Computational Cognitive Science

Lecture 18: Large Cognitive Models 1

Frank Keller

19 November 2026

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Slide credits: Benjamin Peters

Overview

From Small to Large Models

The Centaur Model

Evaluating Large Cognitive Models

Assessment

Neural Foundation Models

Reading: Binz et al. (2025).

From Small to Large Models

Models

Many of the models we have focused on so far

- have few parameters
- are interpretable
- can be fitted on small datasets.

But advances in AI, hardware, and big datasets have made it feasible to develop and train much larger models.

Large models in AI and across the sciences

AI models are trained on vast datasets and can be applied across a wide range of tasks. These models are sometimes called foundation models, as they provide a foundation for many downstream tasks.

- LLMs (e.g., GPT or Qwen series)
- Image generation models (e.g., Stable Diffusion)
- models for music, robotics, astronomy, radiology, genomics, coding, math, and chemistry

Typically, models are first pre-trained on enormous datasets and then specialized through fine-tuning on a much smaller dataset.

Large models in cognitive science?

The pre-training/fine-tuning approach is general. Can we apply it in cognitive science?

- Can we scale up our models dramatically?
- Can large models help us understand cognition?
- Does this allow us to build a unified model of cognition, rather than many small domain-specific models?
- This has been the dream in AI/Cogsci for a long time (Newell 1990).

In Lecture 17, we used LLMs to compute expectations in sentence processing. But we found that better prediction does not necessarily mean better cognitive modeling.

Today we ask a broader version of the same question: Can large foundation models become general models of human cognition and neural activity?

The Centaur Model

Large dataset

Binz et al. (2025) create the Psych-101 dataset:

- Combines data from 160 tasks across 60k participants; 10 million choices.
- Tasks include multi-armed bandits, decision-making, memory, supervised learning.
- Tasks were transcribed into language to make them understandable to language models.

(Multi-armed bandits are experimental paradigms used to study decision making under uncertainty.)

Large 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).

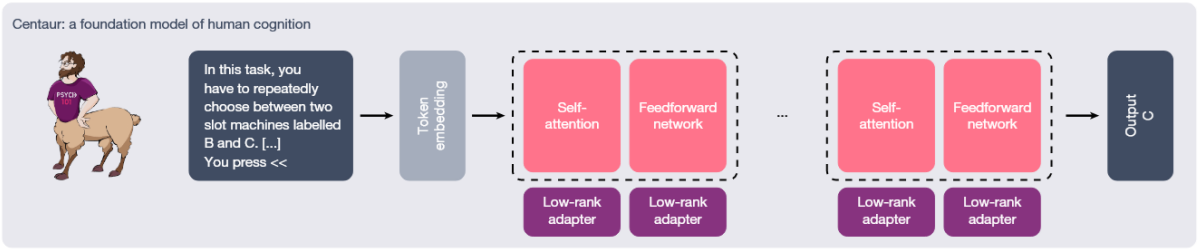
This model has already learned a lot of things:

- 70 billion parameters;
- pretrained on roughly 15 trillion tokens of text using next-token prediction;
- the pretraining gives the model broad linguistic and world knowledge.

Centaur

The authors create a fine-tuned version of the Llama model called Centaur.

- Take the pre-trained Llama model and fine-tune it on the Psych-101 dataset.
- Instead of updating all 70B parameters, use quantized low-rank adaptation (QLoRA) and train only small adapter matrices.
- This adds 0.15% new trainable parameters (i.e., 105M parameters).



Evaluating Large Cognitive Models

Models

How well does their model compare against established cognitive models?

Binz et al. (2025) compare three models:

- Off-the-shelf Llama model (no fine-tuning)
- Centaur model
- A domain-specific cognitive model

Which model better captures human responses on the tasks?

They report the difference in log-likelihood (ΔLL) of Centaur and Llama relative to a domain-specific cognitive model for each experiment (zero line in the graph).

Domain-specific cognitive models

The domain-specific models are classic cognitive models for each task. For example, for multi-armed bandits they use the Rescorla-Wagner model:

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

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

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

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

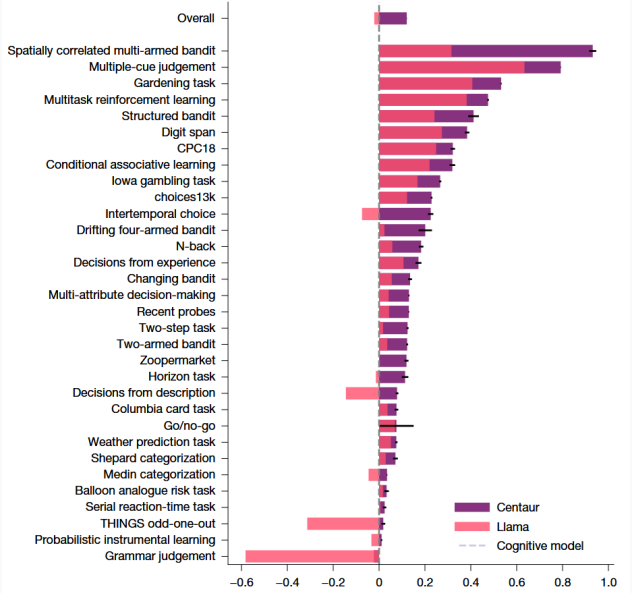
$$V_{i,1} = d$$

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

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

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 participants (ΔLL)



Model comparison: held-out participants

The authors test fine-tuned Centaur on held-out participants from the same tasks and datasets that were used for fine-tuning:

- Centaur outperformed Llama (no fine-tuning) on all tasks.
- Centaur outperforms the domain-specific cognitive model on all but one task.

Interestingly, the exception is the grammar-judgment task.

Time for a short quiz on WooClap!



Model comparison: held-out participants

Large models (when trained with enough data) are incredibly good at learning distributions.

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

Stronger generalization: Evaluation on 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.

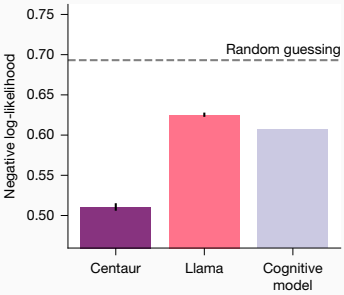
Model comparison: held-out tasks

Two-step task (modified)

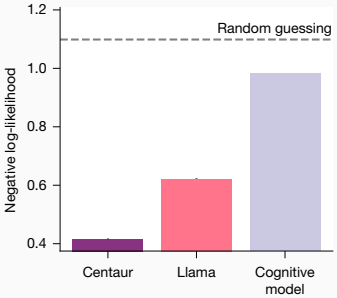
You are playing the role of a musician living in a fantasy land. You play the flute for gold coins to an audience of genies, who live inside magic lamps on Pink Mountain and Blue Mountain. Pink Mountain has genies H and J, and Blue Mountain has genies A and E. Each genie lives in a lamp with the corresponding 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.

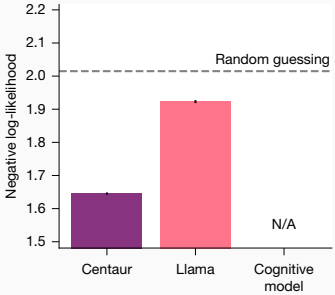
Model comparison: held-out tasks



Two-step task



Three-armed bandit task



Logical reasoning task

Centaur model also outperformed Llama and domain-specific cognitive models on several held-out (modified or new) tasks.

But in Lecture 17, we saw that predictive accuracy and cognitive plausibility are different criteria. Similar criticism applies here.

Assessment

Criticisms of Centaur

Bowers et al. (2026) raise three main concerns:

- Insufficiently severe tests: Centaur performs well on many held-out data points, but may not capture the theoretically important effects and qualitative patterns that distinguish competing cognitive theories.
- Lacks basic human mechanistic constraints: The model can show strongly non-human behavior under simple manipulations, e.g., unrealistically large memory capacity or following implausible instructions.
- Prediction is not explanation: Even perfect prediction would not by itself reveal how human cognition works.

Centaur may be a powerful predictive model without being a theory of cognition.

Large predictive models

So what do we learn from big data and big models about human cognition?

- Proof in principle: showing that certain properties can be learned from raw data. Nature vs. nurture debate.
- Another system to be studied: If these models possess certain abilities, we can study them with the tools of cognitive science. Like cross-species comparison, this should give us a better understanding of cognition in general:
  - What kind of data does an agent need for a certain behavior?
  - What kind of neural architecture?
  - Which goals/objectives?
- 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.

Neural Foundation Models

A Foundation Model of Visual Cortex

Wang et al. (2025) develop a foundation-model idea of neural activity.

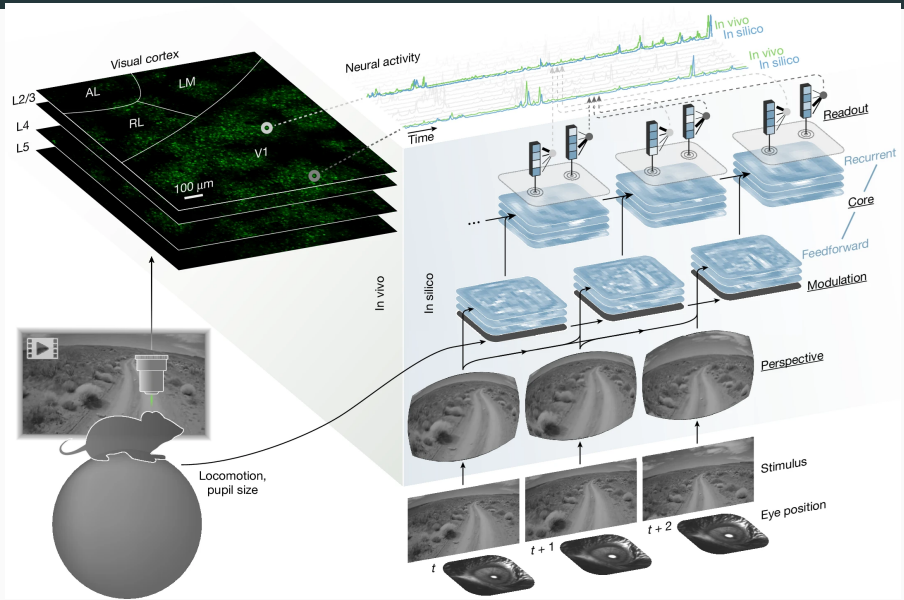
The model is trained to predict neural responses from:

- natural visual input;
- eye position;
- locomotion and pupil size;
- neuron identity and cortical location.

$$\boxed{\text{stimulus} + \text{behavior}} \longrightarrow \boxed{\text{neural activity}}$$

Can shared visual computations generalize across animals?

A Foundation Model of Visual Cortex



What Makes It a Foundation Model?

Instead of training a separate neural network for every mouse:

1. Pool neural responses to natural videos from several mice.
2. Train a shared foundation core:

$$\text{visual input} \longrightarrow \boxed{\text{shared latent representation}}$$

3. For a new mouse, freeze the foundation core.
4. Learn only mouse-specific components:
   - perspective/eye-position transformation;
   - behavioral modulation;
   - neuron-specific readouts.

Shared computation is learned once; individual differences require much less new data.

Generalization to New Mice and New Stimuli

The foundation core is transferred to new mice and tested on several stimulus classes:

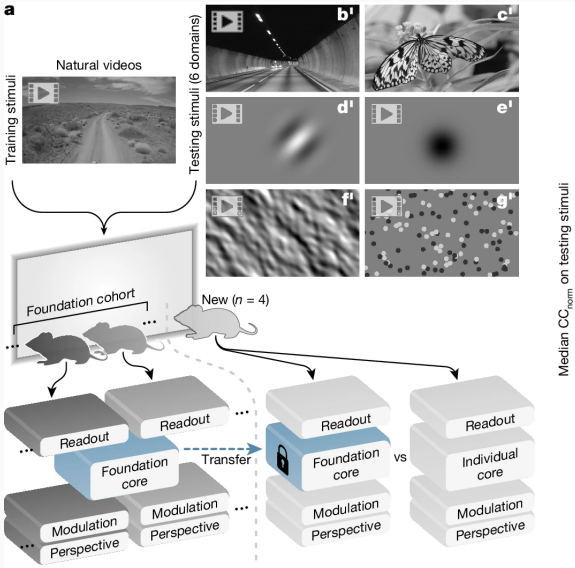
- natural videos and images;
- drifting Gabor filters;
- flashing dots;
- visual noise and motion stimuli.

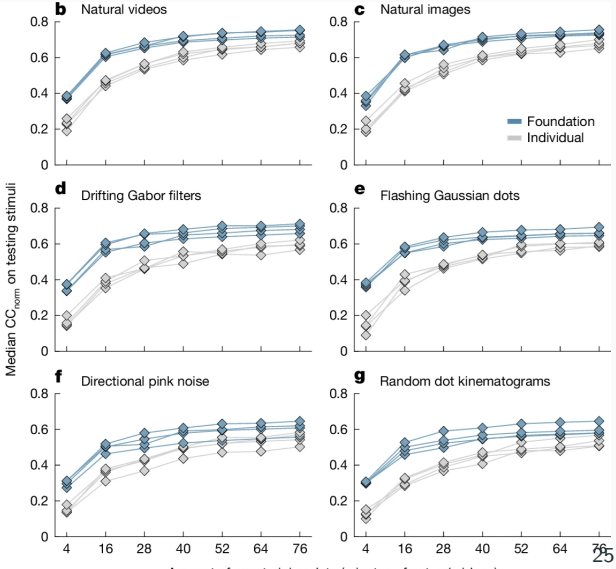
Compared with models trained separately for each mouse, the foundation model:

- needs less new neural data;
- predicts responses more accurately;
- transfers to new stimulus types.

The learned representation transfers across both subjects and stimulus domains.

Generalization to New Mice and New Stimuli





Summary

- Large cognitive models use the same basic recipe as modern AI: pre-training on large datasets followed by adaptation to new tasks or individuals.
- Centaur fine-tunes an LLM on behavioral data from many different cognitive tasks and predicts human choices better than domain-specific models.
- But, as in Lecture 17, predictive accuracy does not guarantee cognitive plausibility: a model fits behavior well but may rely on non-human capacities or mechanisms.
- Large models can nevertheless be useful as:
  - proofs of what can be learned from data;
  - systems for comparison with human cognition;
  - generators of hypotheses about cognitive mechanisms.
- Wang et al. extend the foundation-model idea to neuroscience: shared visual-cortex representation can be adapted to new individuals and stimulus types.

References

Binz, Marcel, Elif Akata, Matthias Bethge, Franz Brändle, Frederick Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Seif Haridi, Ashwini K. Jagadish, Li Ji-An, Alon Kipnis, Sudeep Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Sreejan S. Nath, Joshua C. Peterson, Martin Rmus, Evan M. Russek, Tor Saanum, Julian A. Schubert, Lukas M. Schulze Buschoff, Nishant Singhi, Xinyu Sui, Matthias Thalmann, Fabian J. Theis, Viet Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Karoline Witte, Shuchen Wu, Dirk U. Wulff, Hao Xiong, and Eric Schulz. 2025. A foundation model to predict and capture human cognition. *Nature* 644(8078):1002–1009.

Bowers, Jeffrey S., Guillermo Puebla, Sushrut Thorat, Konstantinos Tsetsos, and Casimir J. H. Ludwig. 2026. On the misuse of prediction in NeuroAI: A case study of Centaur. PsyArXiv v9w37_v5. Preprint, Version 5.

Newell, Alan. 1990. *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.

Wang, Eric Y., Paul G. Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A. Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, Dat Tran, Jiakun Fu, Casey M. Schneider-Mizell, MICrONS Consortium, R. Clay Reid, Forrest Collman, Nuno Maçarico da Costa, Katrin Franke, Alexander S. Ecker, Jacob Reimer, Xaq Pitkow, Fabian H. Sinz, and Andreas S. Tolias. 2025. Foundation model of neural activity predicts response to new stimulus types. *Nature* 640(8058):470–477.