Computational Cognitive Science

Lecture 18: Large-scale models

Benjamin Peters

School of Informatics

University of Edinburgh

13 November, 2025

Models

Many of the models we have focused on so far

- have few parameters
- are interpretable
- can be fitted to reaonsable data sizes in simple experiments.

Models

Advances in ML and hardware and big datasets have made it feasible to develop and train much larger models.

Large-scale models in ML and across sciences

Models in ML are trained on vast datasets with the purpose to apply it across a wide range of tasks. These models are often called *foundation models*, because they provide the foundation for other down-stream tasks.

- LLMs (e.g., GPT or Llama series)
- Image generative models (e.g., Stable Diffusion)
- music, robotics, astronmy, radiology, genomics, coding, math, chemistry

Typically self-supervised pre-training on enormous datasets (Stable Diffusion: ~ 2.3 billion images, GPT-4: ~ 2 billion pages of text). Then specialized fine-tuning for specific tasks with much smaller dataset.

Large-scale models in cognitive science?

- Can we make use the large-scale approach in cognitive science?
- Can large-scale models help us understand human cognition?
- We may want to build a single model of cognition rather than many small domain-specific model (Allen Newell[1])

Today, we will look at very recent large-scale models and data approaches in cognitive science (and neuroscience).

[1] Newell, Allen. 1990. Unified Theories of Cognition. Harvard University Press, Cambridge, Massachusetts.

Reading

- Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., ... Schulz, E. (2025). A foundation model to predict and capture human cognition. Nature, 644(8078), 1002–1009. https://doi.org/10.1038/s41586-025-09215-4

Large-scale dataset

Psych-101 dataset

- The authors combined datasets from 160 tasks across 60k participants. 10 million choices.
- Tasks: multi-arm bandit task, decision-making, supervised learning, ...
- Tasks were trascribed into language

Large-scale dataset

Example for tasks

PSYCH101

Multi-armed bandits

In this task, you have to repeatedly choose between two slot machines labelled B and C. When you select one of the machines, you will win or lose points. Your goal is to choose the slot machines that will give you the most points.
You press <<C>> and get –8 points.
You press <<B>> and get 0 points.
You press <<B>> and get 1 points.

Decision-making

You will choose from two monetary lotteries by pressing N or U. Your choice will trigger a random draw from the chosen lottery that will be added to your bonus.
Lottery N offers 4.0 points with 80.0% or 0.0 points with 20.0%.
Lottery U offers 3.0 points with 100.0%.
You press <<U>>.

Memory

You will view a stream of letters on the screen, one letter at a time. You have to remember the last two letters you saw since the beginning of the block. If the letter you see matches the letter two trials ago, press E, otherwise press K.
You see the letter V and press <<K>>.
You see the letter X and press <<K>>.
You see the letter V and press <<E>>.

Supervised learning

In each trial, you will see between one and three tarot cards. Your task is to decide if the combination of cards presented predicts rainy weather (by pressing P) or fine weather (by pressing L).
You are seeing the following: card 3, card 4. You press <<L>>. You are right, the weather is fine.
You are seeing the following: card 1, card 4. You press <<P>>. You are right, the weather is rainy.

Markov decision processes

You will be taking one of the spaceships F or V to one of the planets M or S. When you arrive at each planet, you will ask one of the aliens for space treasure.
You are presented with spaceships V and F.
You press <<V>>. You end up on planet M and see aliens G and W. You press <<G>>.
You find 1 piece of space treasure.

Miscellaneous

You will be presented with triplets of objects, which will be assigned to the keys E, Z, and B. In each trial, please indicate which object you think is the odd one out by pressing the corresponding key.
E: tablet, Z: fox, and B: vent. You press <<Z>>.
E: ivy, Z: coop, and B: drink. You press <<B>>.
E: kite, Z: flan, and B: jar. You press <<E>>.
E: wand, Z: flag, and B: fire. You press <<Z>>.

Pretrained LLM

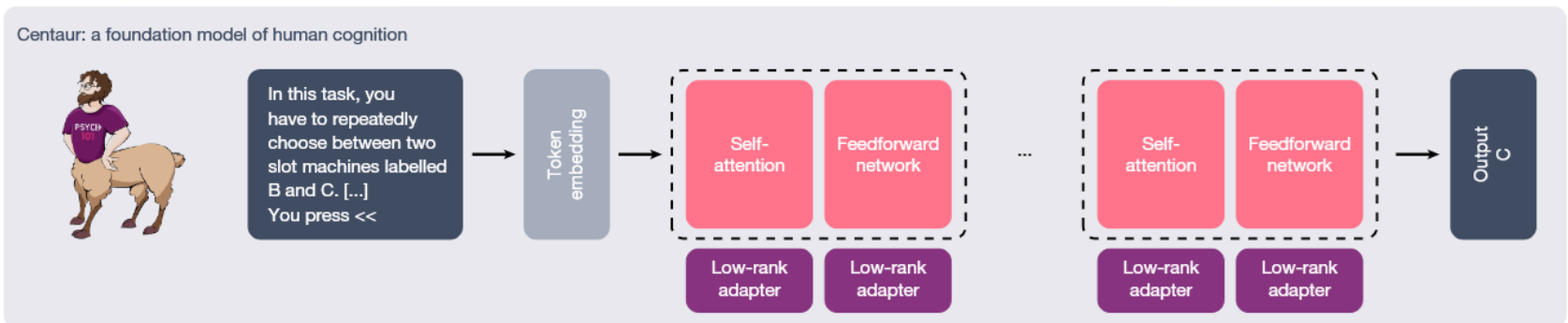
The authors took a pre-trained large language model (LLama 3.1 70B) from Meta. Model training:

- trained on next-token prediction ($\sim$ 15 Trillion tokens, 25 billion pages of text).
- supervised fine-tuning to create human-like responses on prompts.
- reinforcement learning to produce responses that humans prefer.

Centaur

The authors create a fine-tuned version of the Llama model they call ‘Centaur’.

- Take the Llama model and continue training on the Psych-101 dataset (‘fine-tuning’).
- Instead of training all 70B parameters, they only train particular ‘dimensions’ of the model (low-ranking adaptation).
- This adds 0.15% new parameters (i.e., 255M new parameters).



Models

How well does their model compare against established cognitive models. They compare three models:

- Off-the shelve Llama model
- Centaur model
- A domain-specific cognitive model

Which model better captures human responses on the tasks?

Domain-specific cognitive models

These are classic cognitive models for each task. For example, for bandit tasks they use the RW-model:

$$p(c_t = i) \propto \exp\left(aV_{i,t} + bS_{i,t} + cI_{i,t}\right)$$

$$V_{i,t} = \begin{cases} V_{i,t-1} + \text{sigmoid}\left(\alpha^+\right)\left(r_{t-1} - V_{i,t-1}\right) & \text{if } c_{t-1} = i \text{ and } r_{t-1} - V_{i,t-1} \geq 0 \\ V_{i,t-1} + \text{sigmoid}\left(\alpha^-\right)\left(r_{t-1} - V_{i,t-1}\right) & \text{if } c_{t-1} = i \text{ and } r_{t-1} - V_{i,t-1} < 0 \\ V_{i,t-1} & \text{otherwise} \end{cases}$$

$$S_{i,t} = \mathbb{1}\left[c_{t-1} = i\right]$$

$$I_{i,t} = \sum_{k=1}^{t-1} \mathbb{1}\left[c_k = i\right]$$

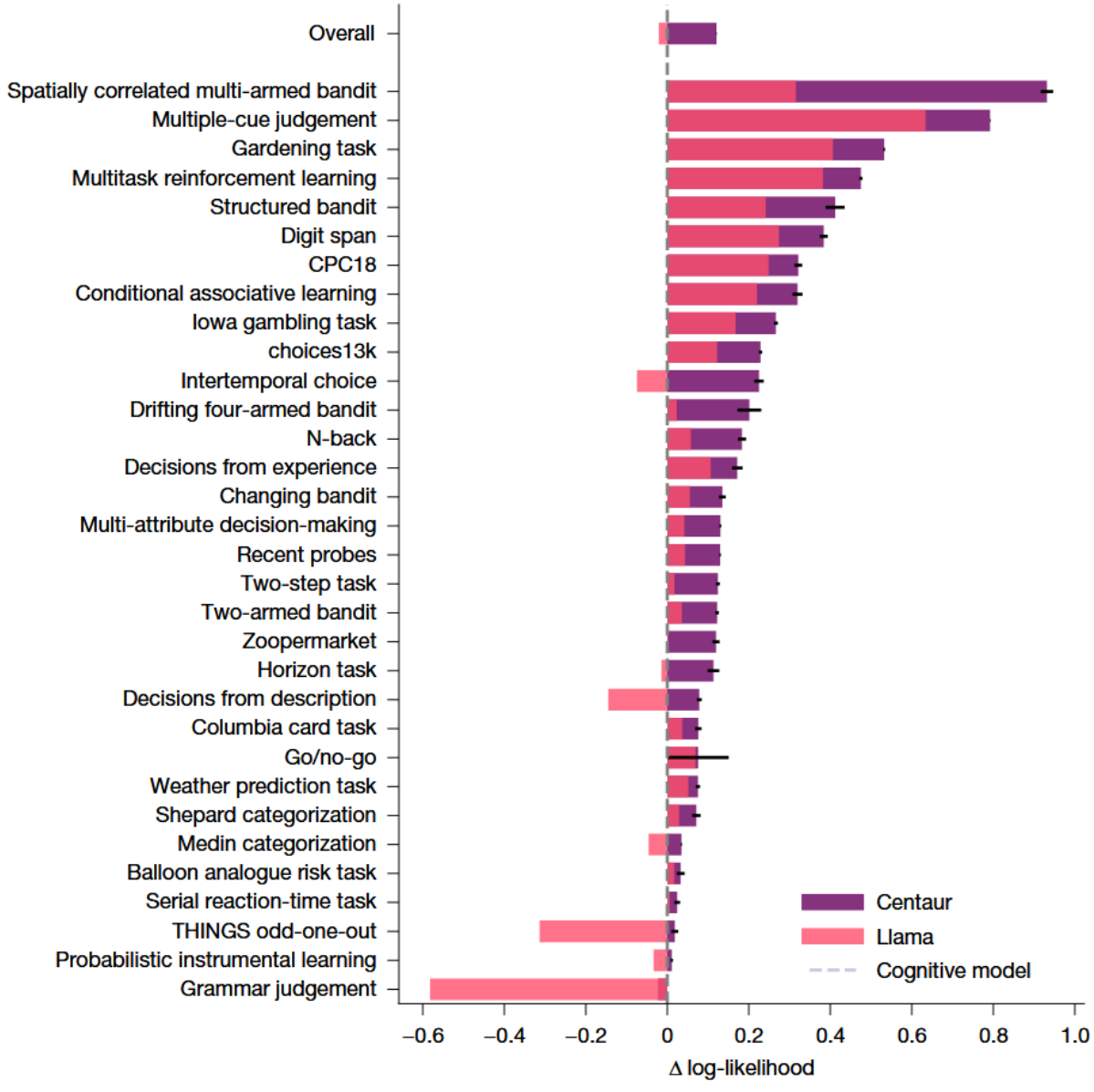
$$V_{i,1} = d$$

$$S_{i,1} = 0$$

$$I_{i,1} = 0$$

where r_t is the reward obtained in trial t. α^+, α^-, a, b, c, and d are free parameters of the model.

Model comparison: held-out subjects



Model comparison: held-out subjects

- Centaur, finetuned with human responses, always outperformed the Llama model on the held-out subjects
- Critically, Centaur outperforms the domain-specific cognitive model on all but one task.

Model comparison: held-out subjects

Large-scale models (with enough data) are incredibly good at learning distributions.

- Held-out subjects still come from the same distribution as the training subjects.
- Perhaps, Centaur "mimicks" subjects from the training distribution without truly capturing the principles by which the human responses were generated?

Stronger generalization: new tasks.

Model comparison: held-out tasks

Example: Two-step task (training)

You are participating in a space treasure game. In this game, you will be visiting two alien planets in search of treasure. Each planet has two aliens on it. The blue aliens live on the blue planet. The red aliens live on the red planet. When you visit a planet, you can choose an alien to trade with by pressing the corresponding button. When you trade with an alien, it will either give you treasure or junk. Your goal is to figure out, and trade with, the aliens that are most likely to give you treasure. To visit a planet, you will choose one rocket ship from two by pressing the corresponding button. They have different designations. Each rocket ship has a planet it will fly to most of the time. But sometimes they will take you to the other planet! Remember the following hints: 1. How likely an alien is to give you treasure will change over time, but this change will be slow. 2. Whether you get treasure depends only on the alien you choose to trade with. 3. If there is an alien you want to trade with, remember to pick the rocket ship that is most likely to take you to that alien's planet

You are presented with two spaceships called S and C. You press <>. You end up on the blue planet. You see a blue alien named D and a blue alien named R. You press <>. You find junk.

Model comparison: held-out tasks

Two-step task (modified)

You are playing the role of a musician living in a fantasy land. You play the flute for gold coins to an audience of genies, who live inside magic lamps on Pink Mountain and Blue Mountain. Pink Mountain has genies H and J, and Blue Mountain has genies A and E. Each genie lives in a lamp with the corresponding letter on it. When you arrive on a mountain, you can pick up a lamp and rub it. If the genie is in the mood for music, he will come out of his lamp, listen to a song, and give you a gold coin. Each genie's interest in music changes with time. To go to the mountains, you chose one of two magic carpets, which you purchase from a magician, who enchants them to fly. Magic carpet K generally flies to Pink Mountain, and magic carpet O generally flies to Blue Mountain. However, on rare occasions a strong wind blowing from that mountain makes flying there too dangerous because the wind might blow you off the carpet. In this case, the carpet is forced to land instead on the other mountain. You can take a magic carpet or pick up a lamp and rub it by pressing the corresponding key. Your goal is to get as many coins as possible over the next 201 days.

You are presented with magic carpets K and O. You press <>. You end up on Pink Mountain. You see lamp H and lamp J. You rub lamp <>. You receive 0 coins.

Summary

- Held-out subjects: Centaur model outperformed Llama and domain-specific cognitive models on held-out subjects.
- Held-out tasks: Centaur model also outperformed Llama and domain-specific cognitive models on a set of these held-out (modified) tasks.

What do you think?

A model of human cognition?

Critcisms raised (e.g., Bowers et al., 2025, link):

- theory-free
- No severe testing: does not actually predict human behaviour well on tasks outside the training set:
  - Centaur was evaluated on digit-span tasks of a certain span length on which it captured human responses better than domain-specific cognitive models.
  - But Centaur can retain lists up to 64 perfectly. It also fails a non-human like way (lists are either fully retained or forgotten).
  - They also instructed the model to have perfect memory.

Large-scale predictive models

How can we learn from big data and big models about human cognition?

As models:

- In principle proof: showing that certain properties can be learned from raw data. Nature vs. Nurture debate.
- As another system to be studied: If these models possess certain abilities, we can study them with tools of cognitive science. Like cross-species comparison, cross-domain (animal vs. machine) gives us a better understanding of cognition in general and about animals/humans specifically. Questions that can be asked: what kind of data does an agent need to express a certain behaviour? What kind of neural architecture? What are the goals/objectives of an agent?
- Hypothesis generator: All parts of the models are accessible and can be intervened on. In principle, we can discover new algorithms of cognition, which themselves can serve as hypotheses for human/animal cognition.

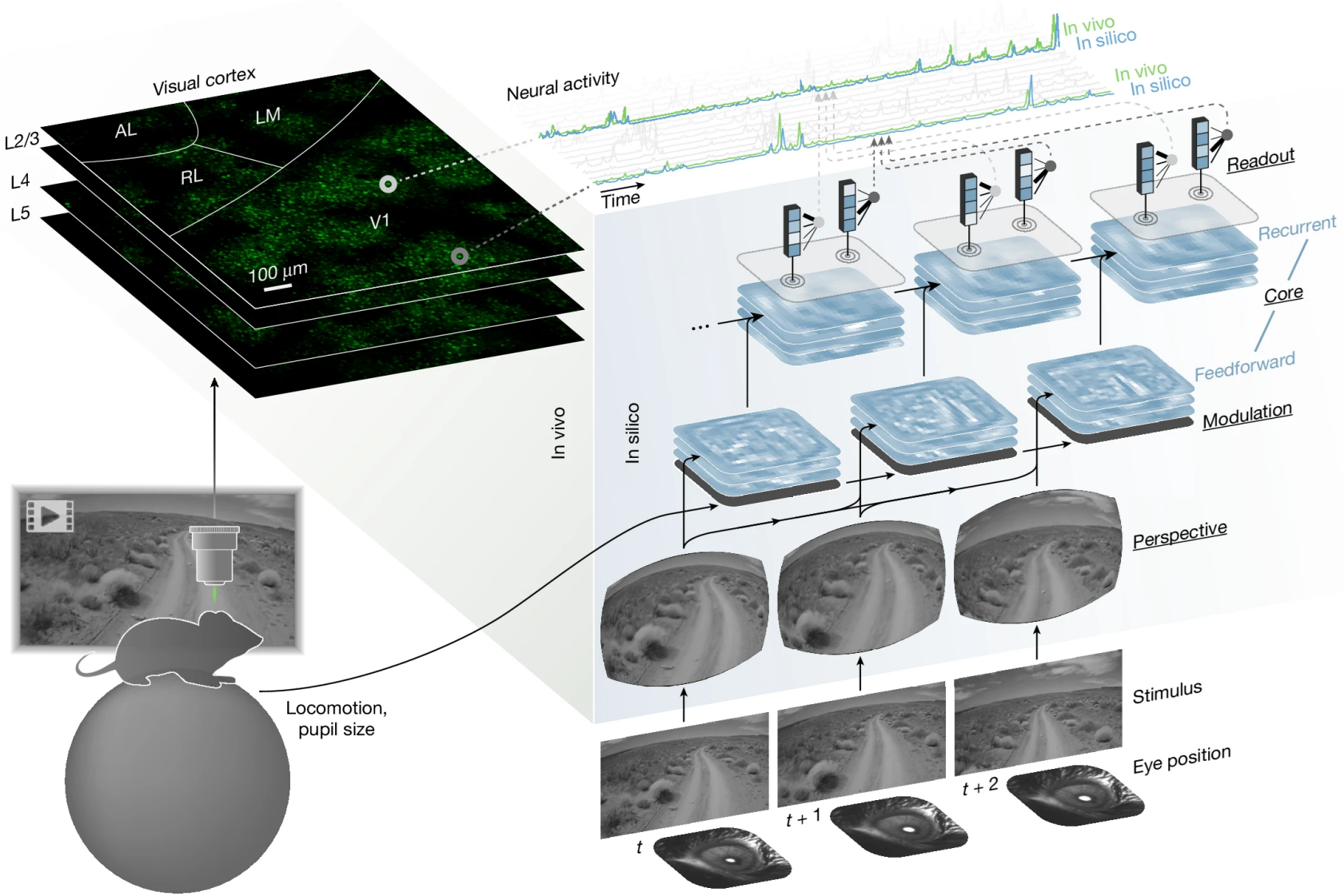
Large-scale predictive models

How can we learn from big data and big models about human cognition?

As tools:

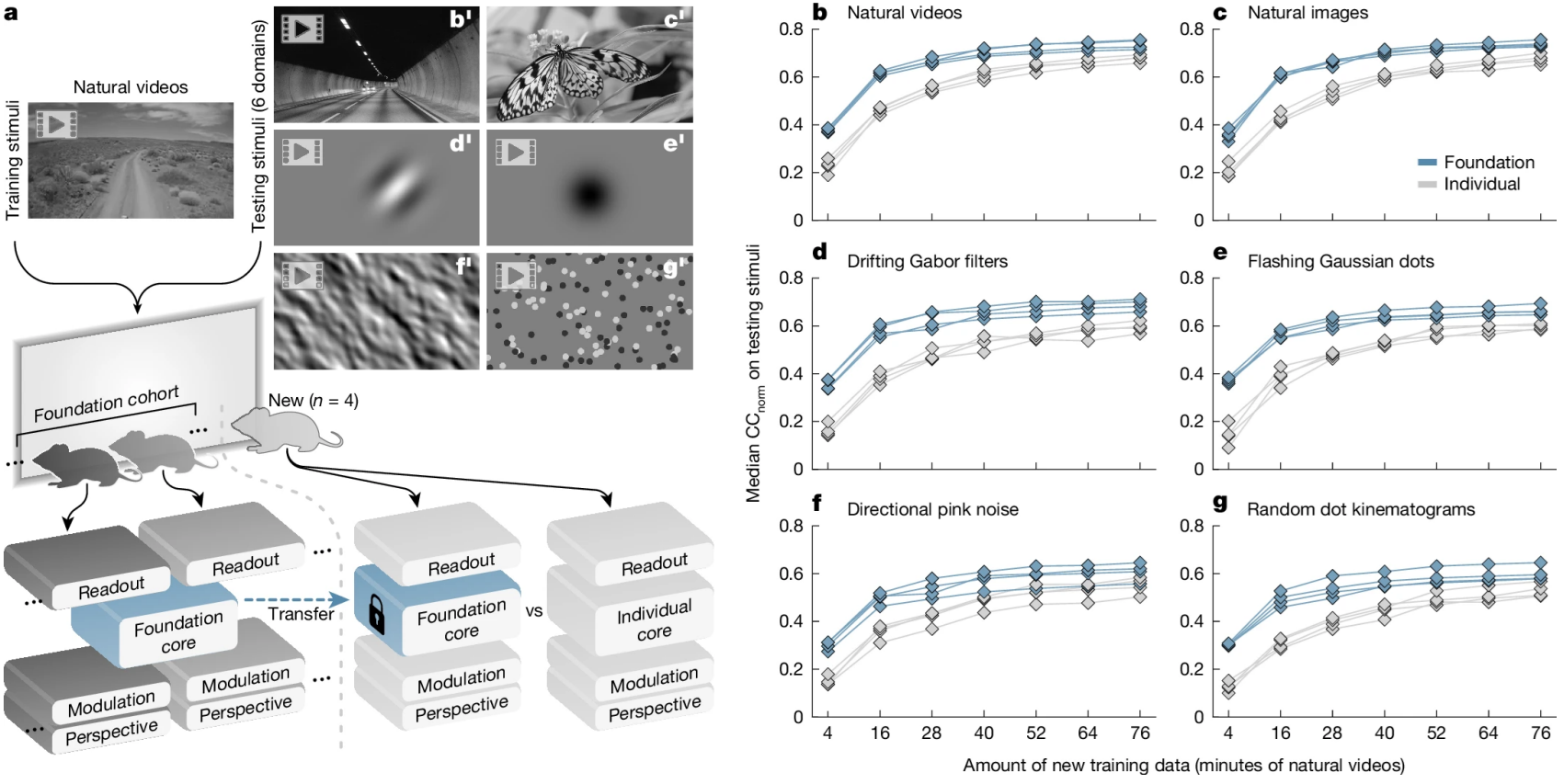
- Can generate novel hypotheses/ideas. Provide inspiration, out-of-the box thinking.
- LLMs can help to quantify qualtitative research. Image/video generative models can create novel stimuli.
- Powerful black-box models can provide estimates on the noise ceiling (i.e., what is the explainable variance in the data).

Neural Foundation Models



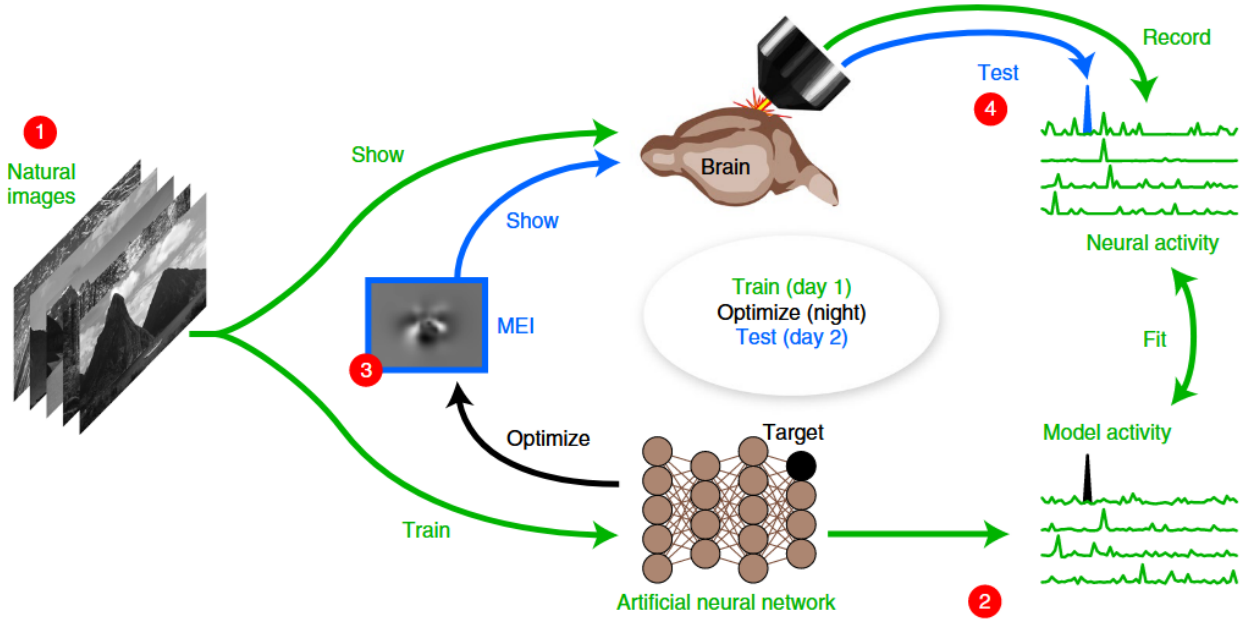
Wang et al., 2025, Foundation model of neural activity predicts response to new stimulus types, link

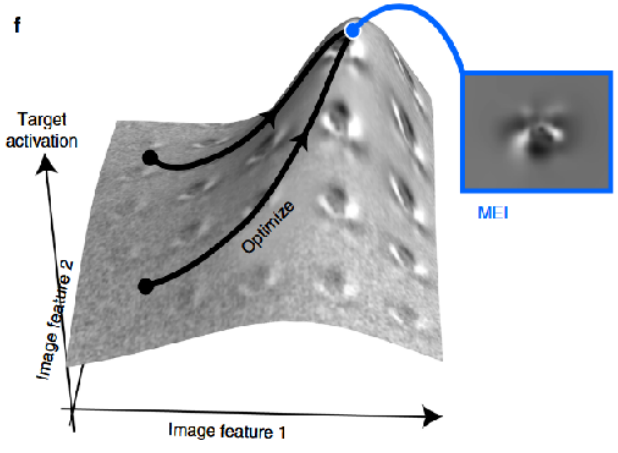
Neural Foundation Models

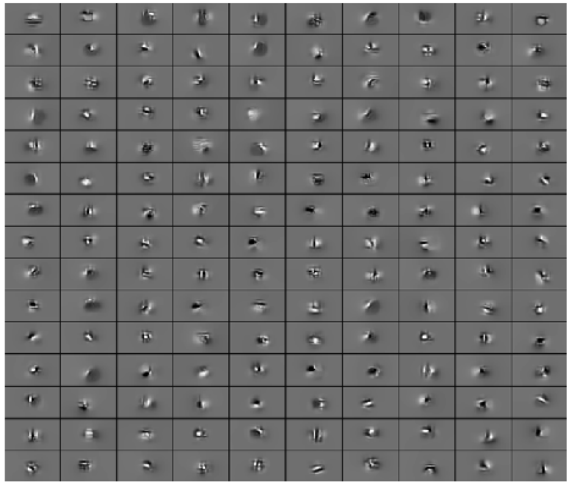


Wang et al., 2025, Foundation model of neural activity predicts response to new stimulus types, link

Discovery of most exciting images







References

- Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., Modirshanechi, A., Nath, S. S., Peterson, J. C., Rmus, M., Russek, E. M., Saanum, T., Schubert, J. A., Schulze Buschoff, L. M., Singhi, N., Sui, X., Thalmann, M., Theis, F. J., Truong, V., Udandarao, V., Voudouris, K., Wilson, R., Witte, K., Wu, S., Wulff, D. U., Xiong, H., & Schulz, E. (2025). A foundation model to predict and capture human cognition. Nature, 644(8078), 1002–1009. https://doi.org/10.1038/s41586-025-09215-4

- Bowers, J. S., Puebla, G., Thorat, S., Tsetsos, K., & Ludwig, C. J. H. (2025). Centaur: A model without a theory (No. v9w37_v2). PsyArXiv. https://doi.org/10.31234/osf.io/v9w37_v2

- Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. Nature Neuroscience, 22(12), 2060–2065. https://doi.org/10.1038/s41593-019-0517-x

- Wang, E. Y., Fahey, P. G., Ding, Z., Papadopoulos, S., Ponder, K., Weis, M. A., Chang, A., Muhammad, T., Patel, S., Ding, Z., Tran, D., Fu, J., Schneider-Mizell, C. M., da Costa, N. M., Reid, R. C., Collman, F., da Costa, N. M., Franke, K., Ecker, A. S., Reimer, J., Pitkow, X., Sinz, F. H., & Tolias, A. S. (2025). Foundation model of neural activity predicts response to new stimulus types. Nature, 640(8058), 470–477. https://doi.org/10.1038/s41586-025-08829-y