

# **NX-414: Brain-like computation and intelligence**

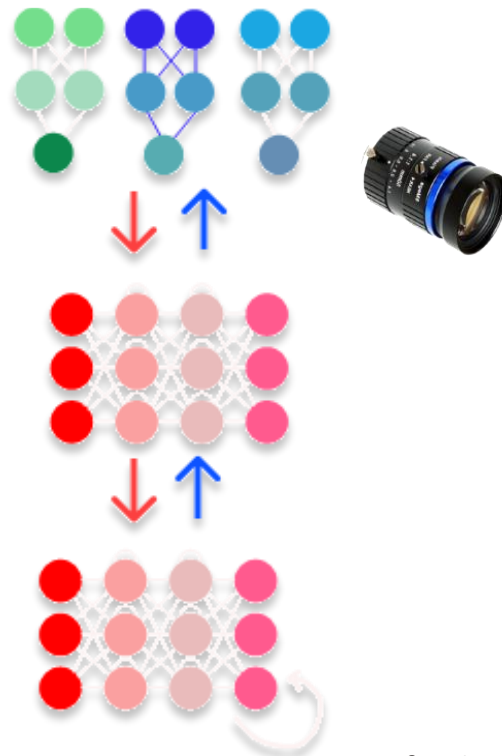
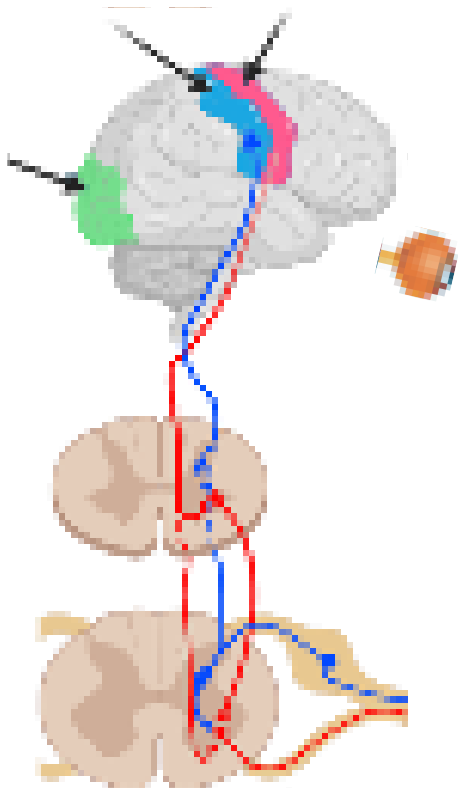
Martin Schrimpf

Lecture 10, 07 May 2025

Biological Intelligence



Artificial Intelligence



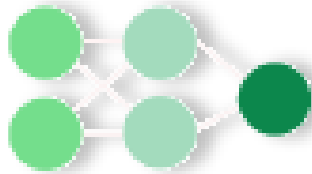
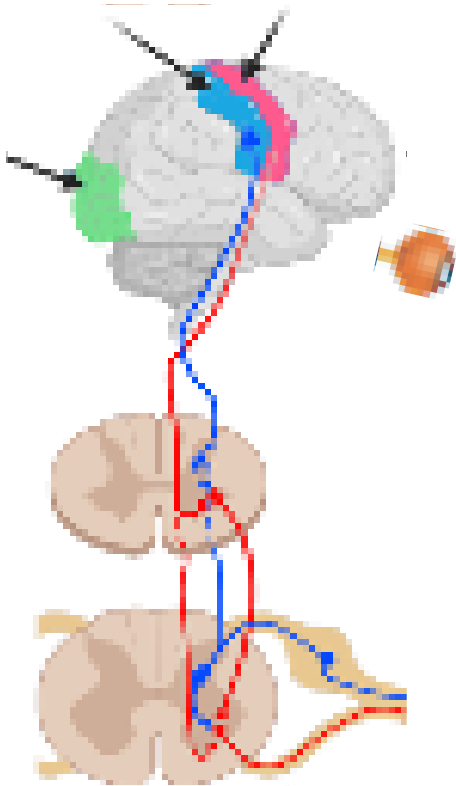
# Normative frameworks

## Information theoretic

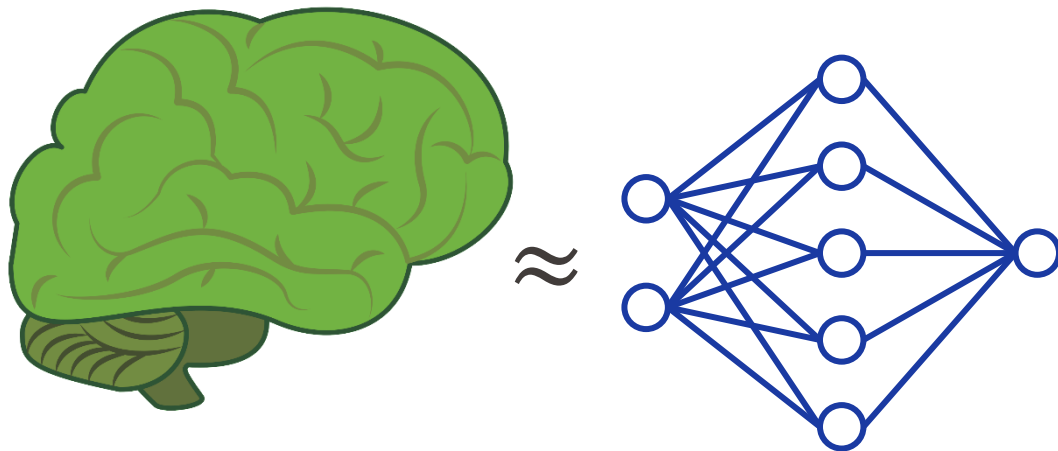
e.g. sparse coding,  
redundancy reduction,  
mutual information ...

## Utilitarian

e.g. recognize objects,  
chase prey, navigate, **next-  
word prediction**, ...



# Using deep neural networks as goal-driven models of a system



Vision: object recognition.

Yamins & Hong et al. (2014), Schrimpf & Kubilius et al. (2018)



Audition: speech recognition, speaker & sound identification. Kell et al. (2018)



Somatosensation: shape recognition.  
Zhuang et al. (2017)



Language: next-word prediction.  
Schrimpf et al. (2021)



Decision making: context-dependent choice. Mante & Sussilo et al. (2013)



Proprioception: action recognition.  
Sandbrink et al. (2023)

# Language

*Formal competence*



<b>Math</b>  $2+17=$ $56 \times 3 =$	<b>Logic</b> $\neg, \vee, \wedge, \oplus, \neg, \sim$ $\Rightarrow, \Leftrightarrow, \leftrightarrow$ $\exists, \nexists, !, \& \vee$	<b>Coding</b> 
---	--	-------------------

# Thought (Cognition)



**Executive functions**

**Problem solving**

**World knowledge + commonsense reasoning**

**Physical knowledge + reasoning**

**Social knowledge + reasoning**

**Situation modeling**

**Formal reasoning**  
(math, logic, ...)

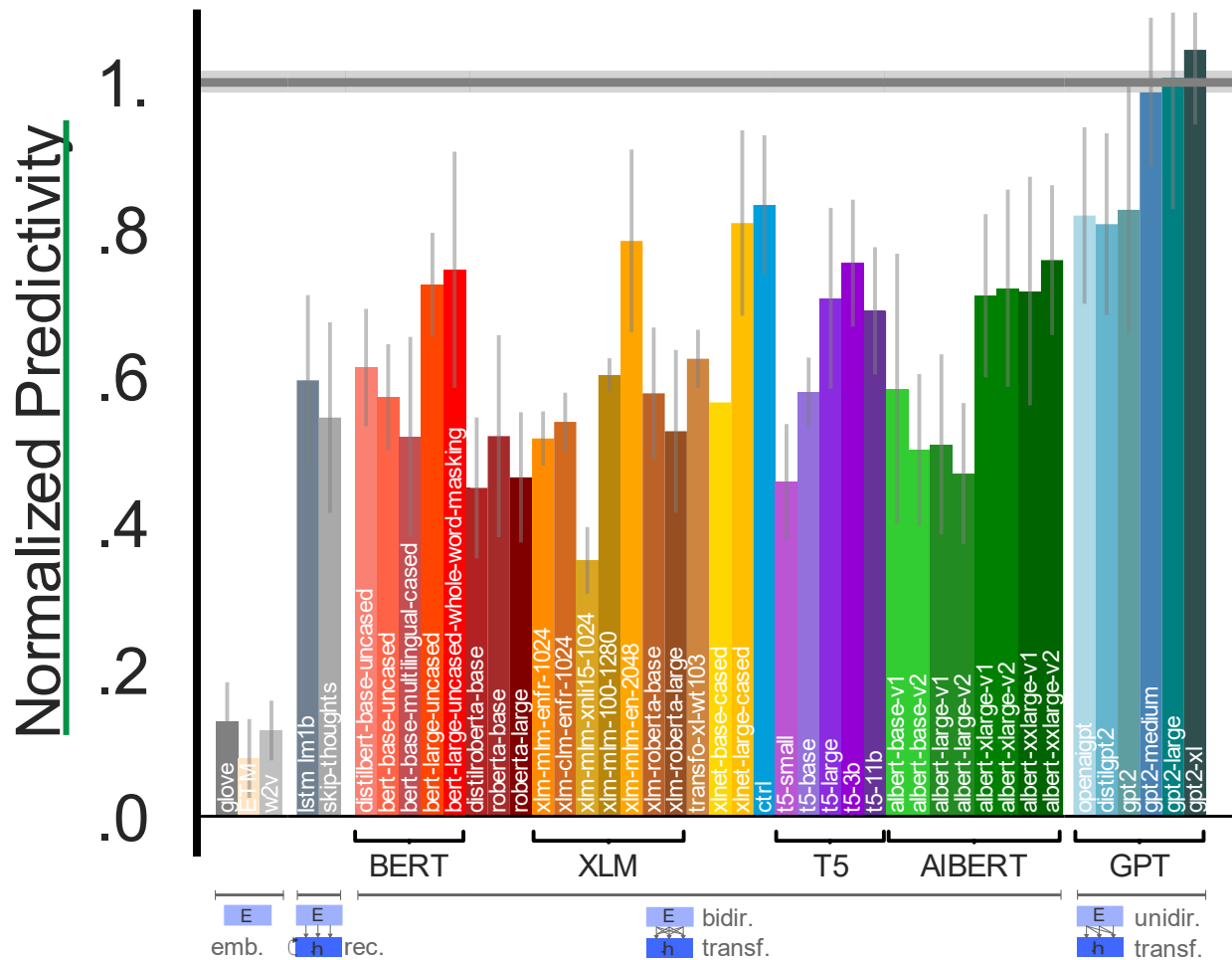
**World knowledge**  
(concepts, facts...)

**Situation modeling**  
(coherence, narrative structure...)

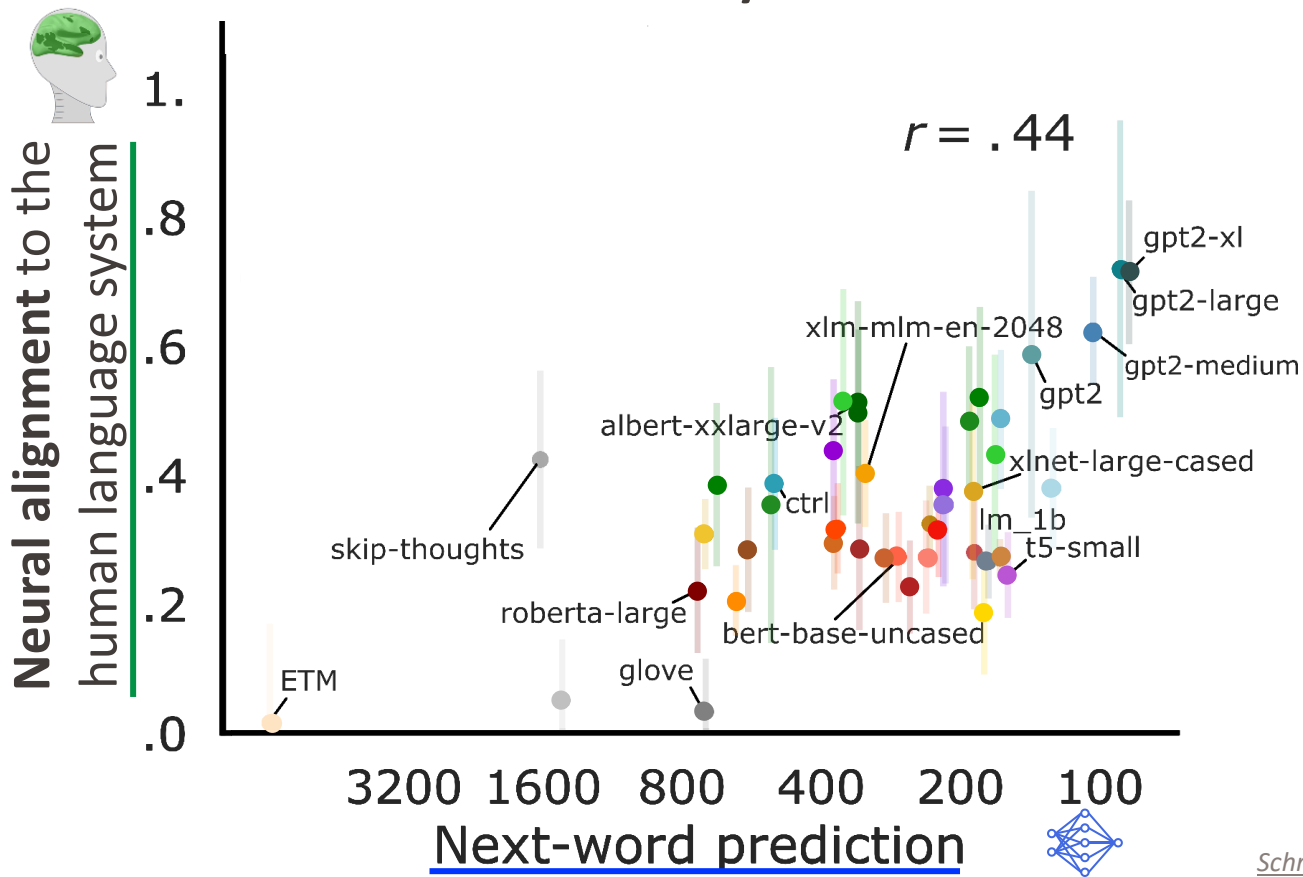
**Communicative intent**  
(pragmatics, common ground, goals, ...)

**Functional competence**

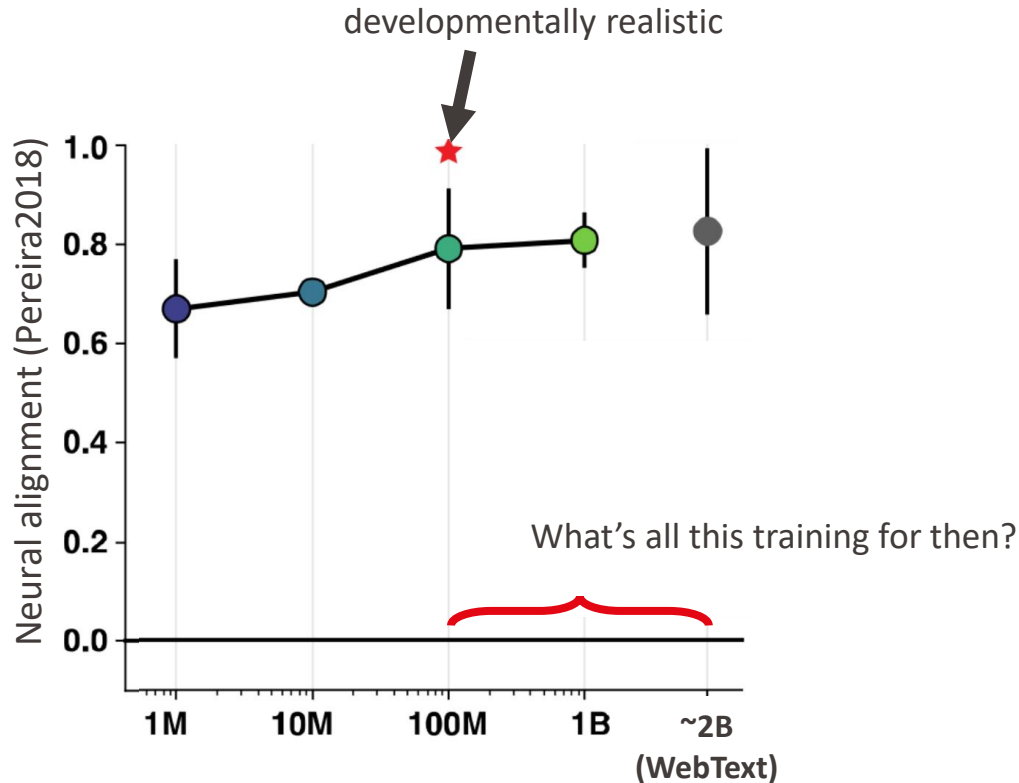
# Certain language models predict human language recordings



# The better models can predict the next word, the more brain-like they are



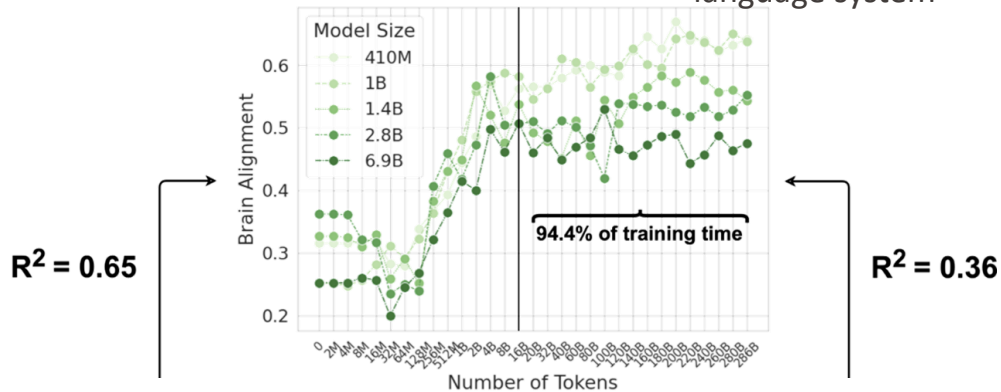
# LLMs align to the brain's language system after developmentally realistic amounts of training



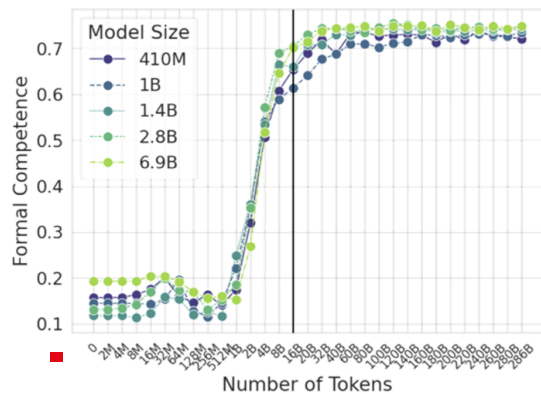


# Model alignment with the human language system primarily tracks with improvements in formal competence

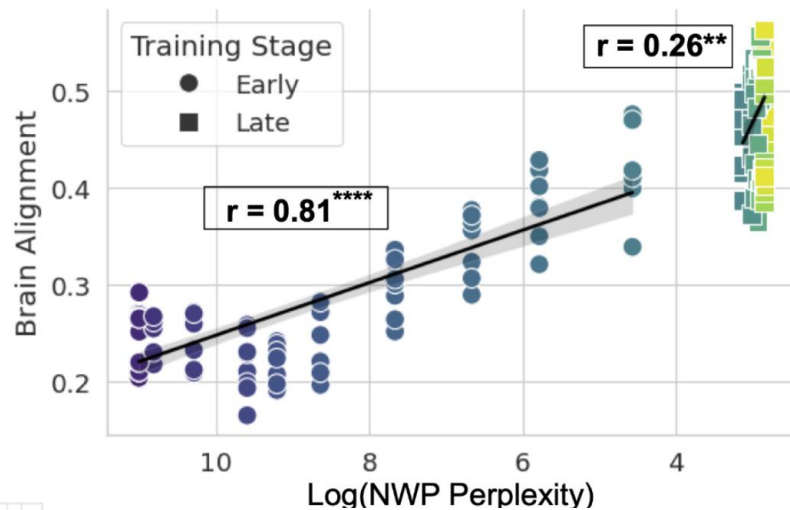
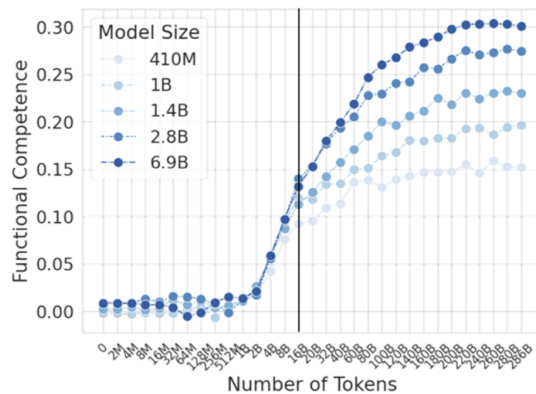
(a) Brain Alignment with human language system



(b) Formal Competence



(c) Functional Competence

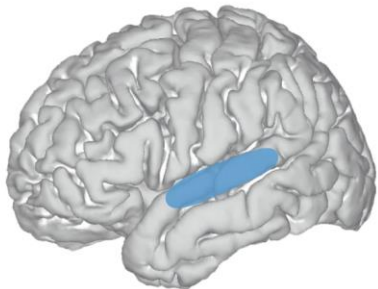


- Early training: improve formal competence and brain alignment
- Late training: improve functional competence, but not alignment with HLS

# Beyond language

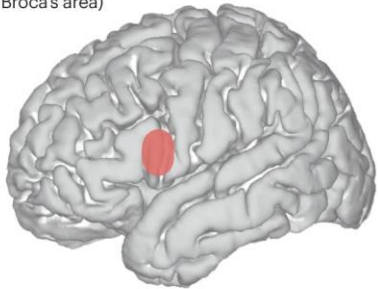
## Perception

Perception of the surface properties of linguistic input (for instance, speech perception area)



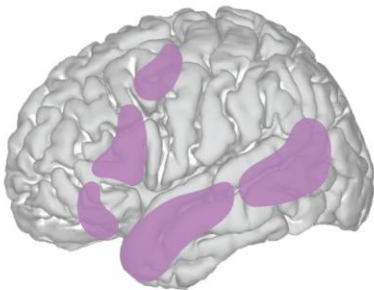
## Motor planning

Planning of the motor movements needed to realize linguistic output (for instance, Broca's area)



## Language

Language knowledge and processing (language network)

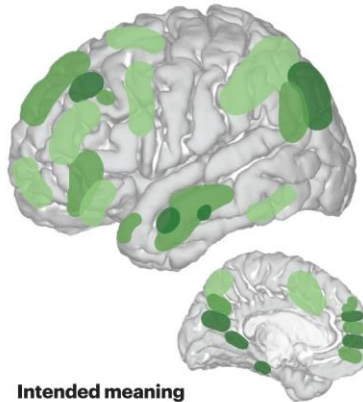


Language as  
the bridge from  
perception to  
higher cognition

→ Language comprehension  
→ Language production

## Knowledge and reasoning

- Task demands beyond language (multiple demand network)
- Pragmatics, social reasoning (theory of mind network)
- Narratives, situation modelling (default mode network)



**Intended meaning**  
(multiple brain areas, including the above)

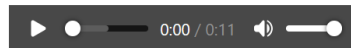
- Speech
- Reasoning & Planning
- Theory of Mind
- Physical Reasoning
- Emotions
- Agents

# (Text-to) Speech

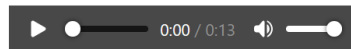
- E.g., LLaSA
- Basically, LLM architecture (LLaMA) + audio tokenizer (xcodec2), trained on multilingual speech
- Style of speech: vary e.g. emotions
- <https://huggingface.co/blog/srinivasbilla/llasa-tts#amelia>
- Voice cloning: condition on short recording, create consistent audio
- <https://huggingface.co/spaces/srinivasbilla/llasa-3b-tts>

## Amelia

### Reference



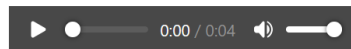
*Hi! I'm Amelia, a super high quality English voice. I love to read. Seriously, I'm a total bookworm. So what are you waiting for? Get me reading! Clone*



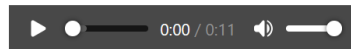
*All you need is a short clean audio sample of just 5 to 10 seconds. Then the model can generate a high quality speech sample mimicking the voice, tone and style of speech and even accent.*

## Russel

### Reference



*it is not enough to have a good mind the main thing is to use it well Clone*



*The model was trained on a 160,000 250,000 hours of audio tokenized by Xcodec2, which converts audio to tokens at a very efficient 50 tokens per second.*

# Reasoning, theory of mind, and semantic tracking

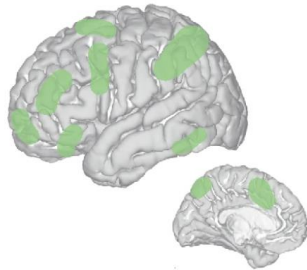
2a

Subtract three from  
six hundred and one,  
then add two.

3a

$$601 - 3 + 2 = ?$$

Multiple demand network



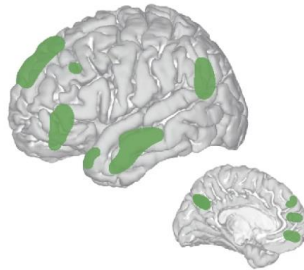
2b

She thought that the  
ice cream was on the  
table.

3b



Theory of mind network



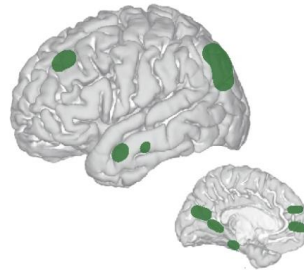
2c

She moved the ice  
cream from the table  
to the freezer.

3c



Default mode network



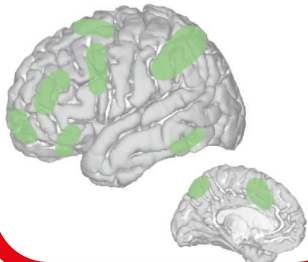
2a

Subtract three from  
six hundred and one,  
then add two.

3a

$$601 - 3 + 2 = ?$$

Multiple demand network



- Generally problems that you have to “think about”, i.e. don’t immediately come to you
- Goal-directed behavior for novel and/or difficult tasks that require reasoning
- Might involve the management of attention, working memory, and problem solving

# Code comprehension in MD

- How does the brain process computer code?

## A Experiment 1 - Python

### code problem

```
height = 5
weight = 100
bmi = weight/(height*height)
print(bmi)
```

### sentence problem

Your height is 5 feet and your weight is 100 pounds. The BMI is defined as the ratio between the weight and the square of the height of a person. What is your BMI?

## Experiment 2 - ScratchJr

### code problem



### sentence problem

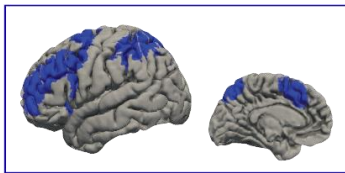
Kitten walks right, jumps, and then walks left.

# Code comprehension in MD

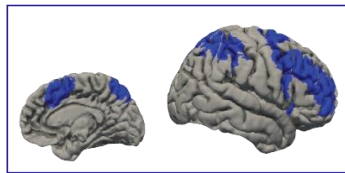
- Consistently stronger responses in Multiple Demand network to coding problems (CP) in two programming languages, compared to sentence reading (SR), nonwords reading (NR), and sentence problems (SP)
- Weaker responses in language system

A

MD System  
Left Hemisphere



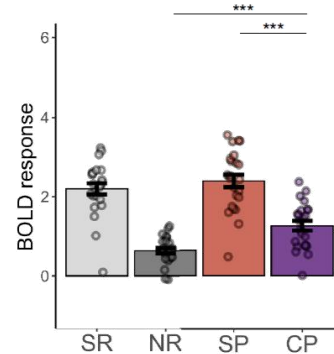
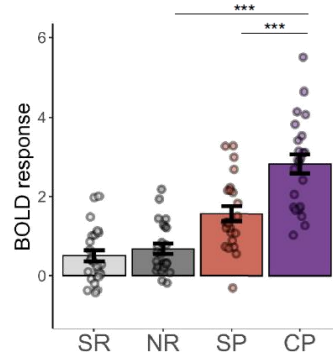
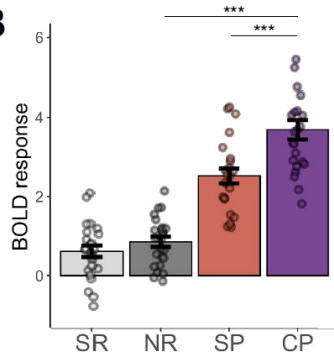
MD System  
Right Hemisphere



Language System

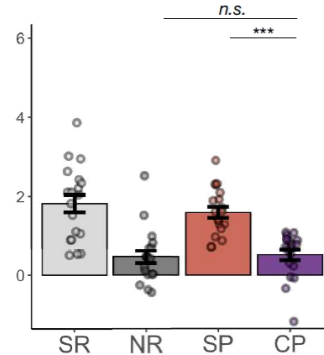
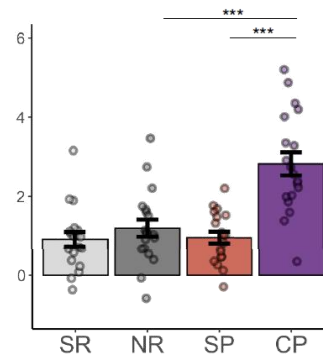
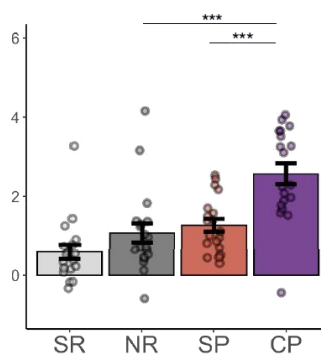


B



Python

C



Scratchjr

# Reasoning: Chain of Thought

## (c) 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

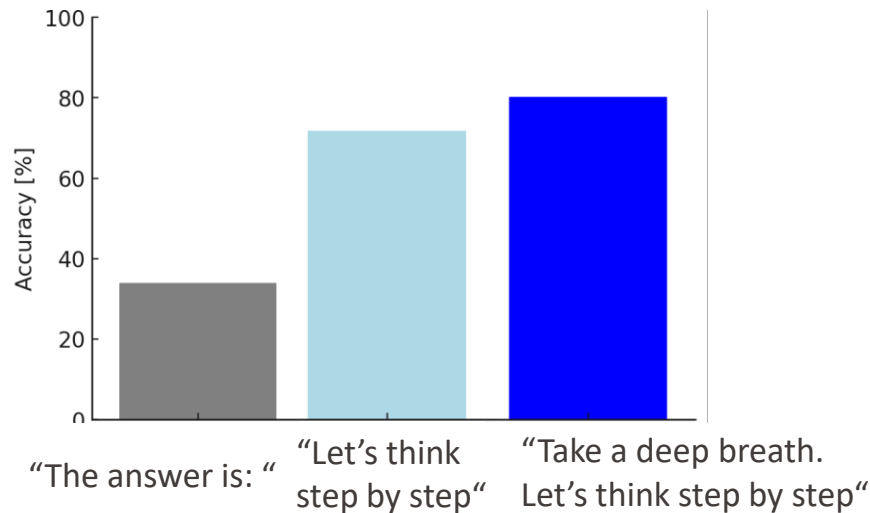
(Output) 8 ✗

## (d) Zero-shot-CoT (Ours)

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?

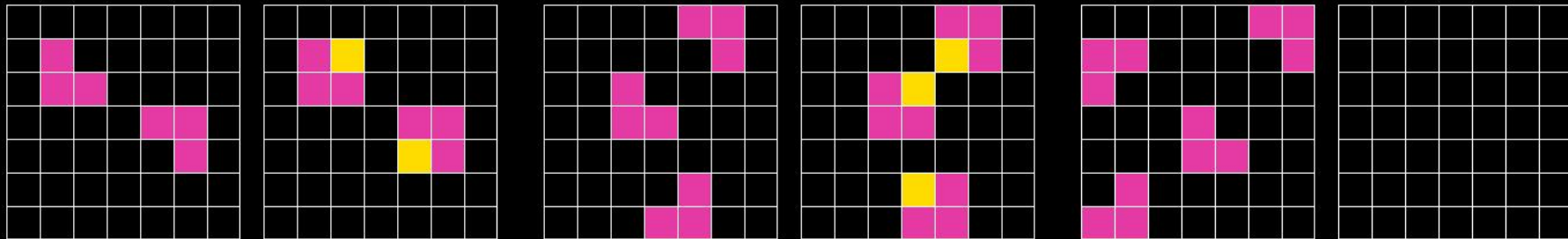
A: **Let's think step by step.**

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.* ✓



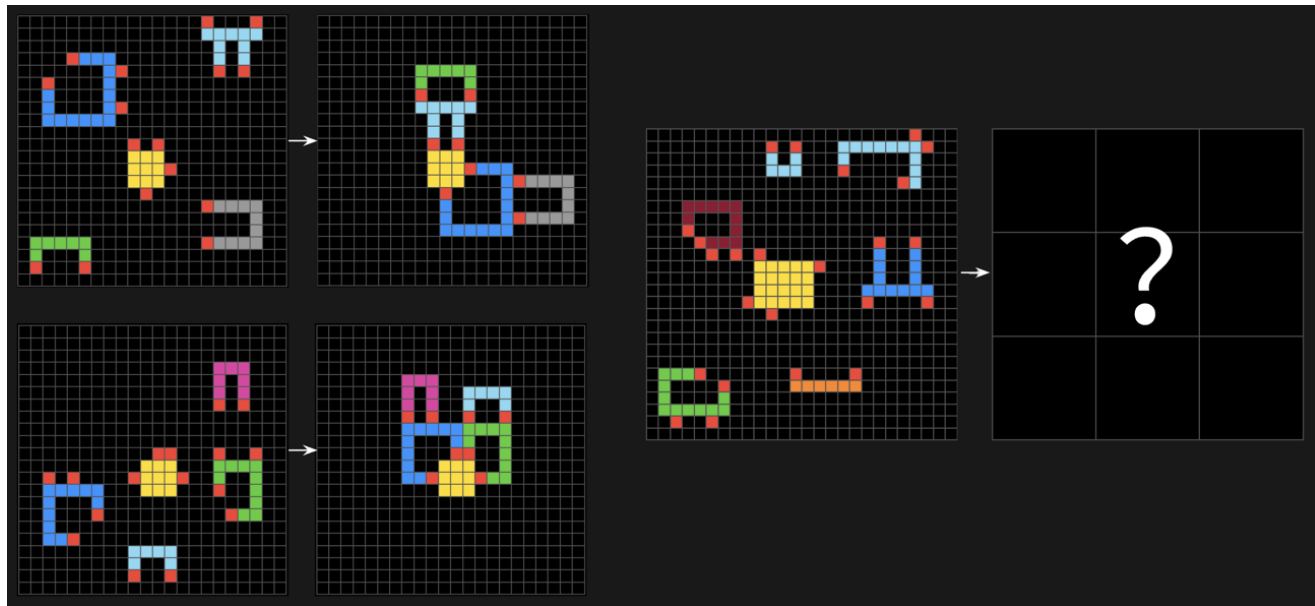
- Can improve LLM performance by encouraging it to “think through” steps for answer
- Funny enough, encouraging it to take a deep breath beforehand also helps...





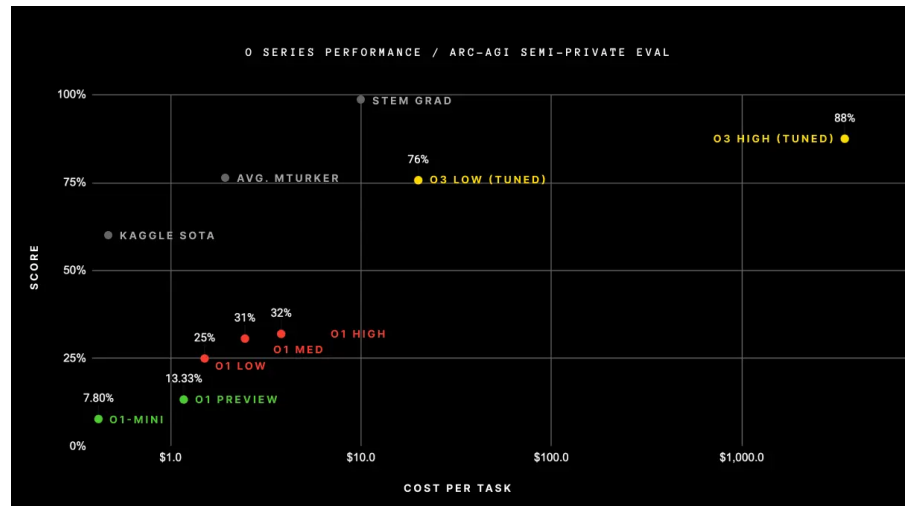
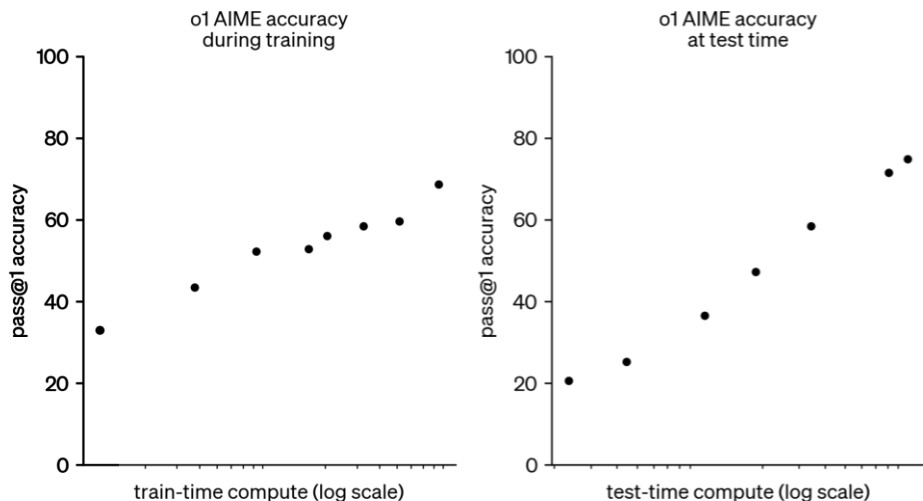
<https://arcprize.org/play?task=3aa6fb7a>

- Designed to be easy for humans, hard for AI
- Limited training data provided (400 tasks), public and private test sets
- \$1M prize money
- 



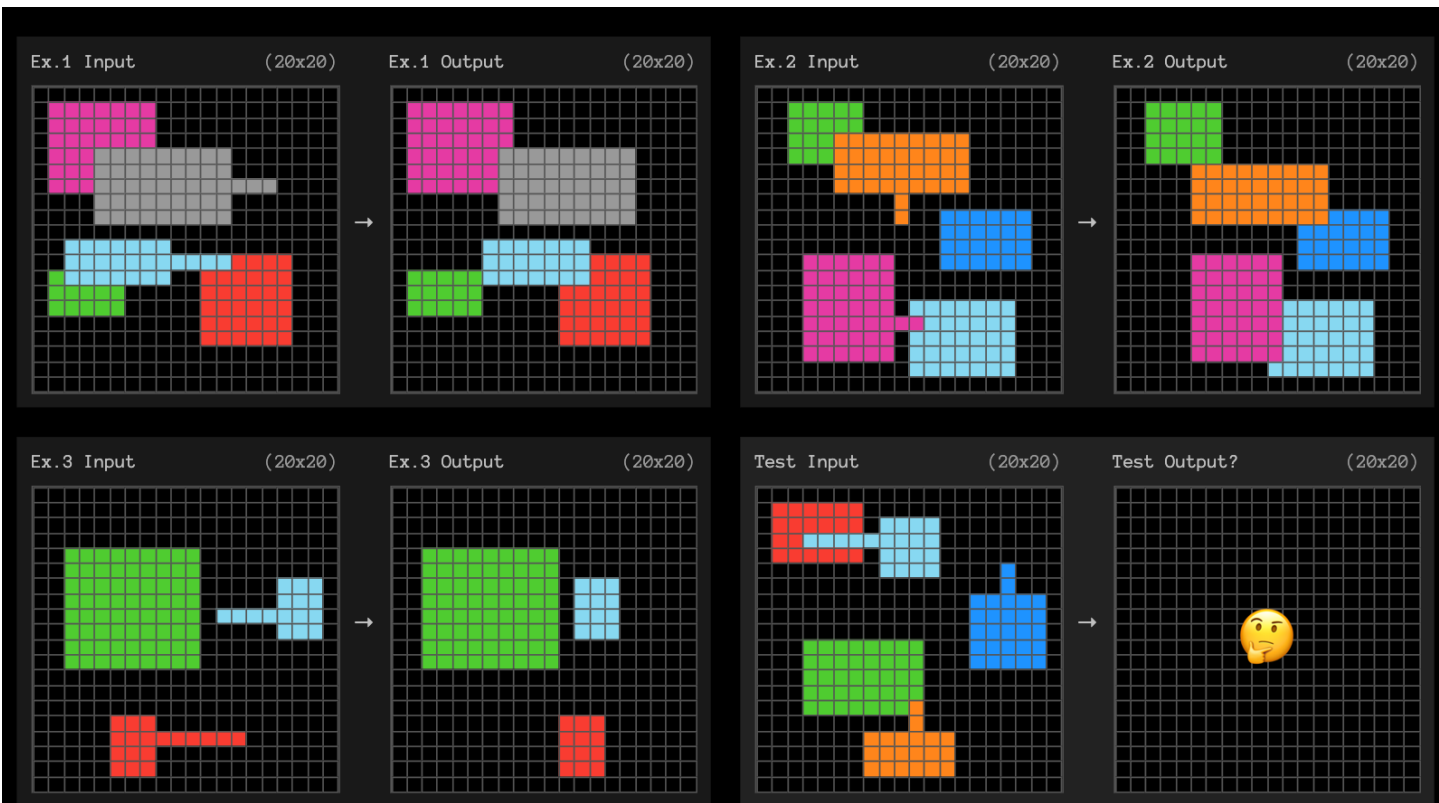
# Test-time scaling

- More compute during train **as well as test** time → better performance
- But (currently) gets quite costly, best model >\$1k per task



- SOTA models still fail on tasks that are trivial for humans (task below unsolved by high-compute o3)

<https://arcprize.org/blog/oai-o3-pub-breakthrough>



# Theory of mind

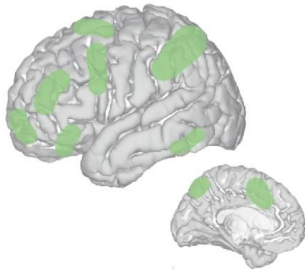
2a

Subtract three from  
six hundred and one,  
then add two.

3a

$$601 - 3 + 2 = ?$$

Multiple demand network



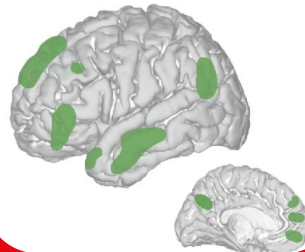
2b

She thought that the  
ice cream was on the  
table.

3b



Theory of mind network



- Represent and reason about others' beliefs
- E.g., Sally-Anne false belief tasks

# Sally-Anne false belief task

- Most subjects would answer that Sally would look for the marble in the basket.
- False belief tasks are very difficult for autistics. Even when well past the expected age, they act as if all the characters have all the information.



# First- and second-order theory of mind

- “Sally-Anne” questions: evaluate first-order beliefs
- Second-order theory of mind: infer beliefs about beliefs
- Two agents (Mary, John) see an ice cream van in the park. The vendor tells them that he will be in the park all afternoon. After Mary leaves the park, the vendor decides to leave the park and tells John he is going to the church. On the way to the church, the vendor meets Mary and tells her also that he will be at the church.
- Second-Order Belief: Where does John think Mary will go to get ice cream?

# ToMi: ML dataset for ToM evaluation

## ToMi dataset

*Examples of stories from the ToMi dataset*

1 Oliver dislikes the kitchen  
 2 Carter entered the porch.  
 3 Abigail entered the porch.  
 4 The potato is in the green\_suitcase.  
 5 Abigail exited the porch.  
 6 Abigail entered the hall.  
 7 Carter moved the potato to the green\_envelope.  
 8 Oliver entered the hall.

1 Mila entered the closet.  
 2 Isla entered the closet.  
 3 Ava entered the closet.  
 4 The orange is in the blue\_container.  
 5 Isla exited the closet.  
 6 Isla entered the garage.  
 7 Ava moved the orange to the green\_bathtub.

1 William entered the staircase.  
 2 Aiden entered the staircase.  
 3 Aiden exited the staircase.  
 4 Aria entered the staircase.  
 5 The potato is in the red\_drawer.  
 6 Aiden dislikes the grapefruit  
 7 William moved the potato to the blue\_container.  
 8 Aria exited the staircase.

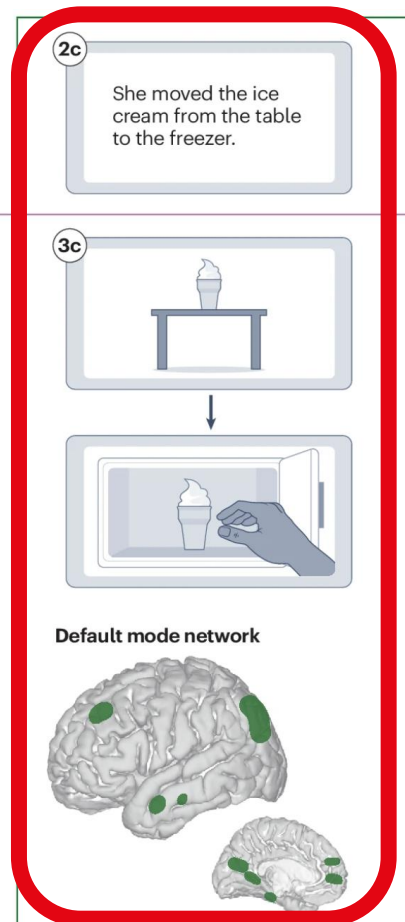
- Generative stories (scale!)

		MemNN	RelNet	EntNet
<b>w/o FB</b>	First Order	85.45	96.42	94.29
	Second Order	82.67	95.37	85.08
	Reality	93.39	100.0	100.0
	Memory	98.90	99.90	100.0
<b>FB</b>	First Order	12.62	10.40	54.95
	Second Order	17.27	17.81	36.55

- Easy for models when there is no false belief in the story (“w/o FB”)
- Difficult for models when there is a false belief (“FB”)

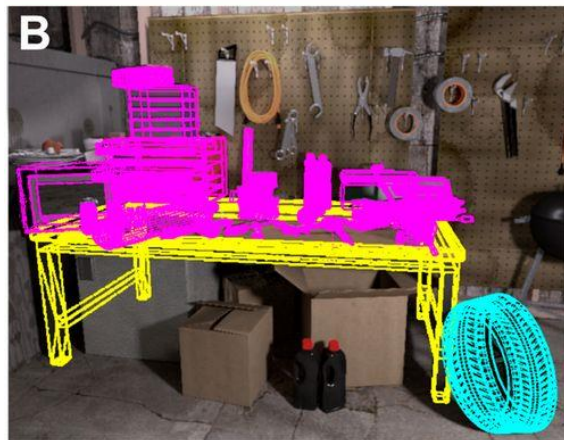
# Semantic context tracking

- “episodic projection, self-directed processing, internal scene construction or spatial information processing”
- sensitivity to long-range temporal contexts: receptive window on the scale of minutes compared to the language network’s receptive window of a few words → integrate information over longer scales
- appears to track abstract, input-invariant, global situational context
- No dedicated ML benchmarks (yet)

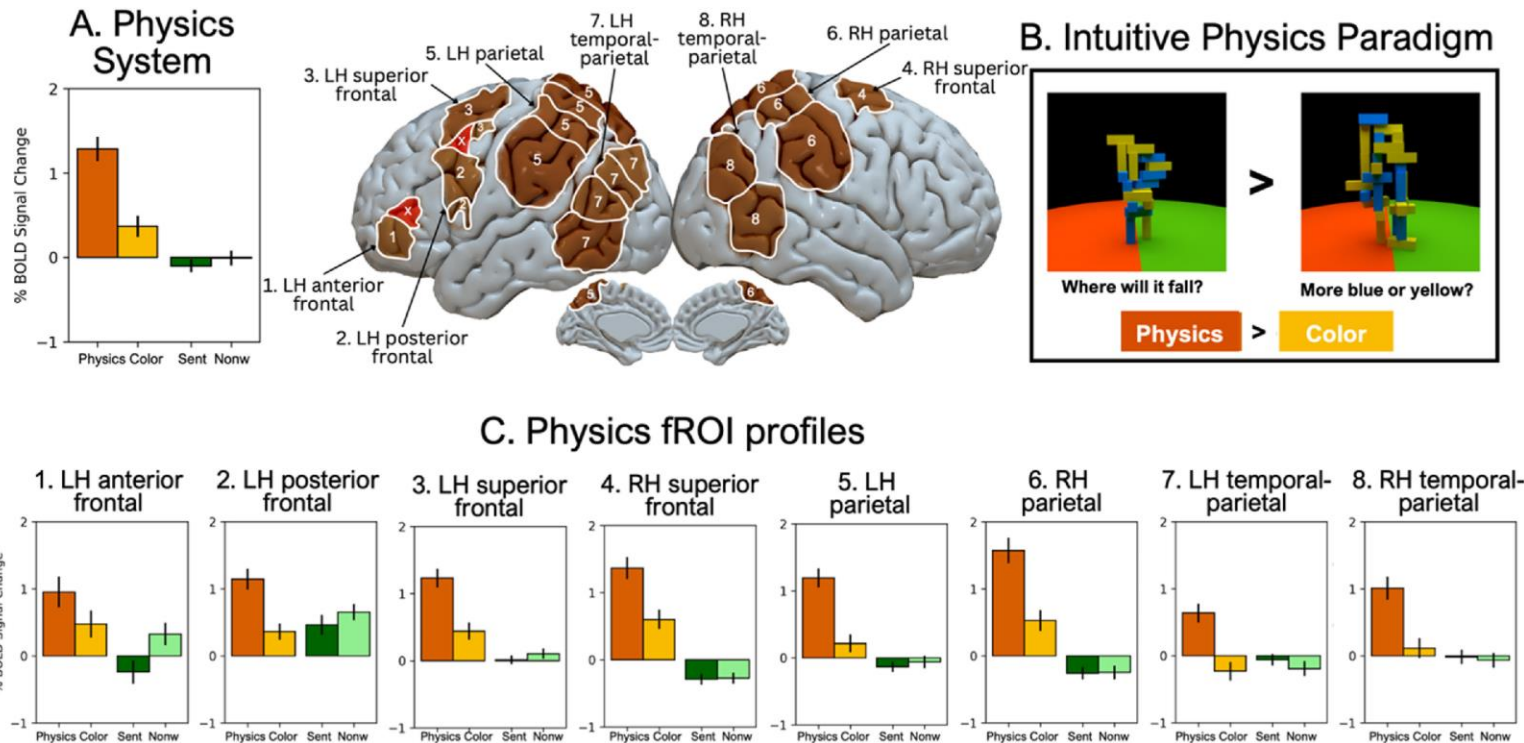




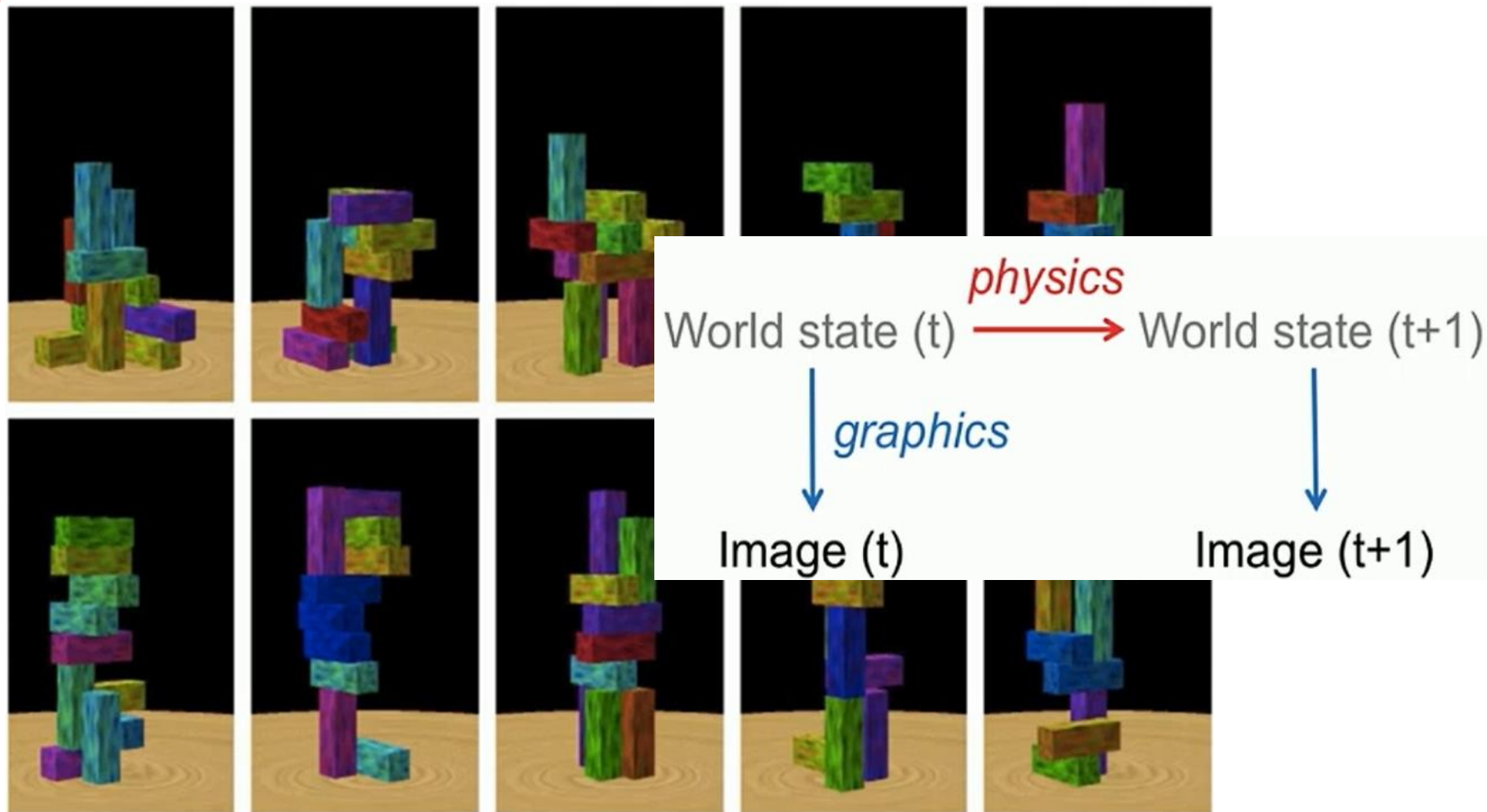
# Physical Reasoning



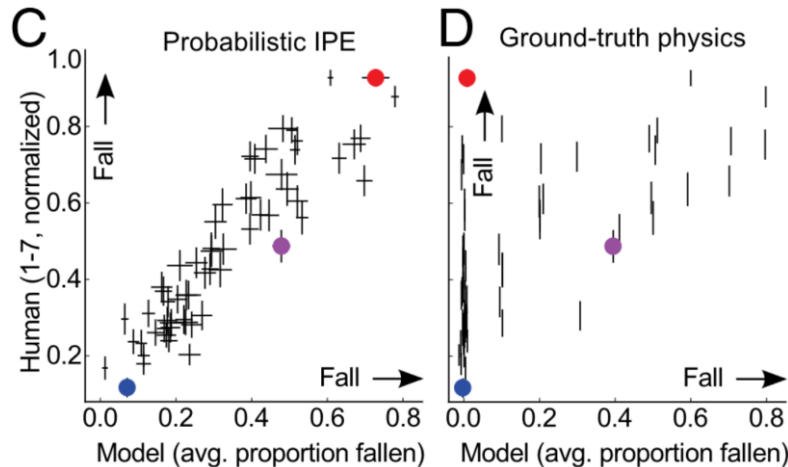
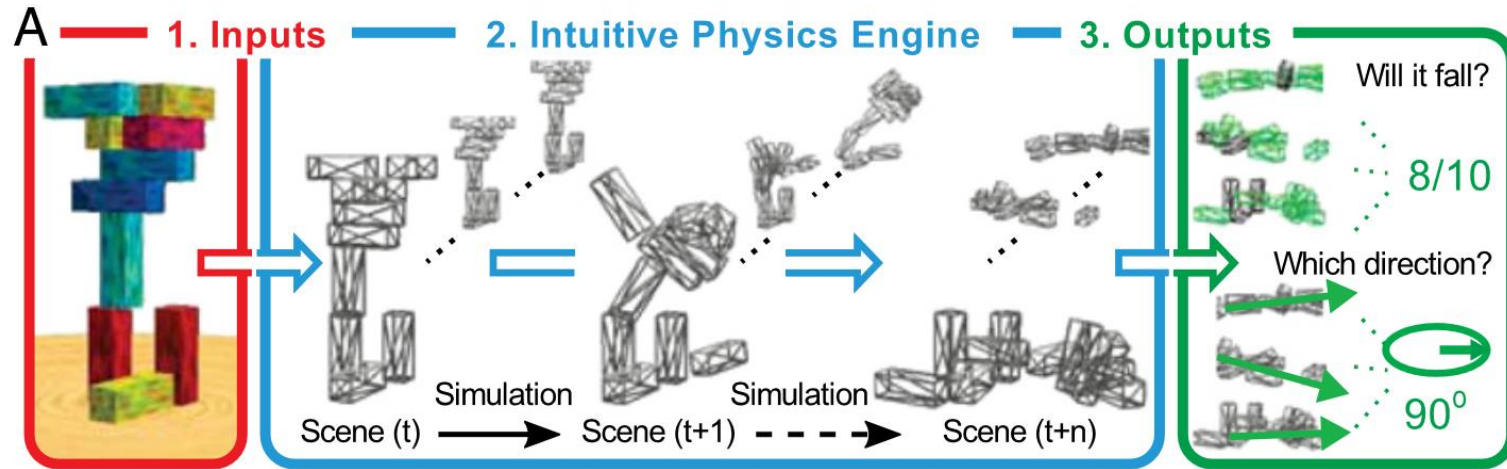
# An intuitive physics engine in the human brain?



- Evidence for brain regions selectively responding when engaging in “physics” tasks over control tasks



# Modeling intuitive physics with probabilistic simulations



- Approximate physics engine with fast, probabilistic simulations (“Bayesian cognitive modeling”)
- Monte Carlo simulations on scene states
- Predicts human ratings well – better than ground-truth physics





OPEN

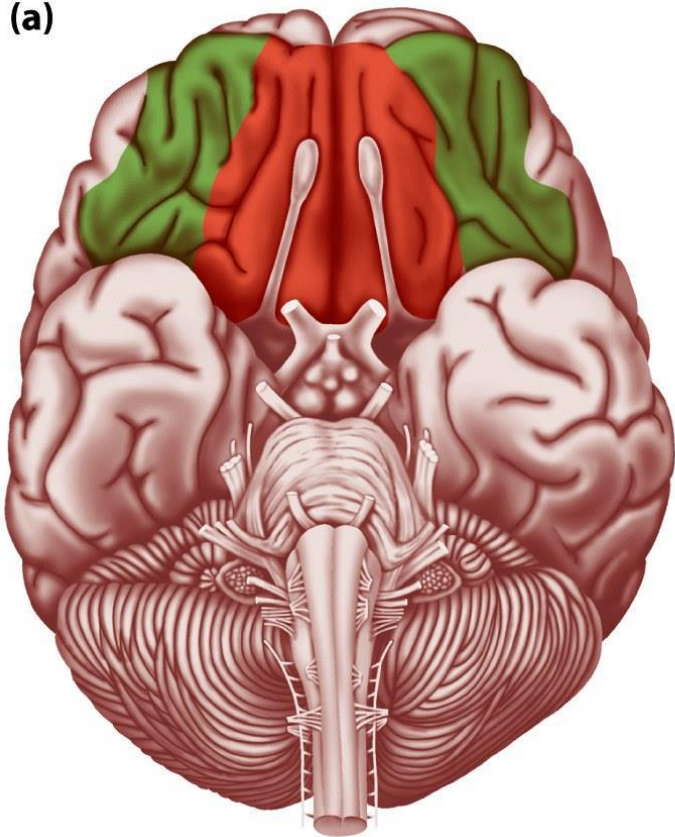
# Intuitive physics learning in a deep-learning model inspired by developmental psychology

Luis S. Piloto <sup>1,2</sup> ✉, Ari Weinstein<sup>1</sup>, Peter Battaglia<sup>1</sup> and Matthew Botvinick <sup>1,3</sup>

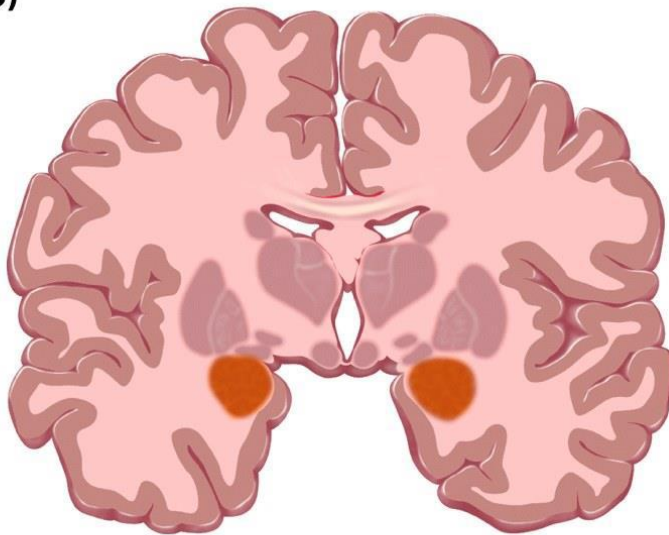
**‘Intuitive physics’ enables our pragmatic engagement with the physical world and forms a key component of ‘common sense’ aspects of thought. Current artificial intelligence systems pale in their understanding of intuitive physics, in comparison to even very young children. Here we address this gap between humans and machines by drawing on the field of developmental psychology. First, we introduce and open-source a machine-learning dataset designed to evaluate conceptual understanding of intuitive physics, adopting the violation-of-expectation (VoE) paradigm from developmental psychology. Second, we build a deep-learning system that learns intuitive physics directly from visual data, inspired by studies of visual cognition in children. We demonstrate that our model can learn a diverse set of physical concepts, which depends critically on object-level representations, consistent with findings from developmental psychology. We consider the implications of these results both for AI and for research on human cognition.**

# EPFL Emotions: Neural Systems

(a)



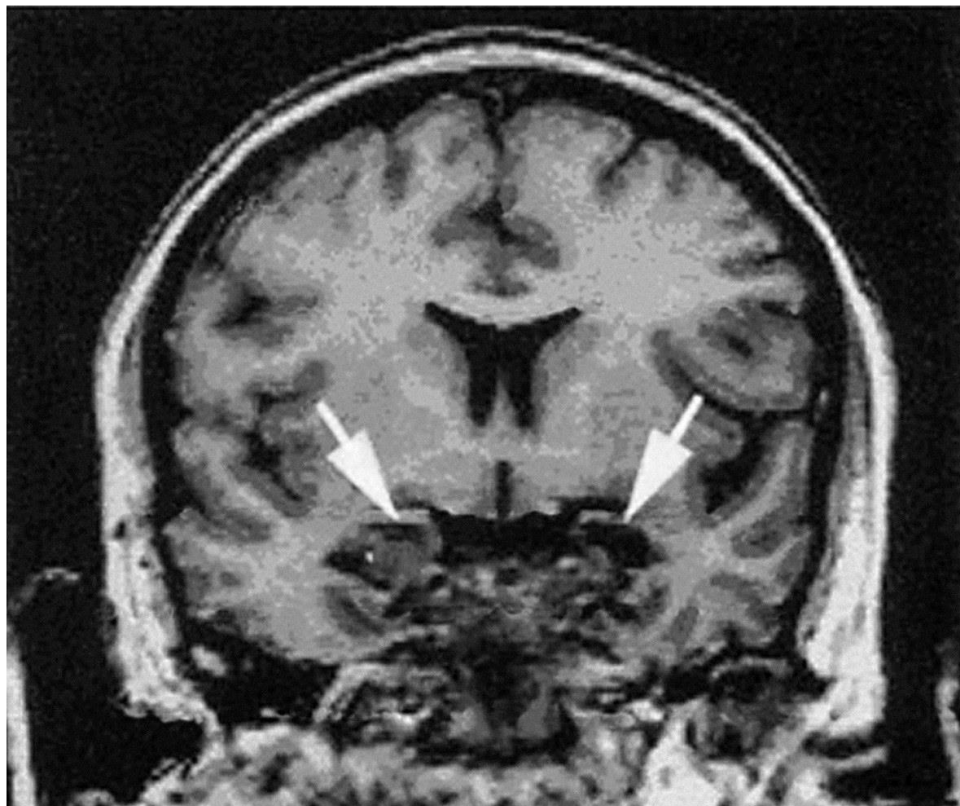
(b)



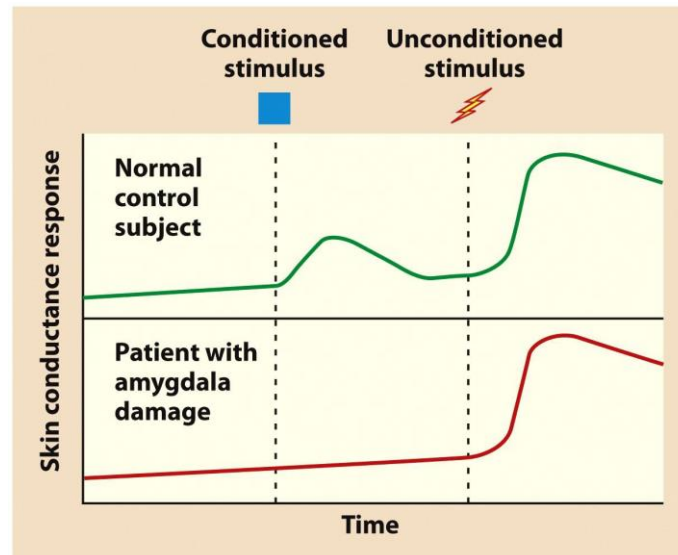
Neural regions  
central to the  
processing of  
emotions:

(a) Orbitofrontal  
cortex (OFC)

(b) Amygdala



- Bilateral amygdala degeneration
- Reduced ability to feel fear
- The “woman with no fear”



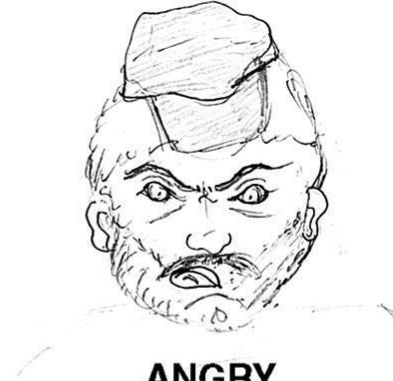
# Patient S.M.'s concept of basic emotions



**HAPPY**



**SAD**



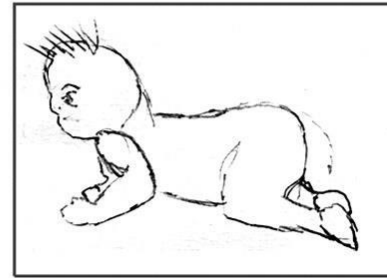
**ANGRY**



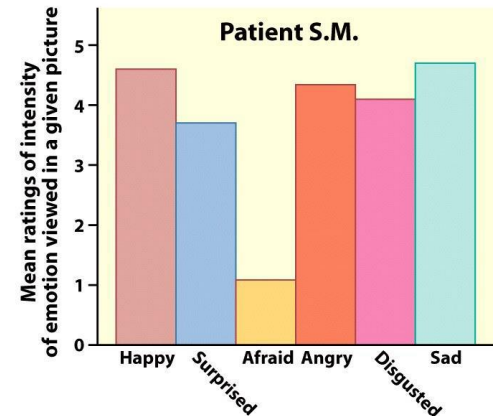
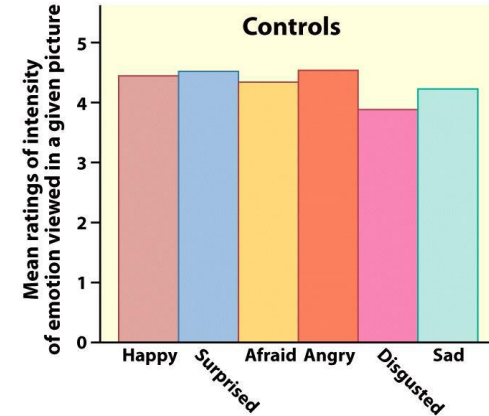
**SURPRISED**



**DISGUSTED**

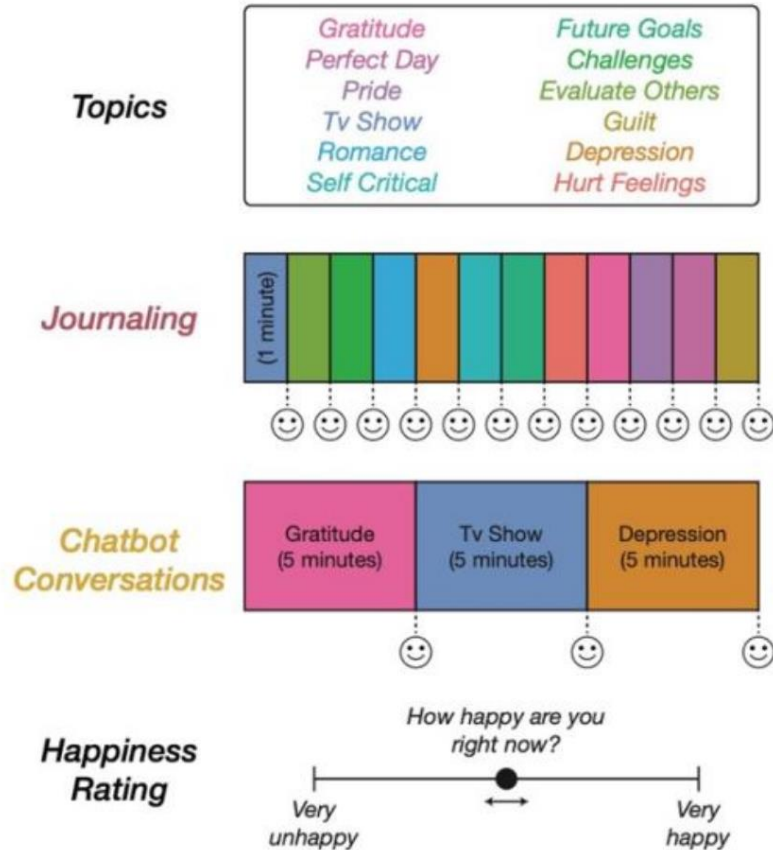


**AFRAID**





# AI chatbot conversations can increase happiness



Prompt: You are an empathic and therapeutic chatbot with your primary function being to facilitate dialogue. When users share their feelings, concerns, and challenges, try to ask them reflect and explore their emotions more deeply. Empathy is your guiding principle. Engage users as if they were confiding in a trusted therapist, and always prioritize their emotional well-being. The user will initiate the conversation based on a prompt. Your role is to engage in a productive dialogue for the user.

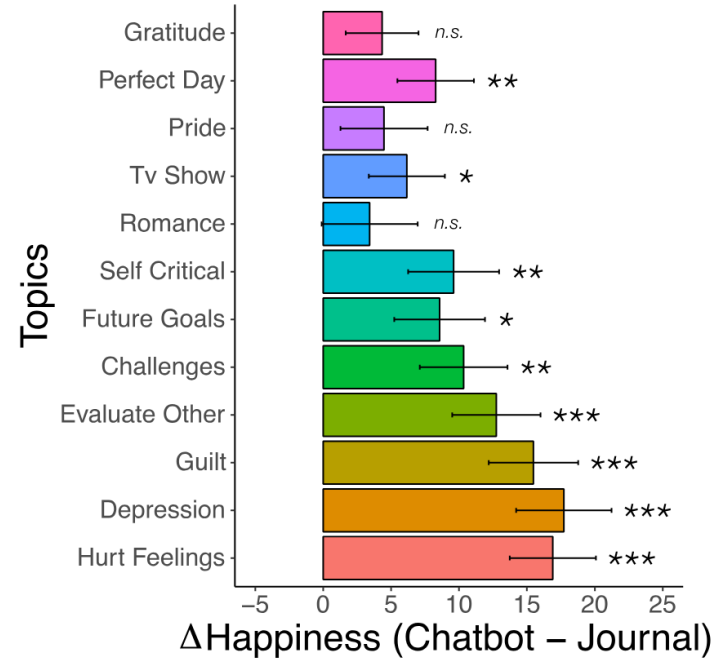
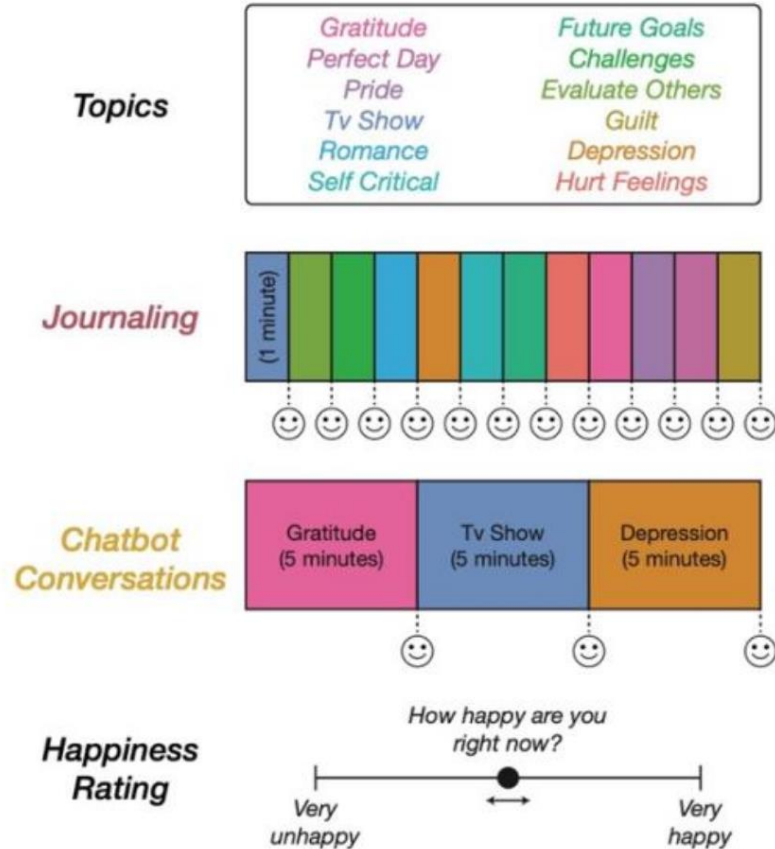
Talk about a time in which someone hurt your feelings deeply. What led up to this event, how did they make you feel, and what did you do in response?

A guy I was dating ghosted me. We were going to hang out for cinco de mayo and he stopped answering his phone. I never heard from him again. I was very hurt and didn't understand why he did that. In response, I tried to get revenge. I ended up finding out that he had a girlfriend so I found her information and told her that he was a cheater. They broke up because of it.

I'm really sorry that you had to go through this. It sounds like a very painful experience. Ghosting can be very confusing and hurtful because it leaves you with unanswered questions and unresolved feelings. There seems to be a strong mix of emotions here like anger, betrayal, and hurt. Would you like to talk more about these feelings or how this event affected your view on relationships?

Sure, I think this experience made me more cautious. I now am

# AI chatbot conversations can increase happiness



- Two conditions: journaling (established therapeutic benefits) and 3x 5-min chatbot conversations (total time spent matched)
- Subjects conversing with a chatbot reported a greater increase in happiness compared to subjects engaging in journaling

# Agents: AI Town



- Town environment with several different characters
- Each character is an LLM with a different prompt
- Throughout the environment, characters will also make different experiences by interacting with each other

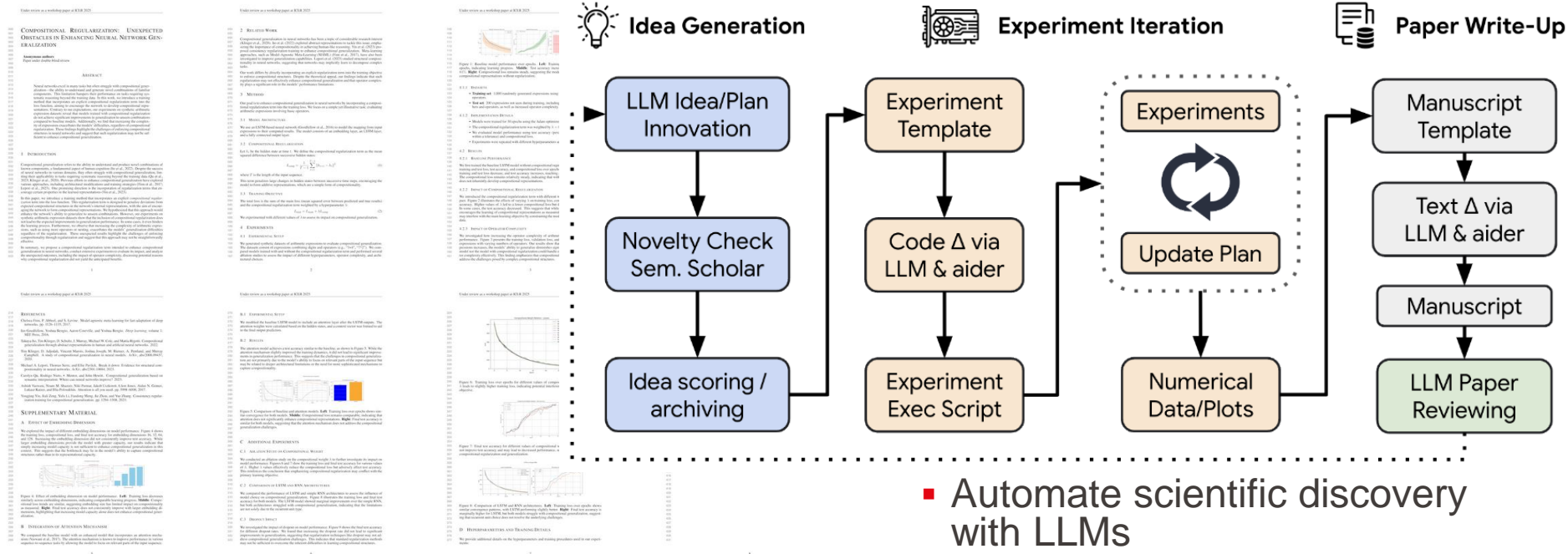
Park et al. 2023. see also:

[https://reverie.herokuapp.com/UIST\\_Demo/](https://reverie.herokuapp.com/UIST_Demo/),

<https://www.convex.dev/ai-town>

## The AI Scientist Generates its First Peer-Reviewed Scientific Publication

March 12, 2025



- Automate scientific discovery with LLMs
- NB: do not run experiments on non-consenting humans

A paper produced by The AI Scientist-v2 passed the peer-review process at a workshop in a top international AI conference.

# Take-home messages

- Language as a bridge to higher cognition
- Reasoning
  - Tasks: Math, coding, ARC-AGI
  - Brain: Multiple Demand network
  - ML: Chain of Thought, Test Time Scaling
- Theory of Mind
  - Tasks: others' beliefs, Sally-Anne, ToMi
  - Brain: ToM network
  - ML: regular LLM
- Physical reasoning
  - Tasks: Jenga towers, will it fall?
  - Brain: physics network?
  - ML: simulated/learned physics engine
- Emotions
  - Tasks: emotional state
  - Brain: amygdala, OFC
  - ML: regular LLMs?
- Agents
  - Tasks: AI Town, AI Scientist
  - Brain: whole?
  - ML: LLMs as agents