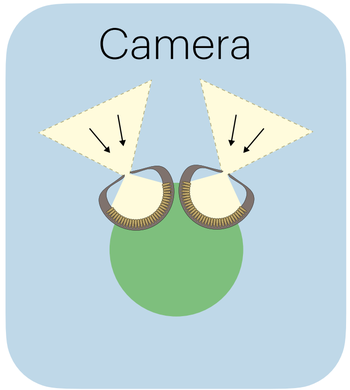
Detection



[Tiwary*, Young*, Klinghoffer, Tasneem, Dave, Poggio, Nilsson, **Cheung****, Raskar** 2025]

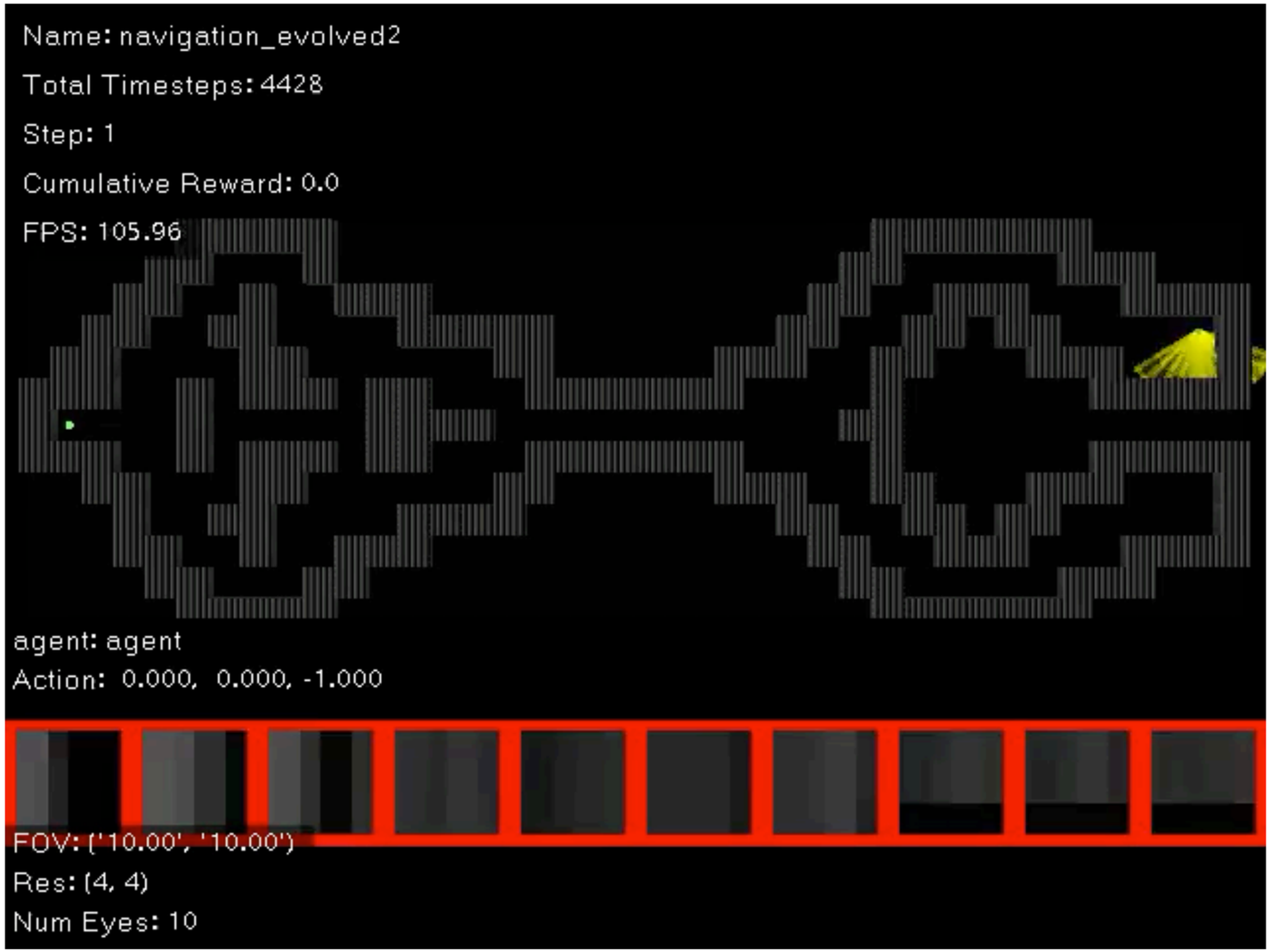
What if the ~~data~~ goals were different?



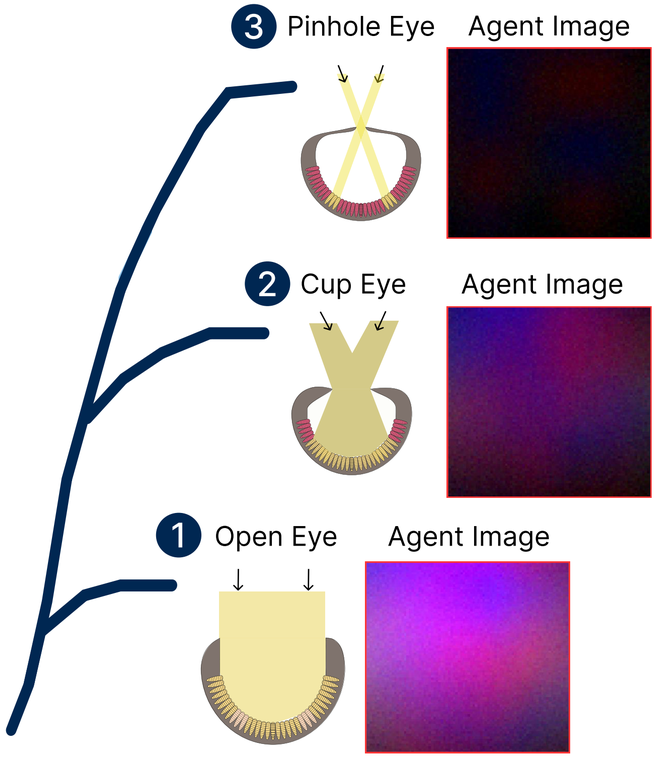


Navigation at Generation 0

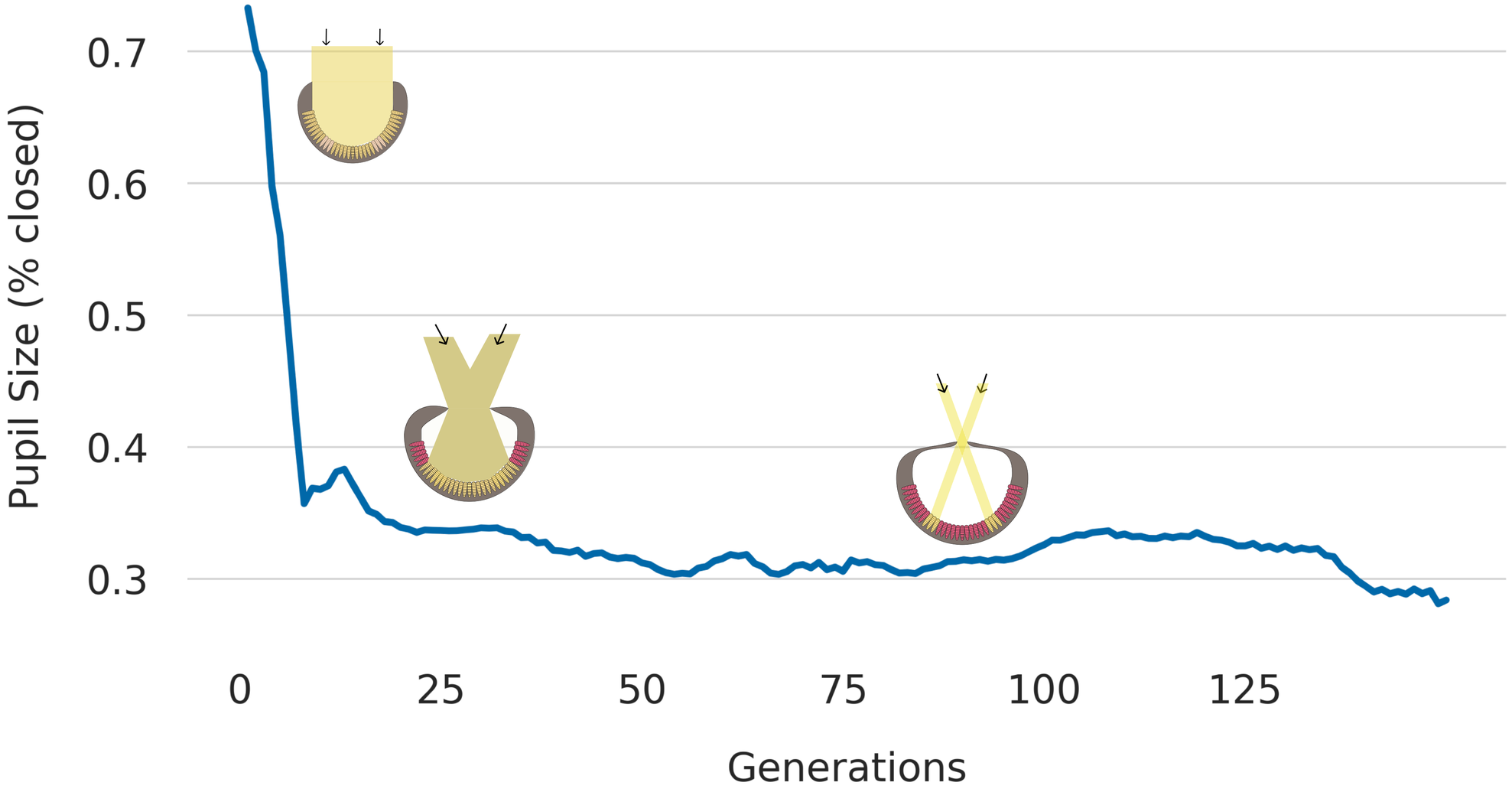
Navigation at Generation 50



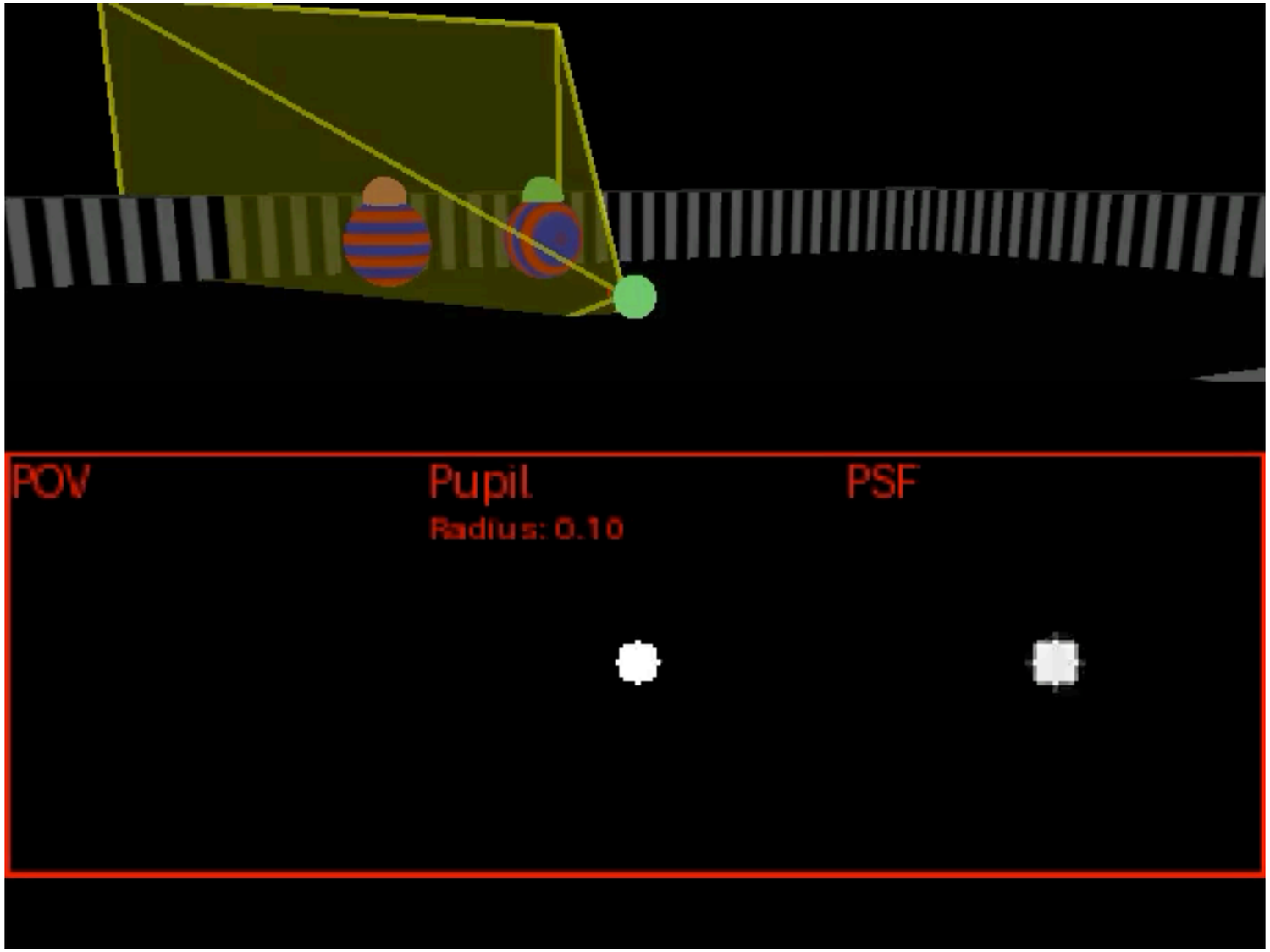
Emergence of Pinhole Eyes



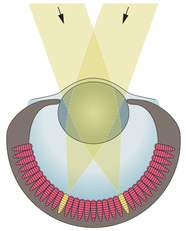
Emergence of Pinhole Eyes



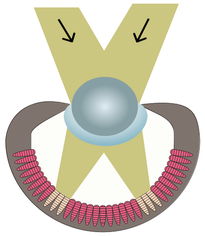
Pinhole Eyes: Acuity at the cost of brightness



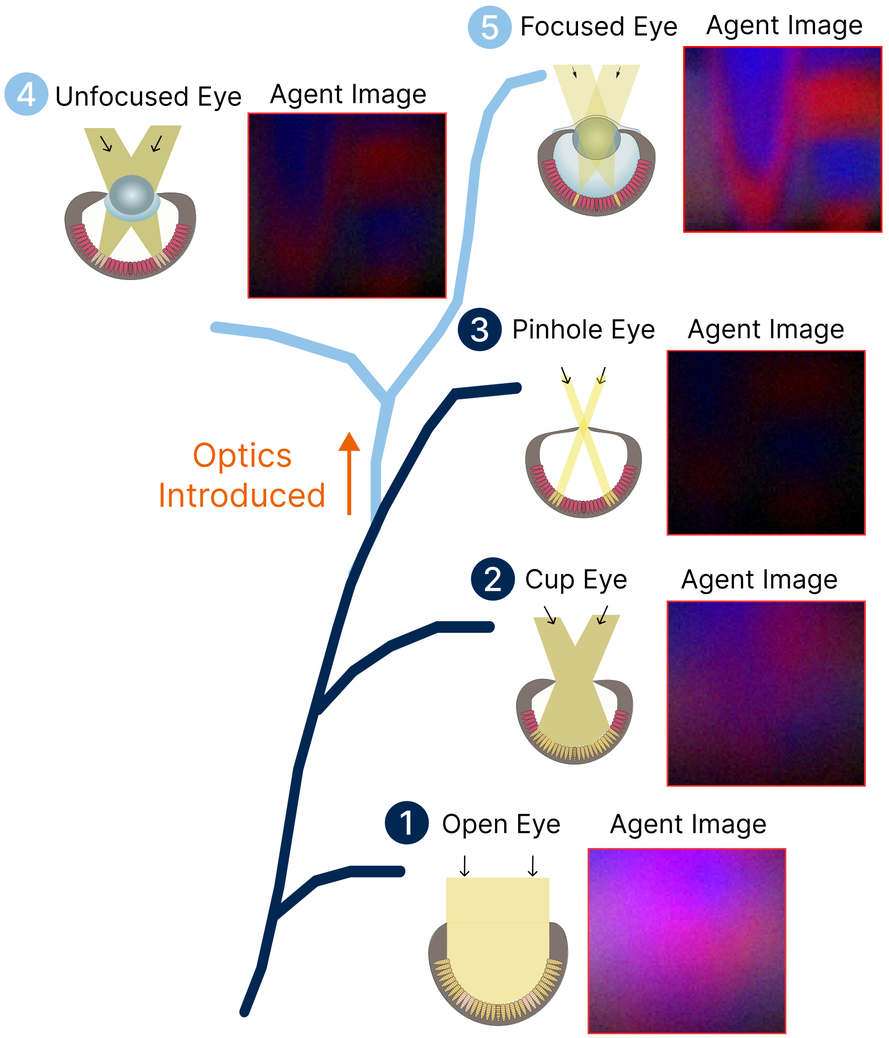
What if eyes could **bend light**?

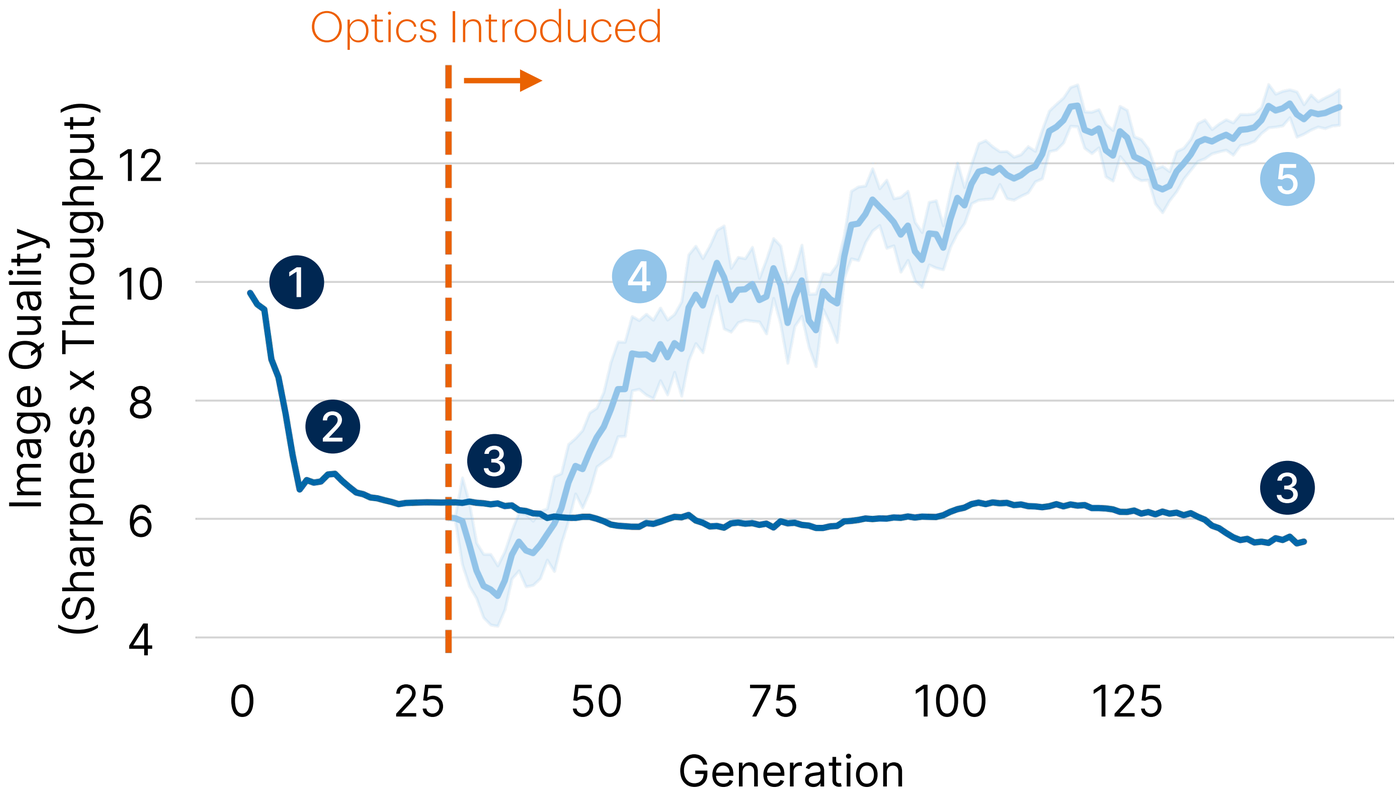


Lens Emerges

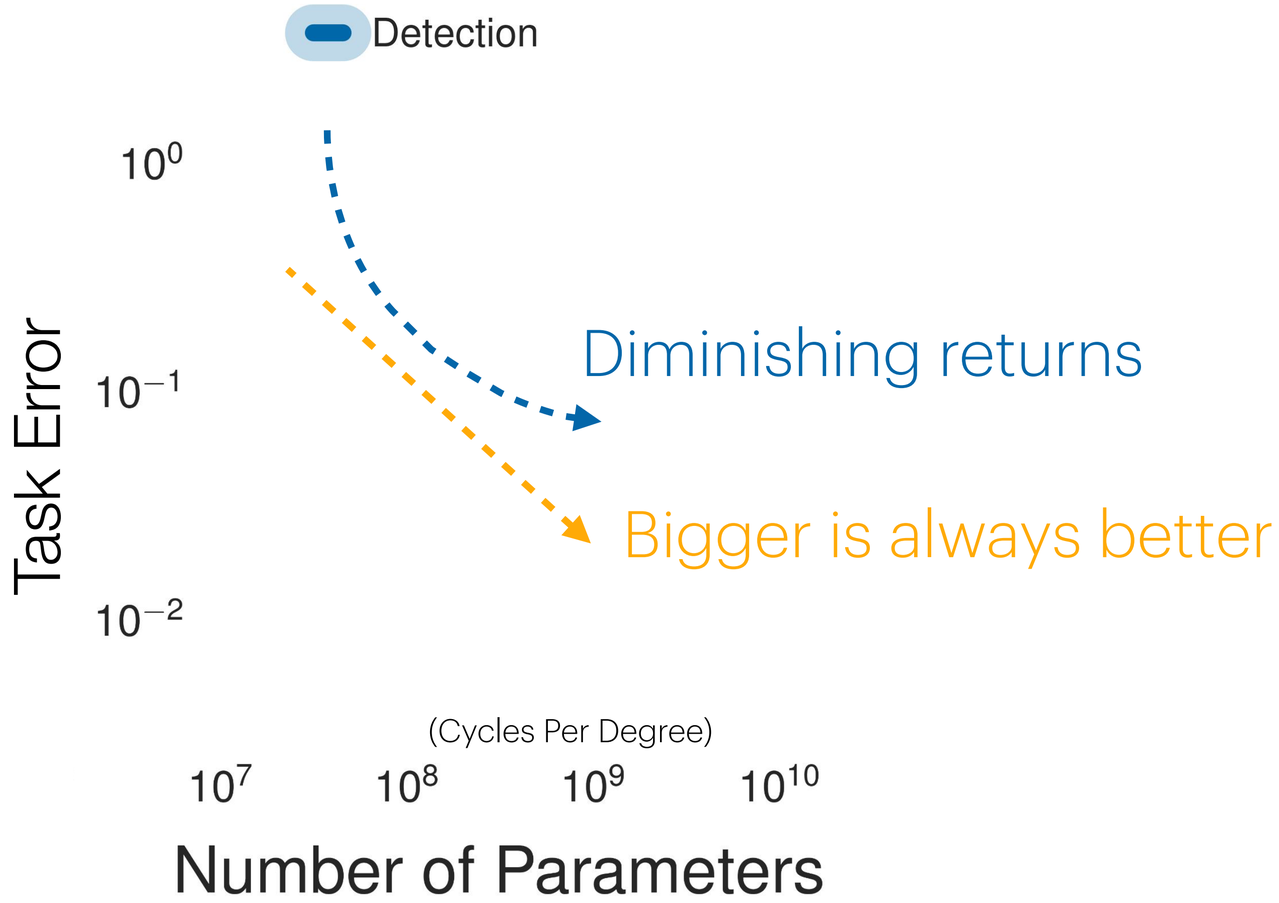


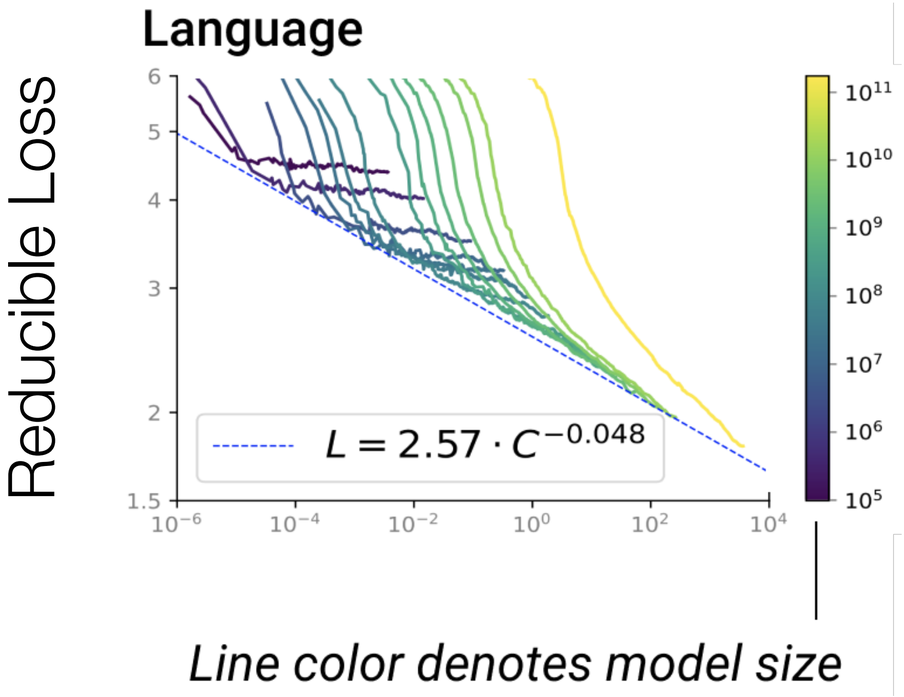
What if eyes could **bend light**?





What if the **brain became larger**?

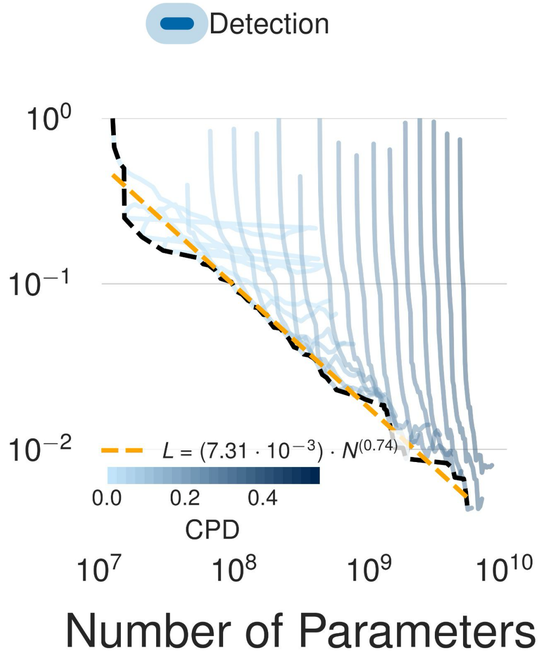




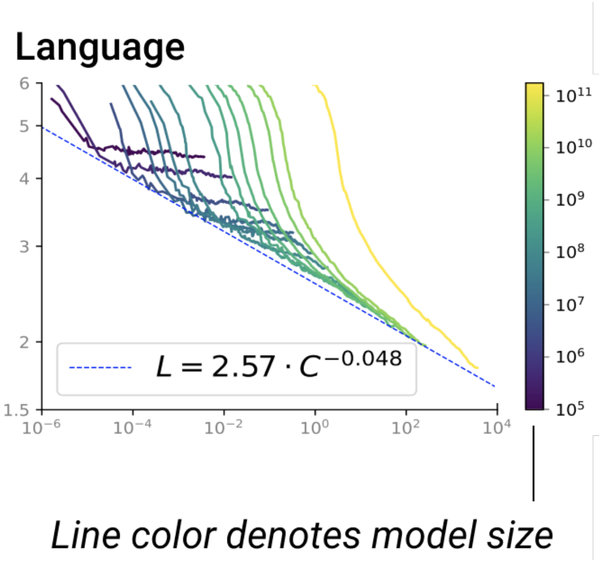
[Henighan*, Kaplan*, Katz* et al. 2020]

Scaling Laws

sim(



,

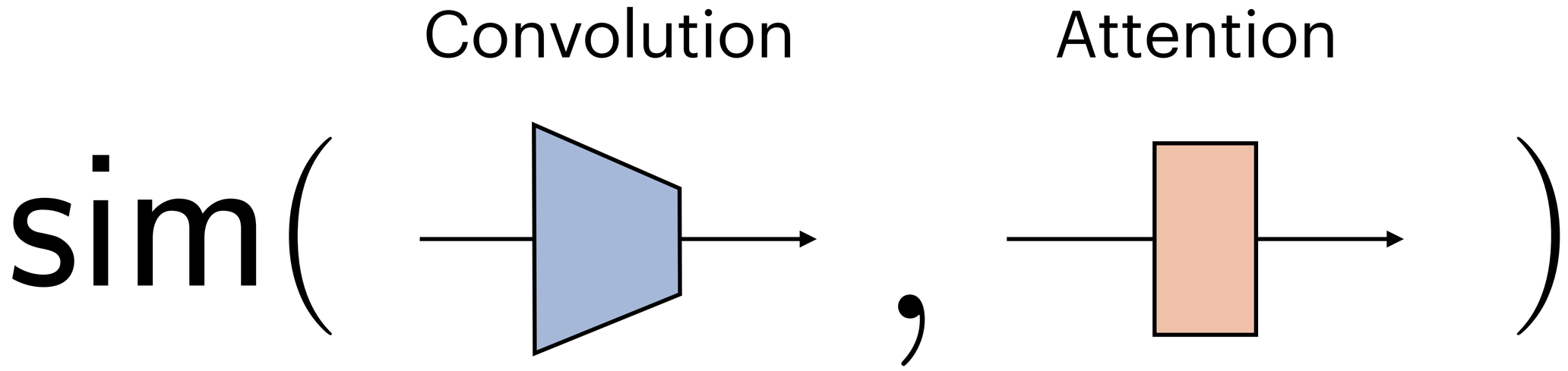


)

K. Tiwary*, A. Young*, T. Klinghoffer, Z. Tasneem, A. Dave, T. Poggio, D.-E. Nilsson

B. Cheung**, R. Raskar**

“What if Eye...? Computationally Recreating Vision Evolution”

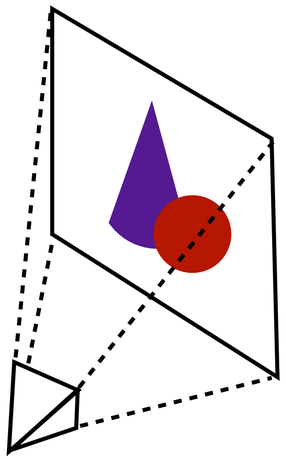


Y. Han, T. Poggio, **B. Cheung**

“System identification of neural systems: If we got it right, would we know?”

Vision

Language

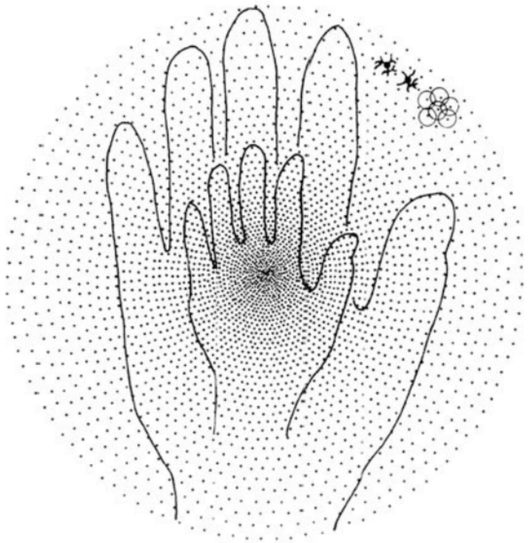
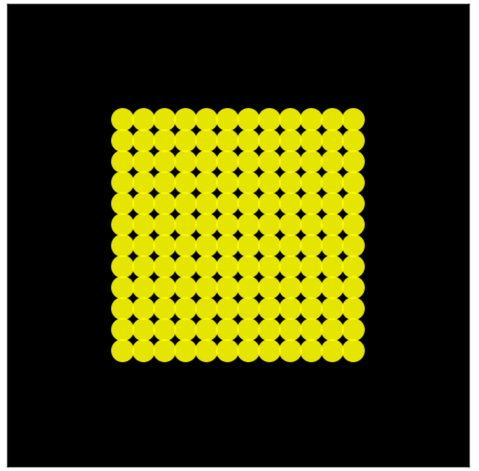
sim( , A red sphere next to a blue cone.)

M. Huh*, **B. Cheung***, T. Wang*, P. Isola*

“The Platonic Representation Hypothesis”

In-vivo

In-silico

sim( , )

B. Cheung, E.A. Weiss, B.A. Olshausen

“Emergence of foveal image sampling from learning to attend in visual scenes”

Future Work

What is alignment?

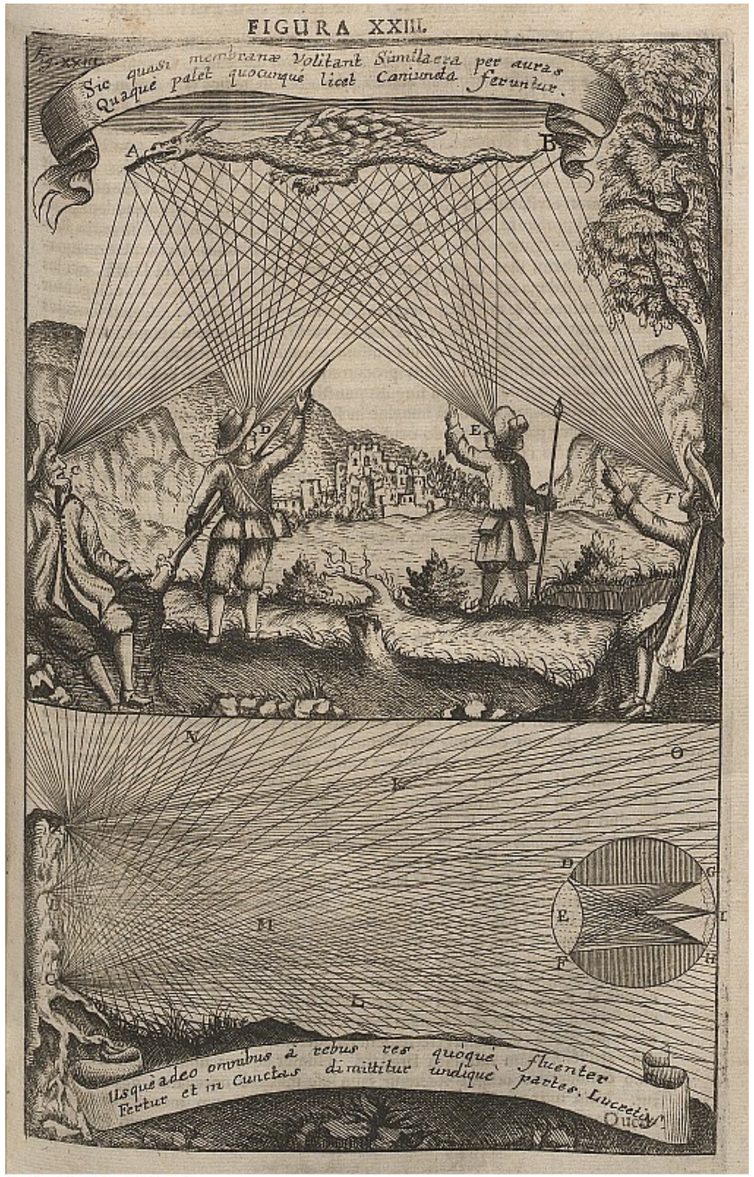
A consistent representation across all ~~modalities.~~ projections of the real world.

What is science?

The search for a **consistent** representation of the real world.

What if we **optimized for alignment** to find consistent representations?

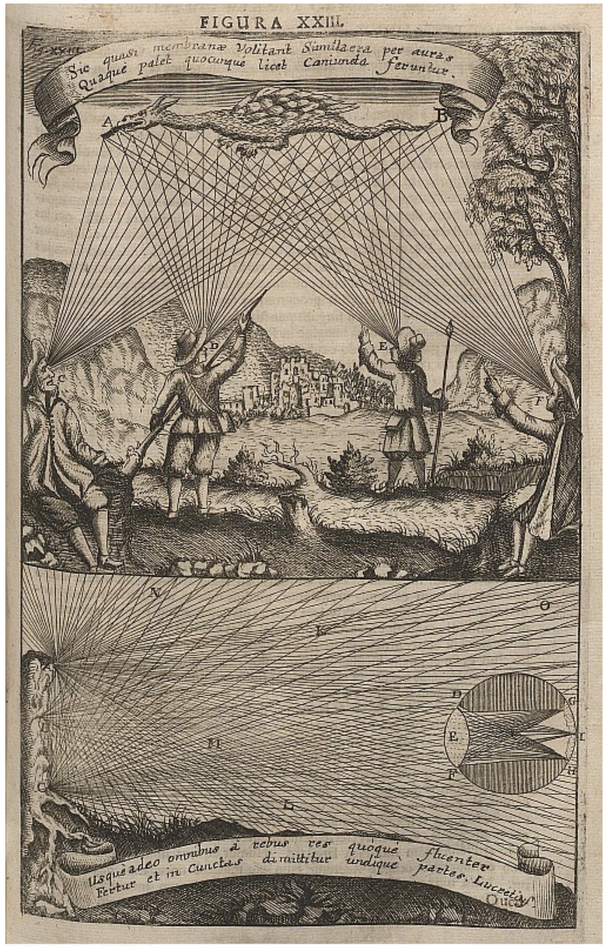
Plato's hypothesis for eyes



Quelle: Deutsche Fotothek

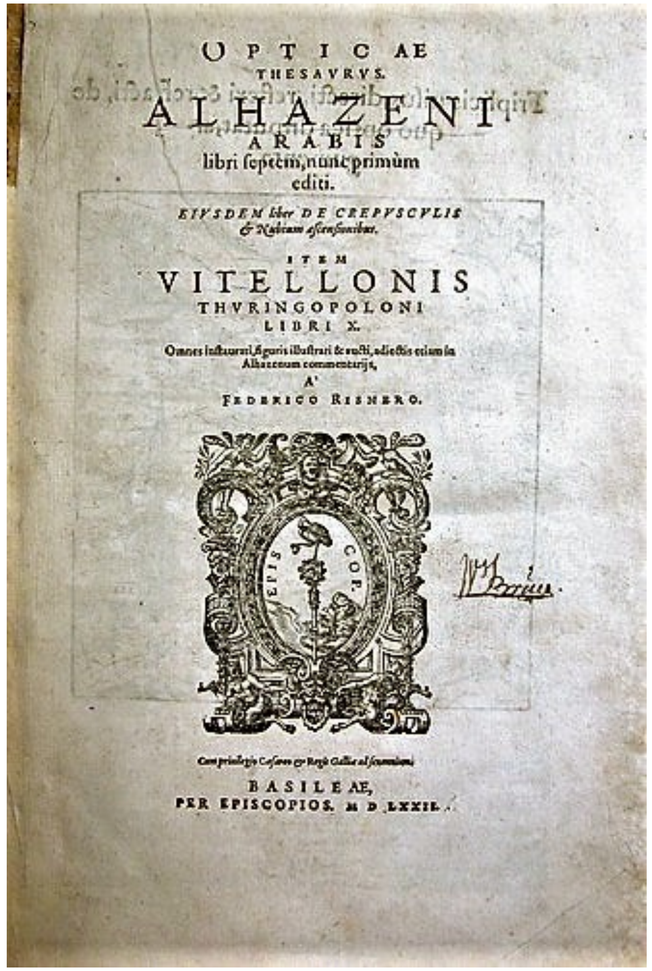
Is my hypothesis **aligned** with my observation?

Emission Theory

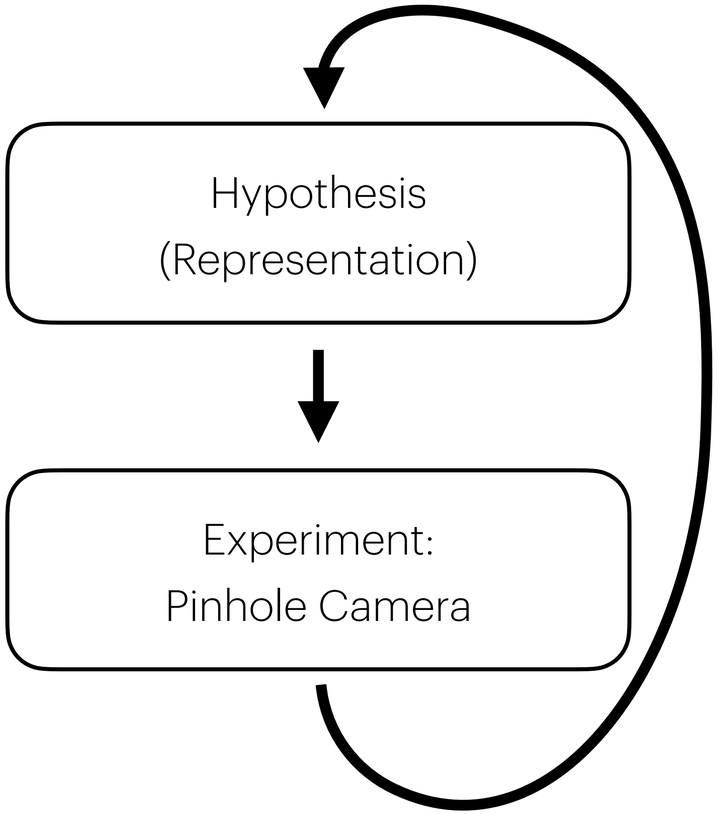


Quelle: Deutsche Fotothek

Intromission Theory

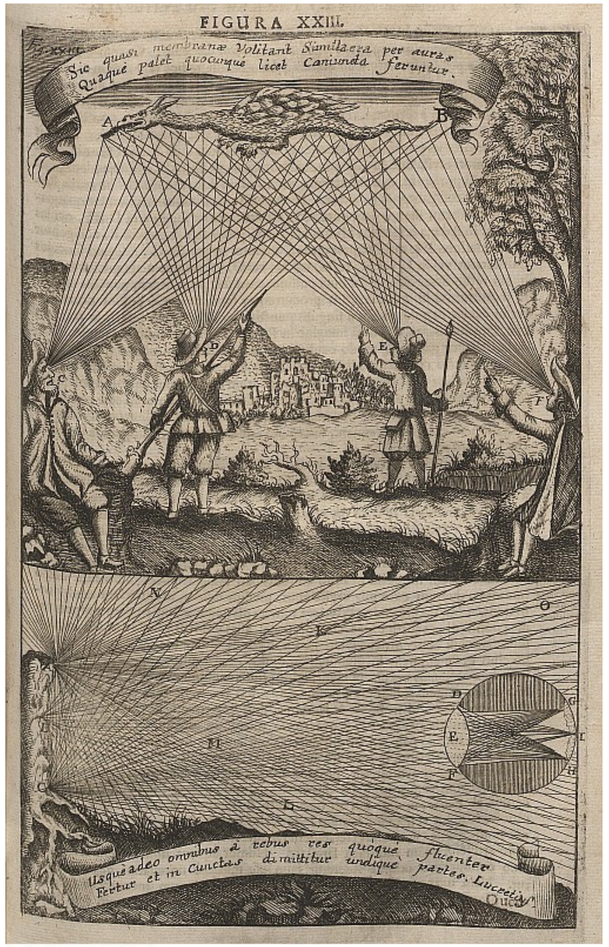

OPTICAE
THESAVRVS.
ALHAZENI
ARABIS
libri septem, nunc primùm
editi.
EIVSDEM liber DE CREPVSCVLIS
& Nubium ascensionibus.
ITEM
VITELLONIS
THVRINGOPOLONI
LIBRI X.
Omnes instaurati, figuris illustrati & multis, adiectis etiam in
Alhazenum commentarijs,
A'
FEDERICO RISNERO.
BASILEAE,
PER EPISCOPIOS. M D LXXII.

[Alhazen 1021]



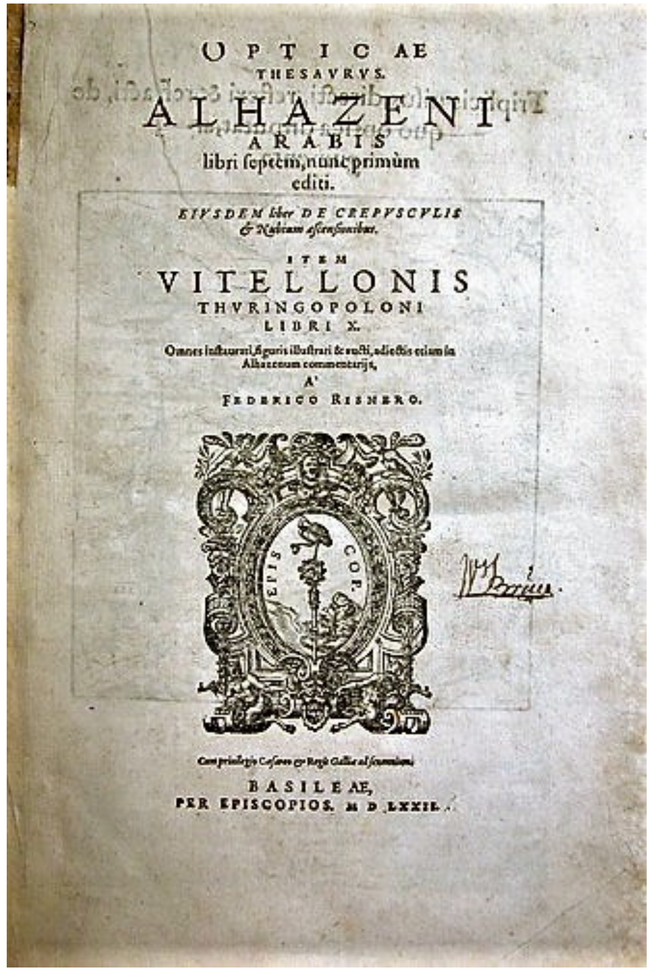
Is my hypothesis **aligned** with my observation?

Emission Theory

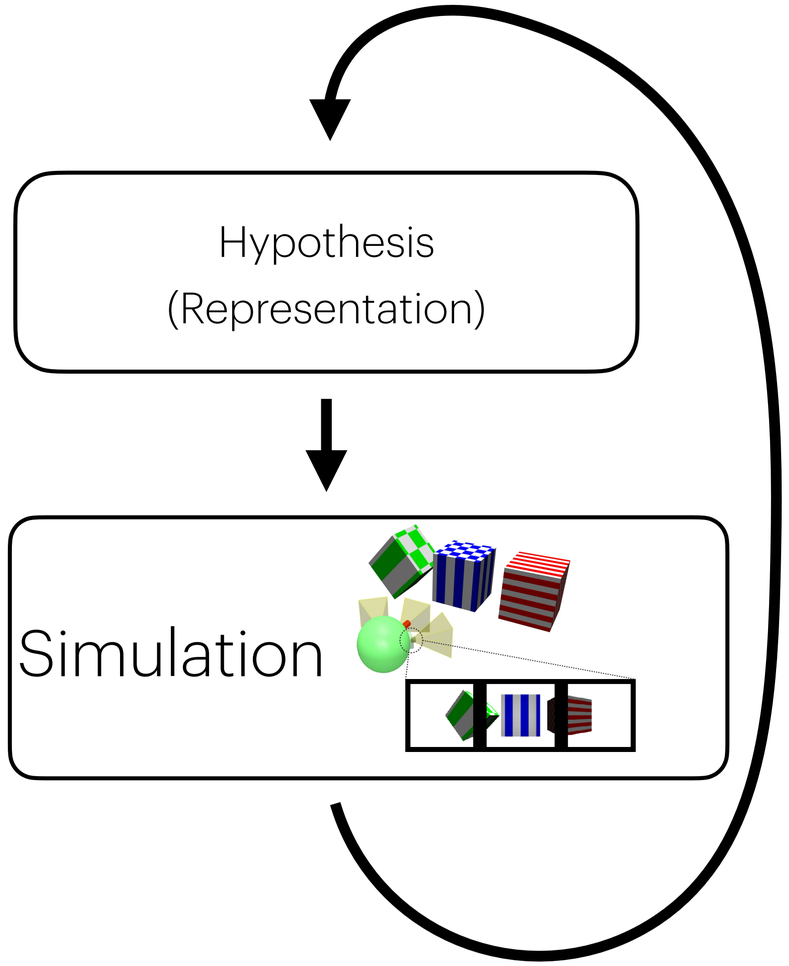


Quelle: Deutsche Fotothek

Intromission Theory

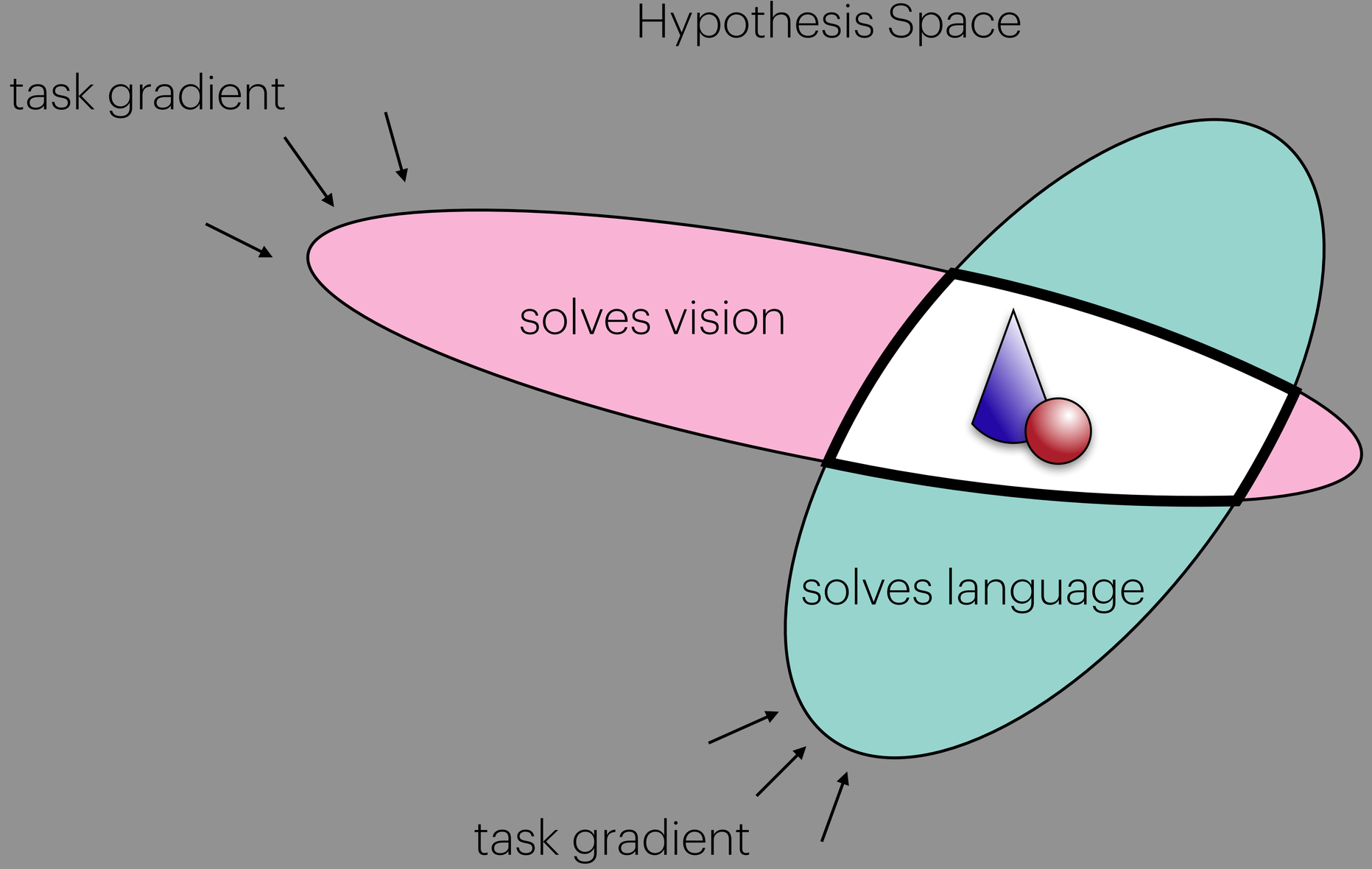

OPTICAE
THESAVRVS.
ALHAZENI
ARABIS
libri septem, nunc primùm editi.
EIVSDEM liber DE CREPVSCVLIS & Nubium ascensionibus.
ITEM
VITELLONIS
THVRINGOPOLONI
LIBRI X.
Omnes instaurati, figuris illustrati & multis, adiectis etiam in Alhazenum commentarijs,
A'
FEDERICO RISNERO.
BASILEAE,
PER EPISCOPIOS. M D LXXII.

[Alhazen 1021]

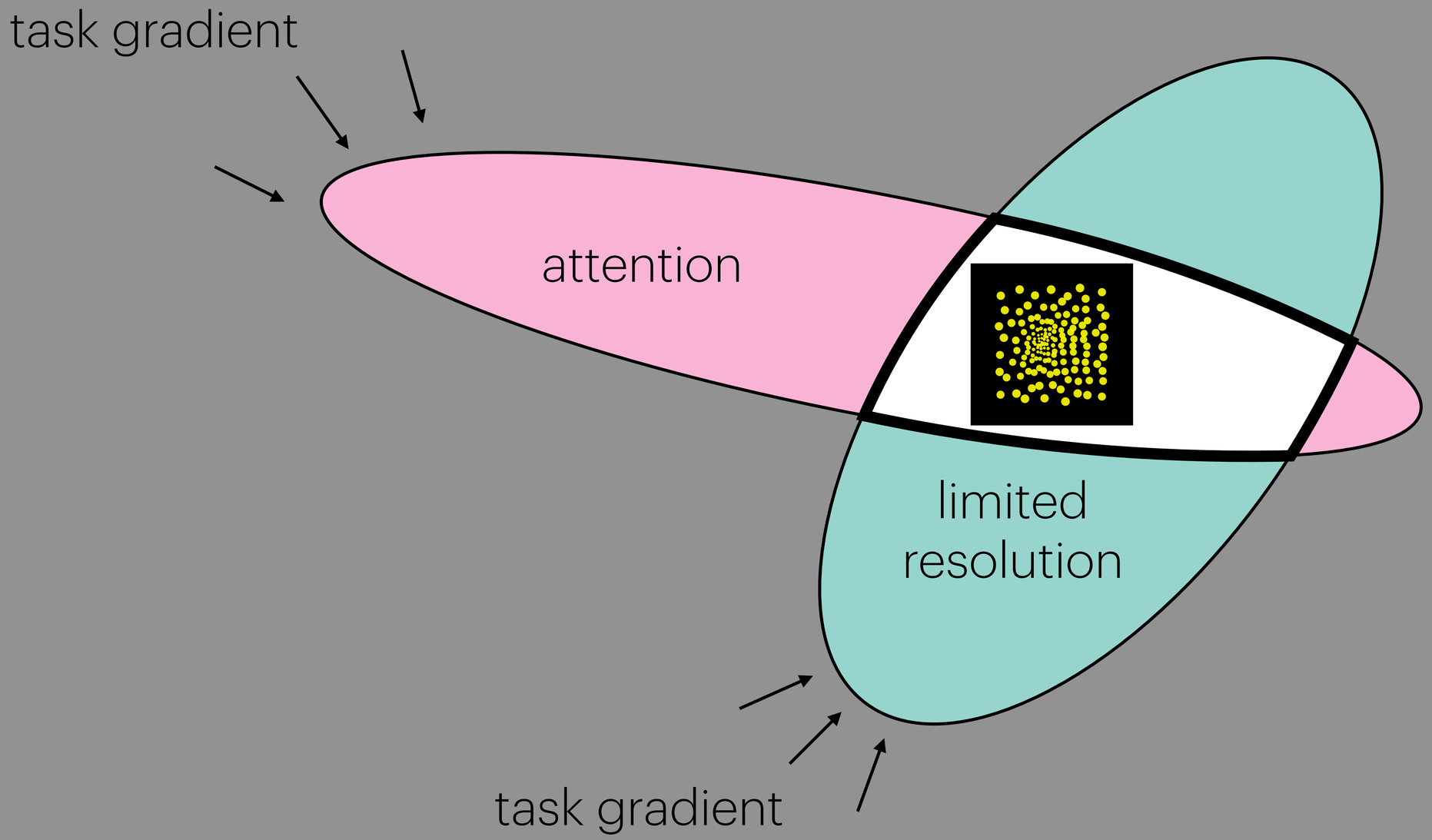


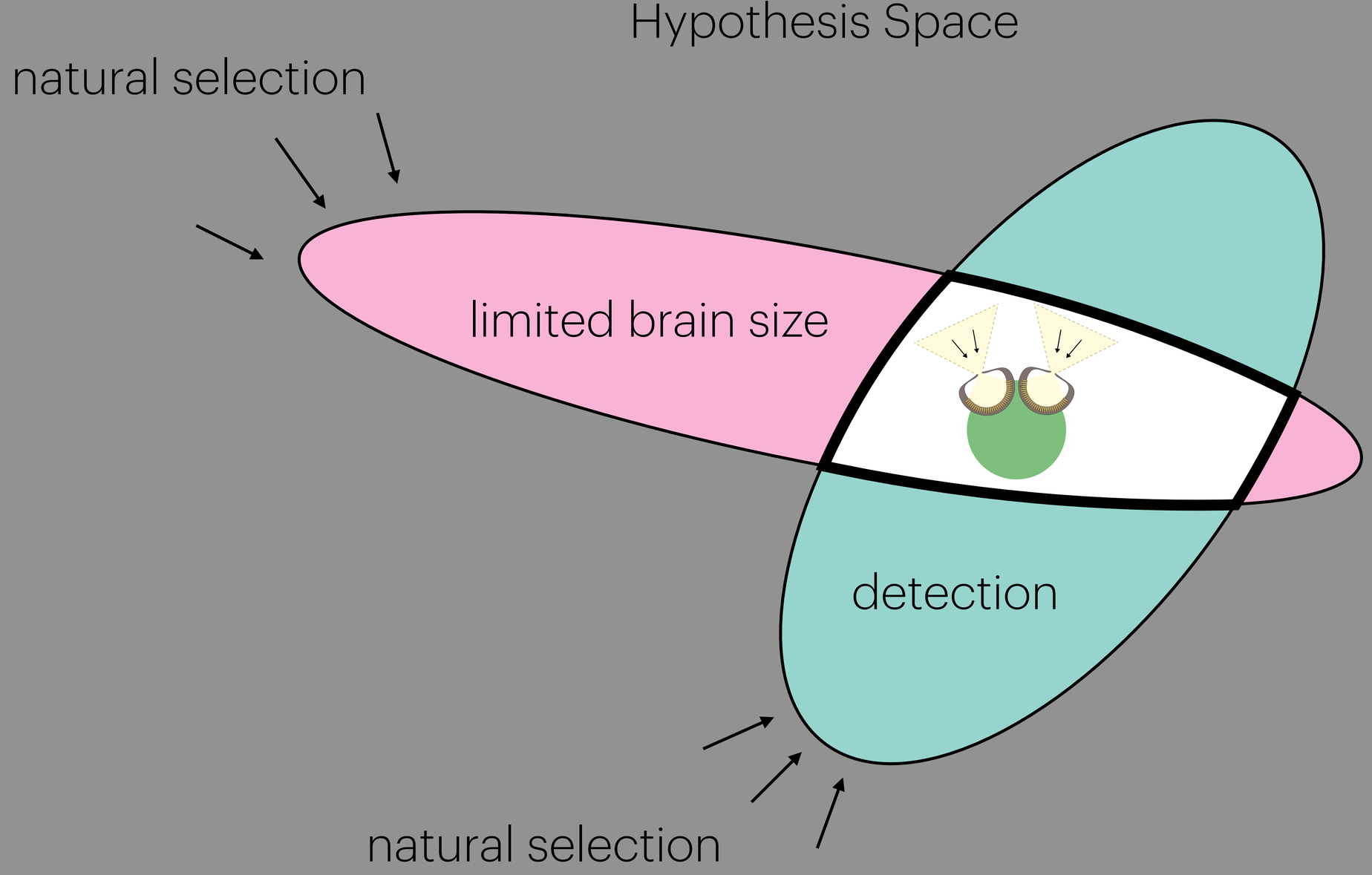
Alhazen (965 - 1040 CE)

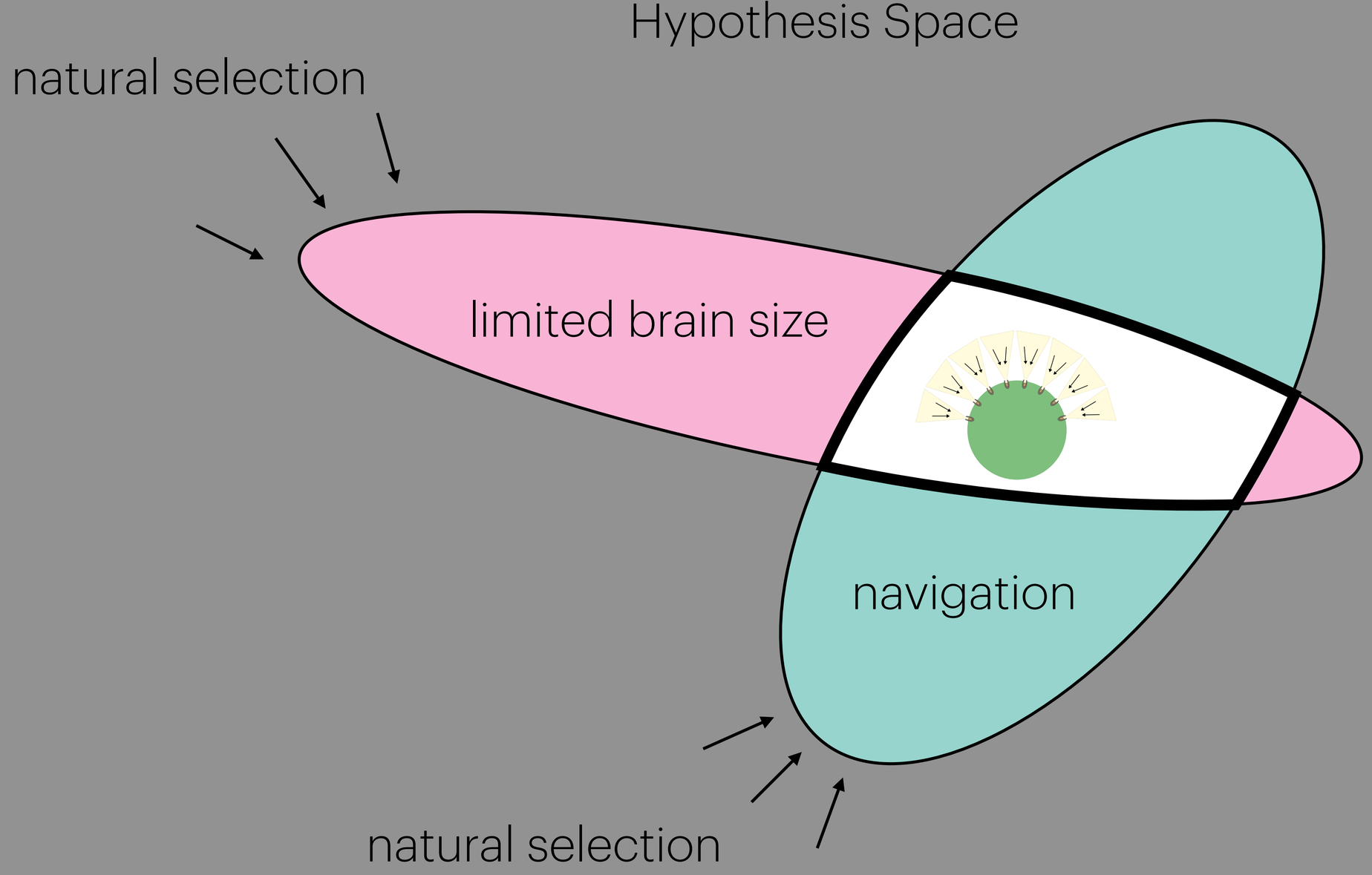
“The duty of the man who investigates the writings of scientists, if learning the truth is his goal, is to make himself an enemy of all that he reads, and ... attack it from every side.”

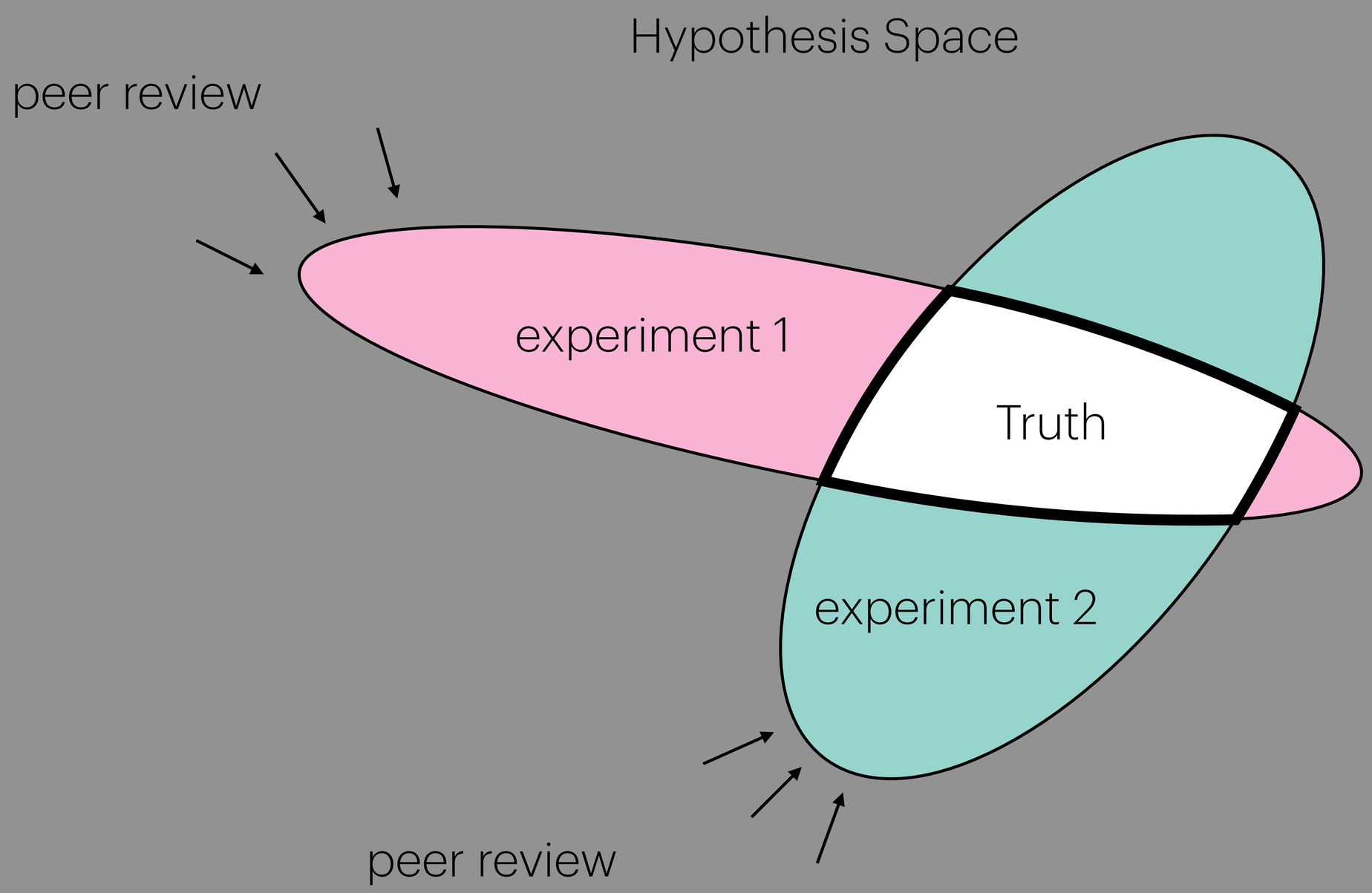

Hypothesis Space
task gradient
solves vision
solves language
task gradient

Hypothesis Space



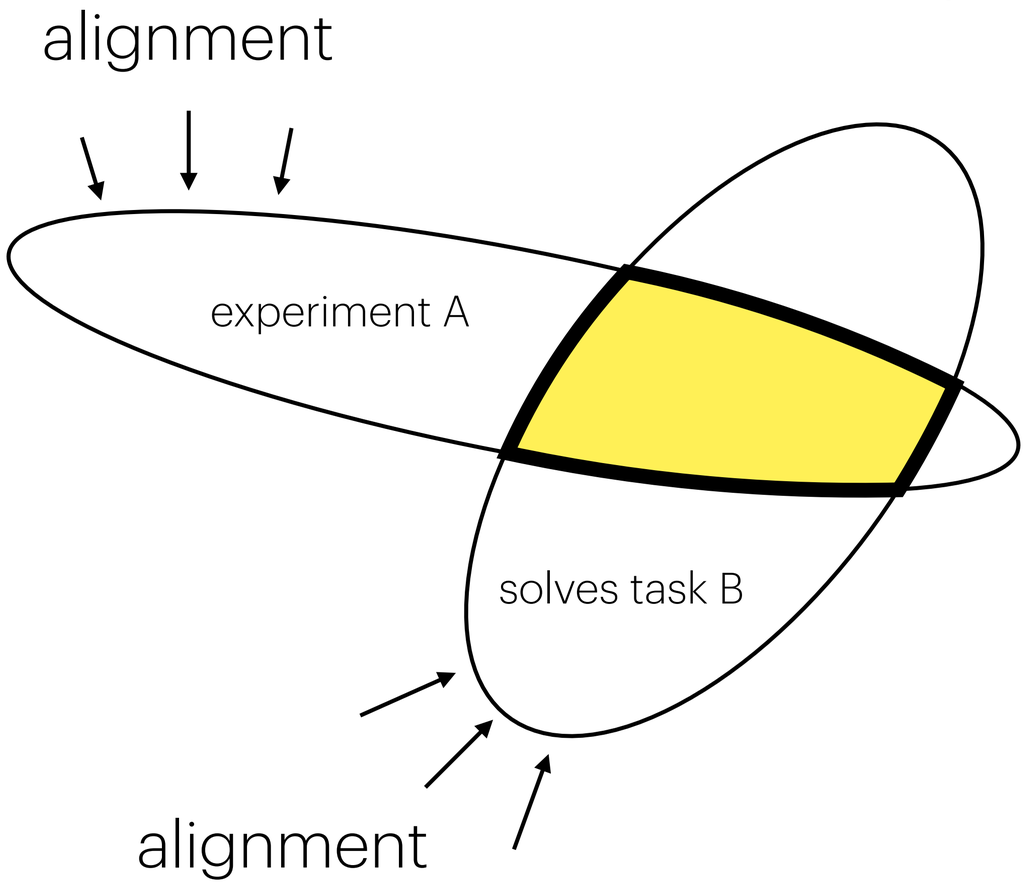

Hypothesis Space
natural selection
limited brain size
detection
natural selection


Hypothesis Space
natural selection
limited brain size
navigation
natural selection


Hypothesis Space
peer review
experiment 1
Truth
experiment 2
peer review

If you can **embed it**, you can **align it**!

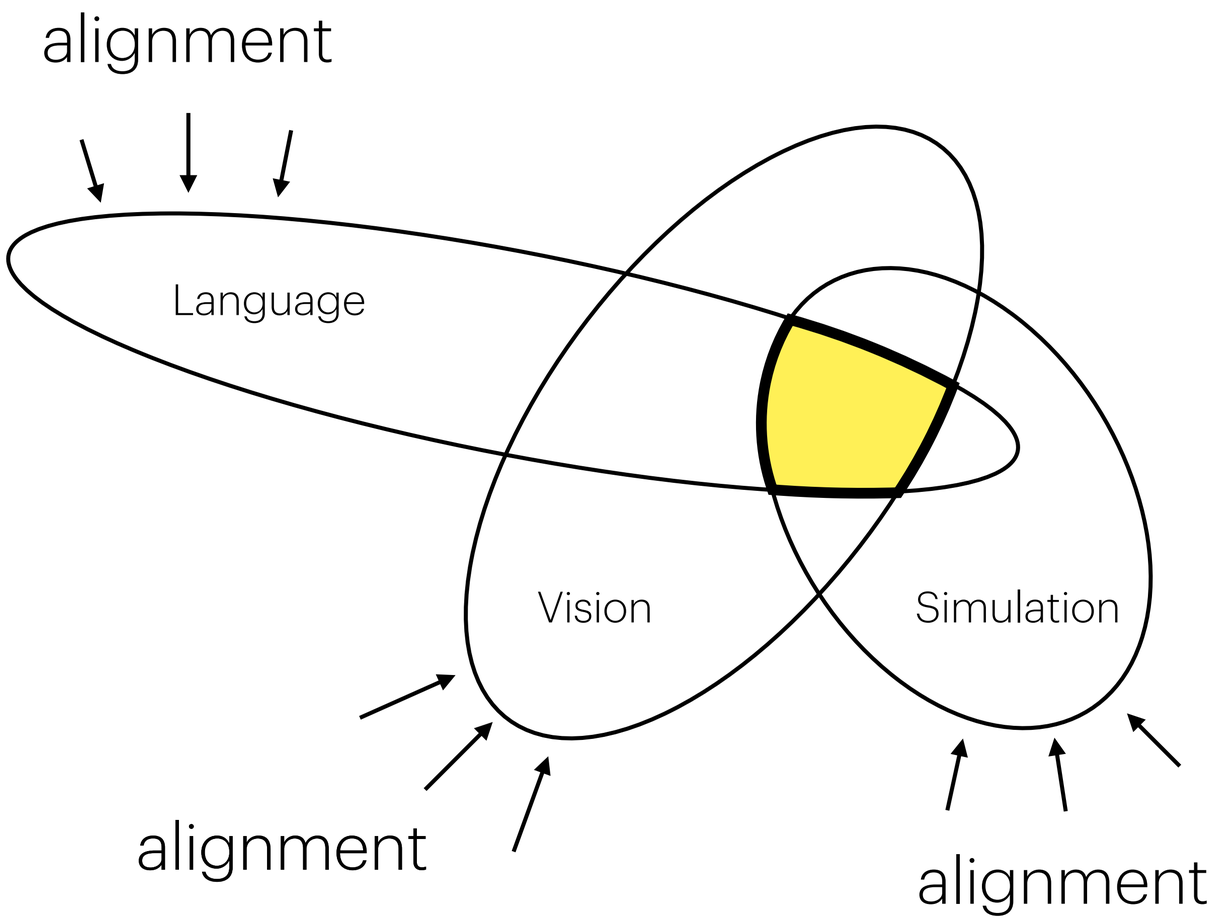
$$f : \mathcal{X} \to \mathbb{R}^n$$

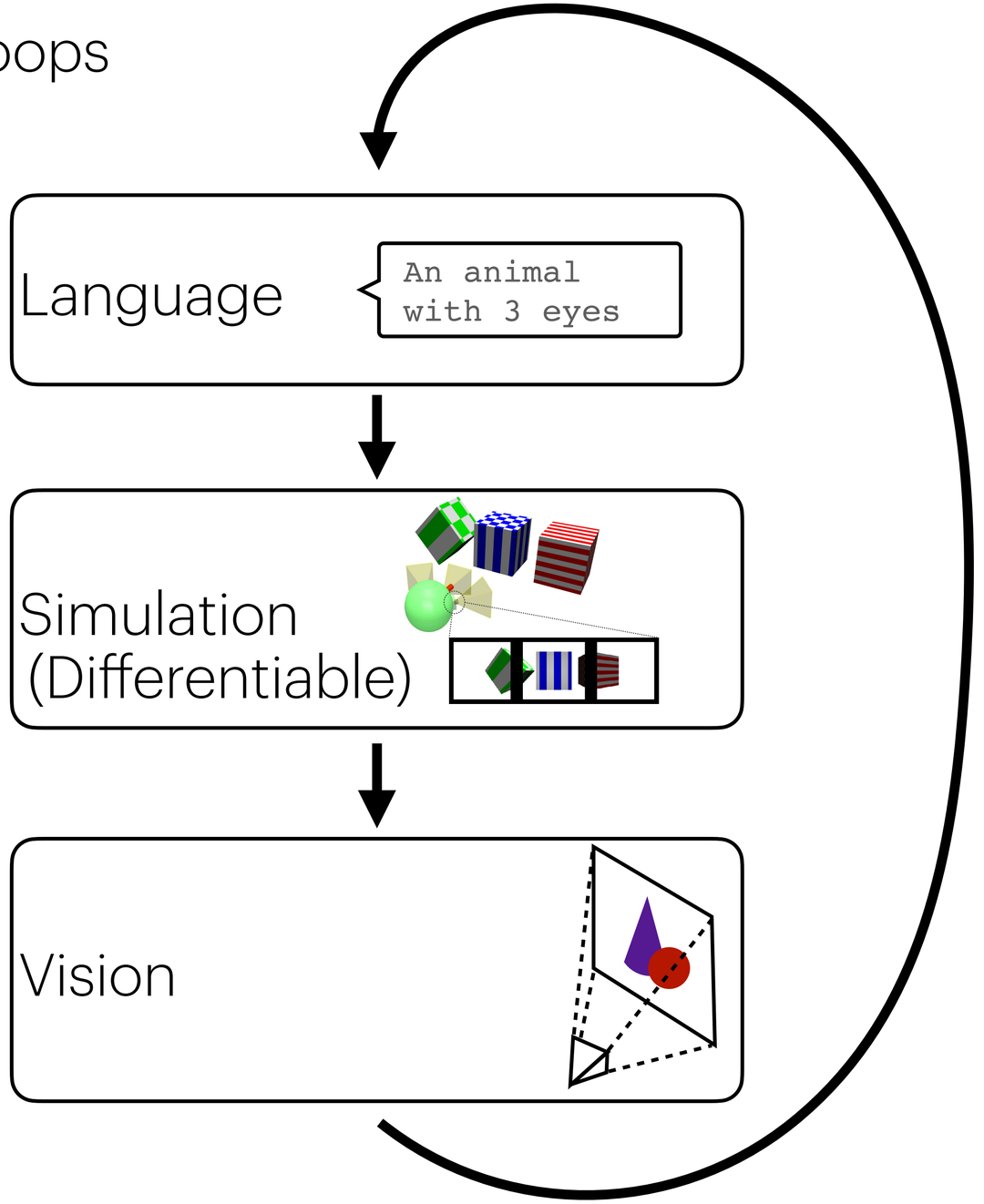


- Experiments
- Modalities
- Neural network representations

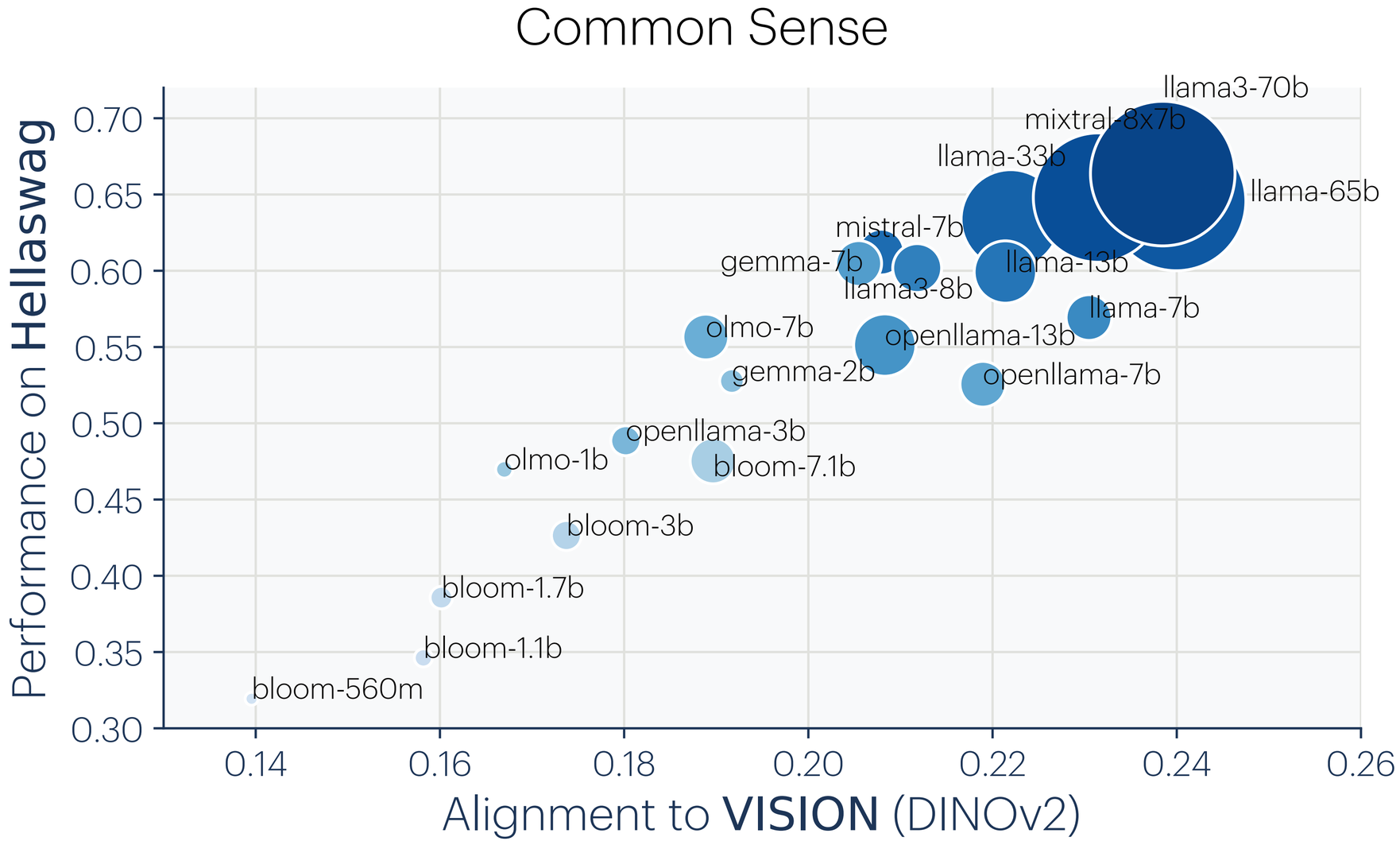
become **interchangeable** and **alignable.**

Alignment Loops

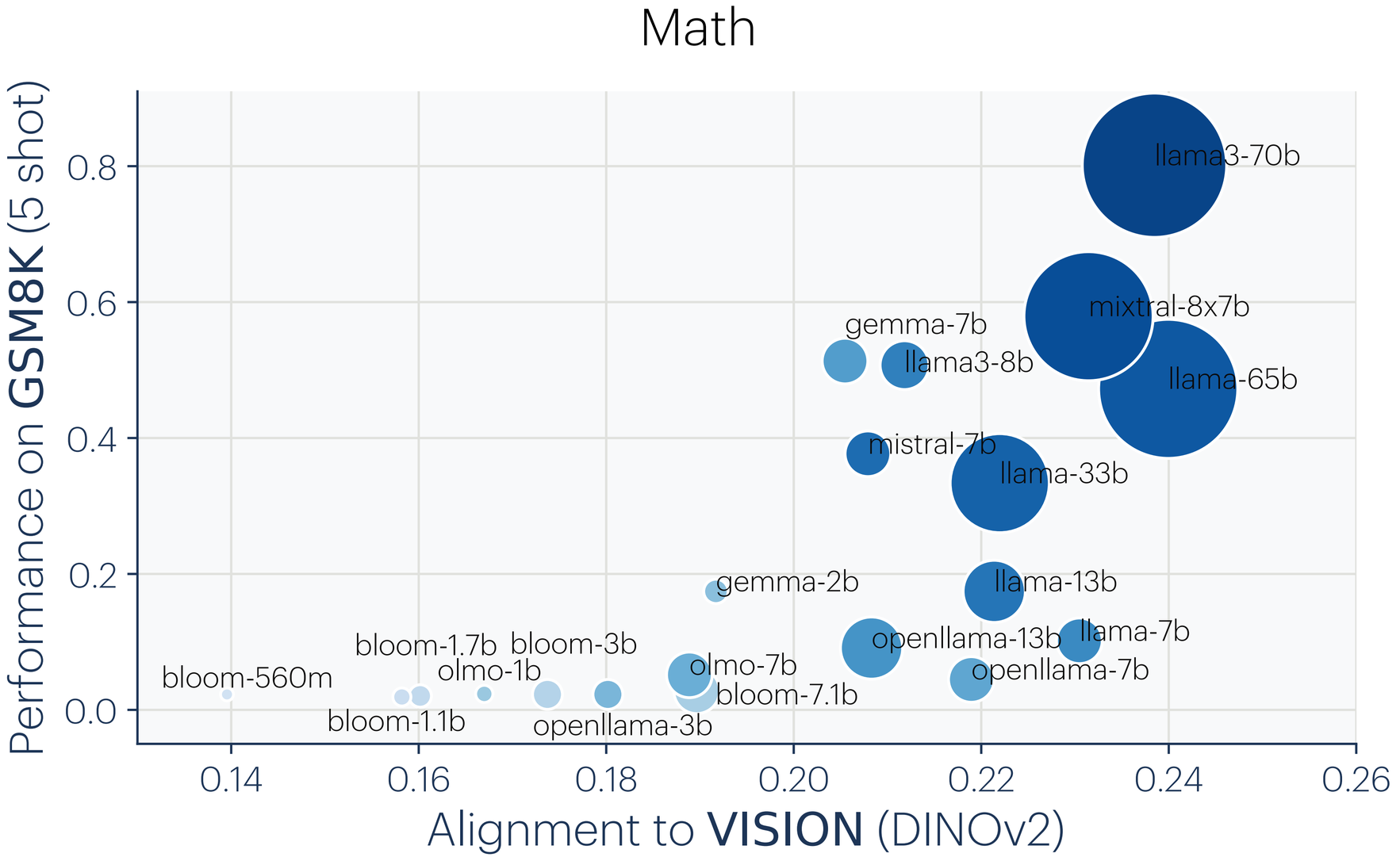




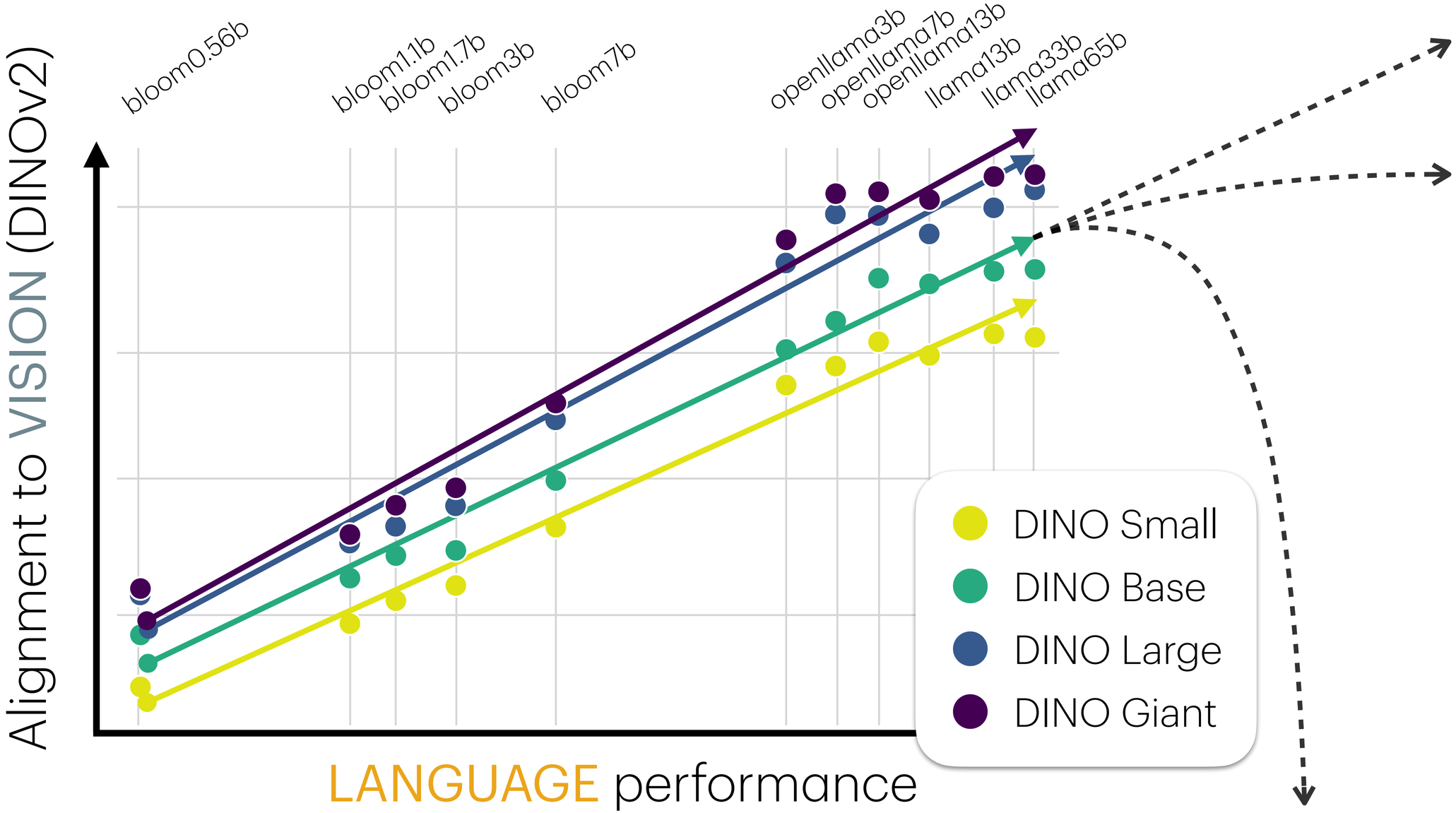
Alignment corresponds to better performance in **self consistent** domains



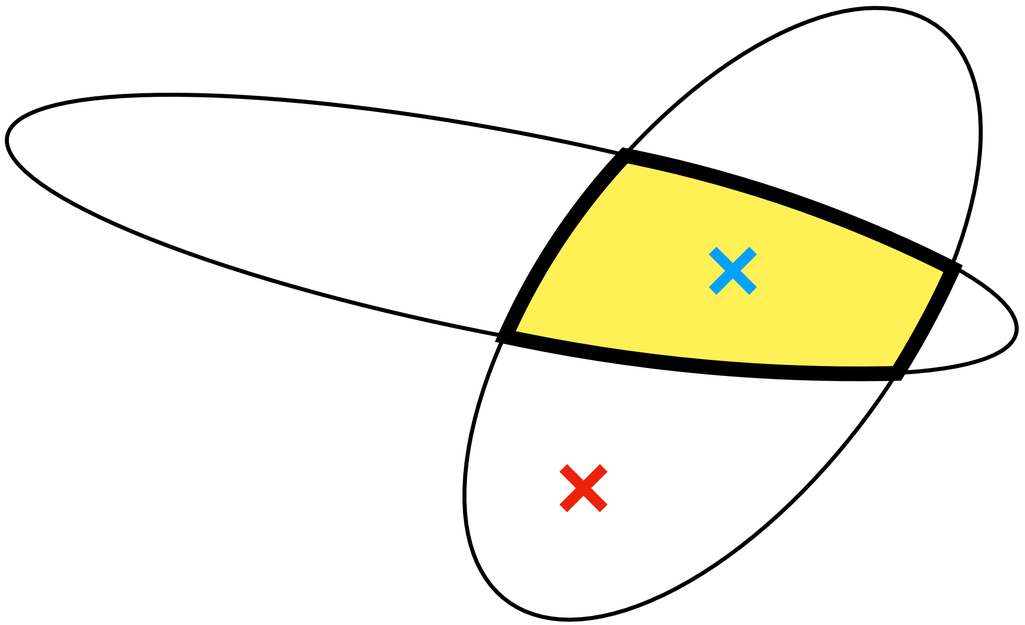
Alignment corresponds to better performance in **self consistent** domains

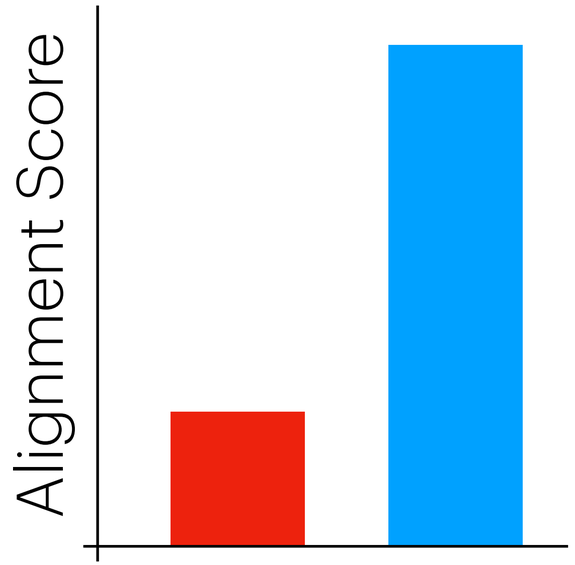


We can choose where this goes

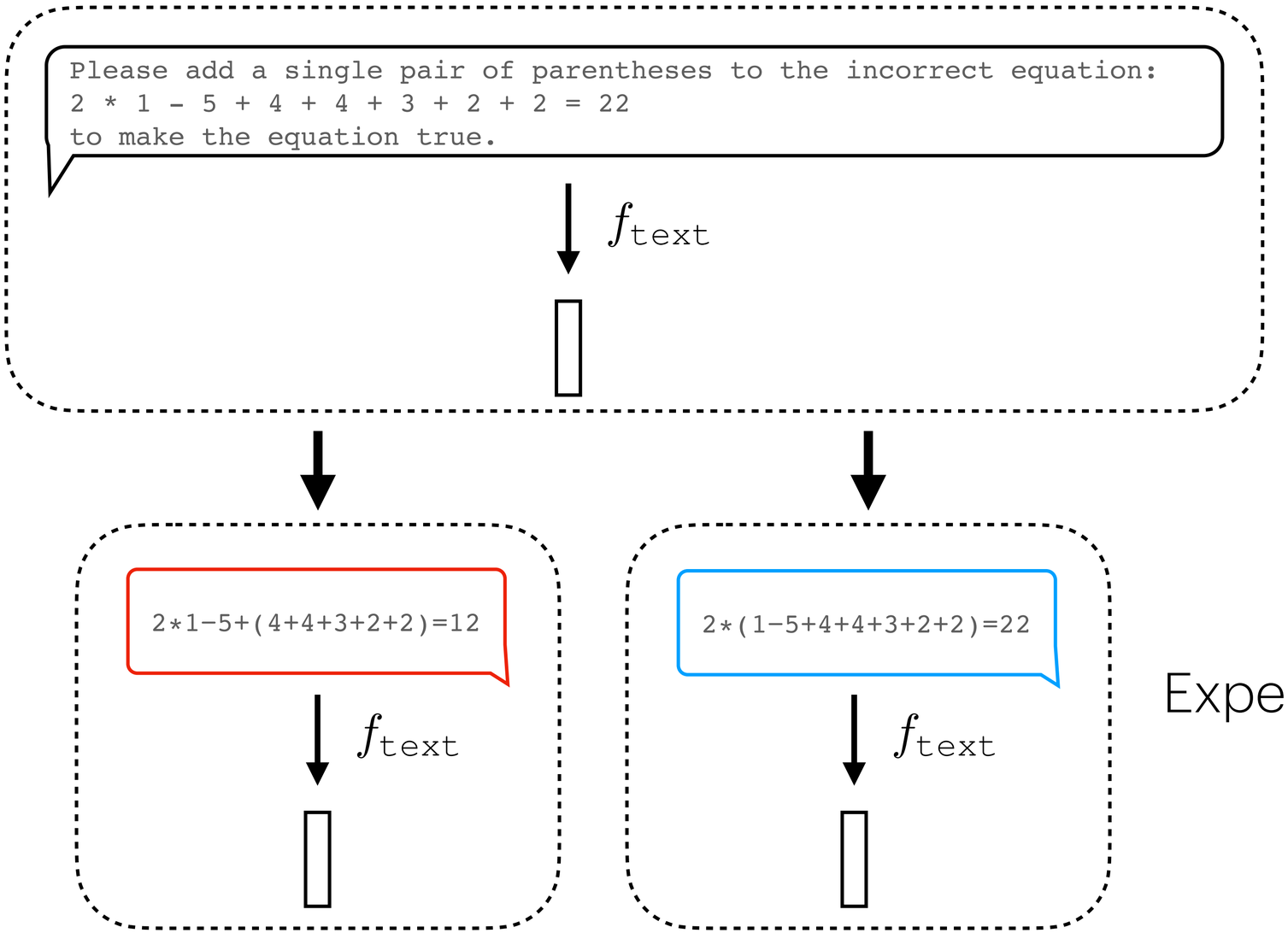


Alignment measures when hypotheses are **not consistent**



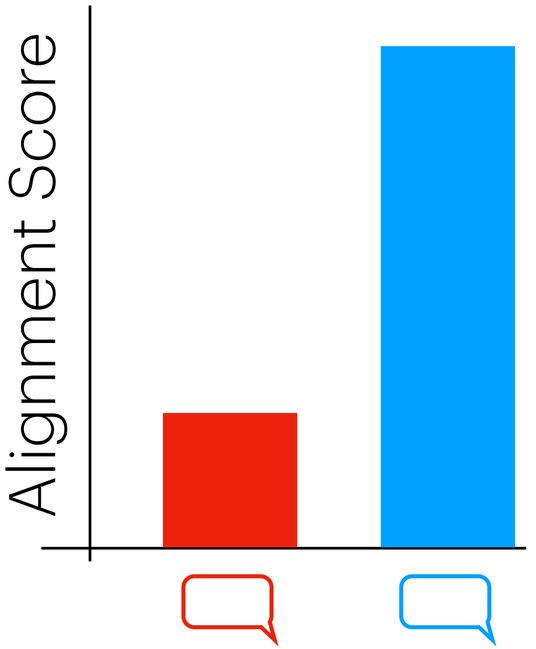


Math reasoning is an alignment loop



Hypothesis

Experiments



Representation Alignment for ‘soft’ verification

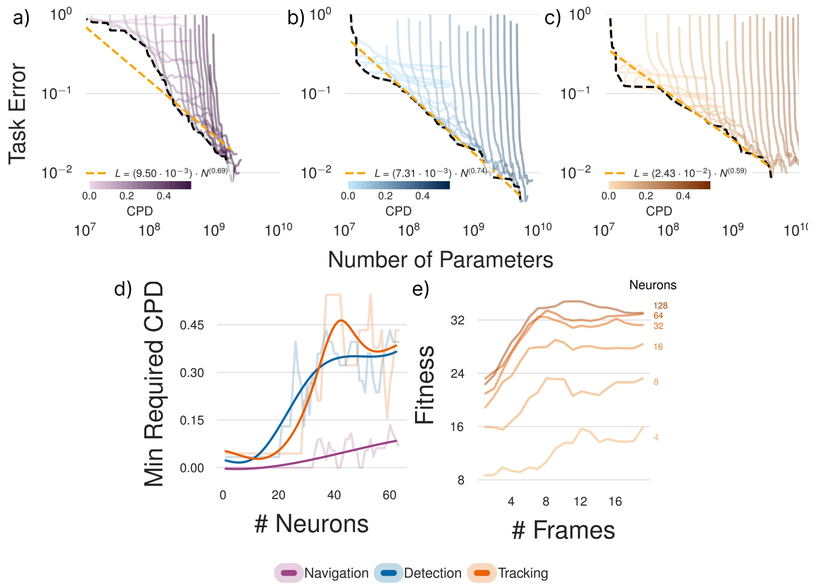
What if Eye...? Computationally Recreating Vision Evolution

Kushagra Tiwary*[1], Aaron Young*[1], Zaid Tasneem[2], Tzofi Klinghoffer[1,6], Akshat Dave[1], Tomaso Poggio[3], Dan-Eric Nilsson[4], Brian Cheung**[3,5], Ramesh Raskar**[1]

Abstract

Vision systems in nature show remarkable diversity, from simple light-sensitive patches to complex camera eyes with lenses [1, 2]. While natural selection has produced these eyes through countless mutations over millions of years, they represent just one set of realized evolutionary paths [3, 4]. Testing hypotheses about how environmental pressures shaped eye evolution remains challenging since we cannot experimentally isolate individual factors [5]. Computational evolution offers a way to systematically explore alternative trajectories [6–10]. Here we show how environmental demands drive three fundamental aspects of visual evolution through an artificial evolution framework that co-evolves both physical eye structure and neural processing in embodied agents. First, we demonstrate computational evidence that task specific selection drives bifurcation in eye evolution - orientation tasks like naviga-

2.6 What if animal brains stayed small throughout evolution?



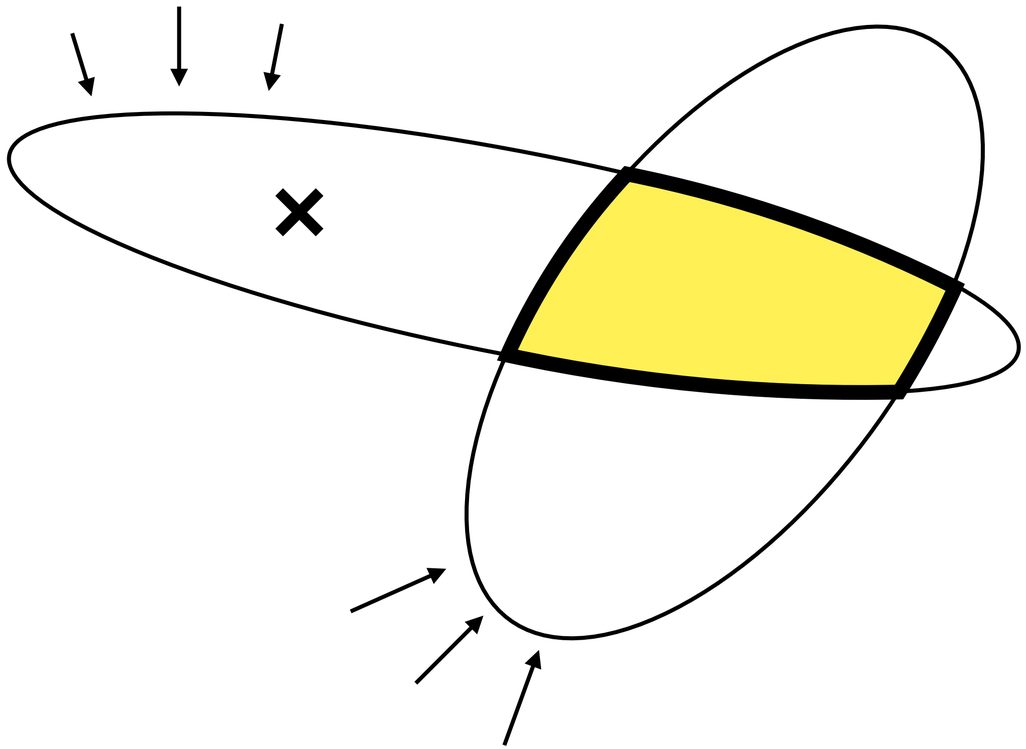
$f_{\texttt{text}}$ $f_{\texttt{text}}$

Similarity is **easier** to verify than exact matching

Alignment Enables:

AI for Science

Science for AI



You can ask questions

cheungb@mit.edu

UCSF in 2026

Ask me about PhD and Post-Docs!