Computational Cognitive Science

Lecture 19: Large Cognitive Models 2

Frank Keller

23 November 2026

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

Overview

Beyond Predictive Accuracy

Learning Linguistic Abstractions

Experiment 1: Verb Classes

Experiment 2: Other Linguistic Constructions

LLMs as Human Proxies

Reading: Jian and Manning (2026).
Optional: Karetnikov et al. (2026), Frank and Goodman (2026).

Beyond Predictive Accuracy

From Lecture 18 to Lecture 19

Lecture 18 showed that large cognitive models can predict:

- human choices across many cognitive tasks (Binz et al. 2025)
- neural responses across animals and stimuli (Wang et al. 2025)

But we also saw a recurring problem:

$$\boxed{\text{good prediction}} \not\Rightarrow \boxed{\text{cognitive explanation}}$$

Today we ask a different question:

Can large models help us test hypotheses about how cognition functions are learned?

Large Models as Cognitive Experiments

There are several ways in which a large model can be useful for cognitive science.

- Behavioral prediction: Does it predict what humans do?
- Learning model: Does it acquire knowledge in a psychologically interesting way?
- Hypothesis test: Can it distinguish predictions made by competing theories?

Jian and Manning (2026) focus on the last two questions. Instead of only studying the final model, they study how behavior changes during learning.

This treats an LLM as an existence proof: if it acquires an ability from a particular learning setup, this demonstrates one possible route to that ability.

possible learning mechanism $\not\Rightarrow$ actual human learning mechanism

Learning Linguistic Abstractions

How Are Linguistic Categories Learned?

Jian and Manning (2026) study categorization in language (see lectures 9 and 10).

Linguistic knowledge involves generalizations over classes of words:

- transitive vs. intransitive verbs;
- verbs with different argument structures;
- verbs occurring in filler–gap constructions.

For example, different verb classes license different kinds of arguments:

(1) a. I gave a book to Emma. recipient
b. We went to the market. destination

The question is not just whether GPT-2 eventually learns these categories, but how they emerge during training.

Two Accounts of Language Learning

Jian and Manning contrast two broad accounts from research on child language acquisition.

Exemplar-first: individual examples $\longrightarrow$ gradual generalization $\longrightarrow$ abstract category

Children initially store item-specific patterns; abstractions emerge from accumulated exemplars.

Abstraction-first: abstract category $\longrightarrow$ item-specific refinement

Learners form class-level generalizations early and subsequently refine their knowledge of individual lexical items.

Why Study Learning Trajectories?

Looking only at a fully trained model cannot easily distinguish the two accounts.

Both may eventually produce correct category-sensitive behavior.

But they make different predictions about what happens first:

- Exemplar-first: differences between individual words should emerge before clear class-level structure.
- Abstraction-first: class-level distinctions should emerge as soon as, or before, item-specific differences.

Training checkpoints therefore turn model learning into an experimental trajectory.

Model and Training Data

Jian and Manning analyze checkpoints from GPT-2 Small:

- decoder-only Transformer;
- 12 layers, 12 attention heads;
- autoregressively trained on OpenWebText;
- 450 checkpoints spanning 400,000 training steps.

At each checkpoint they inspect the model's:

$$P(\text{next token} \mid \text{sentence prefix})$$

This lets them ask when particular distinctions first appear in the model's prediction distributions.

The training objective stays the same; only the learned model changes.

Experiment 1: Verb Classes

Experiment 1: Verb Classes

The first experiment considers four argument-structure classes.

- to-Dative: give, sell, grant
- Motion: go, walk, flee
- Reciprocal: speak, meet, talk
- Spray-load: spray, load, smear

Example prefixes:

(2) a. Chipotle gave away free burritos to the ___
b. Toby accordingly goes to the ___

The continuation should depend on the verb class: recipients for dative verbs, locations for motion verbs, etc.

Question: do predictions first differentiate classes or individual verbs?

Measuring When Categories Emerge

For each verb, the authors obtain a distribution over possible next tokens.

They compare two distributions using Jensen–Shannon divergence $D_{\mathrm{JS}}(P, Q)$.

For this lecture, the intuition is enough:

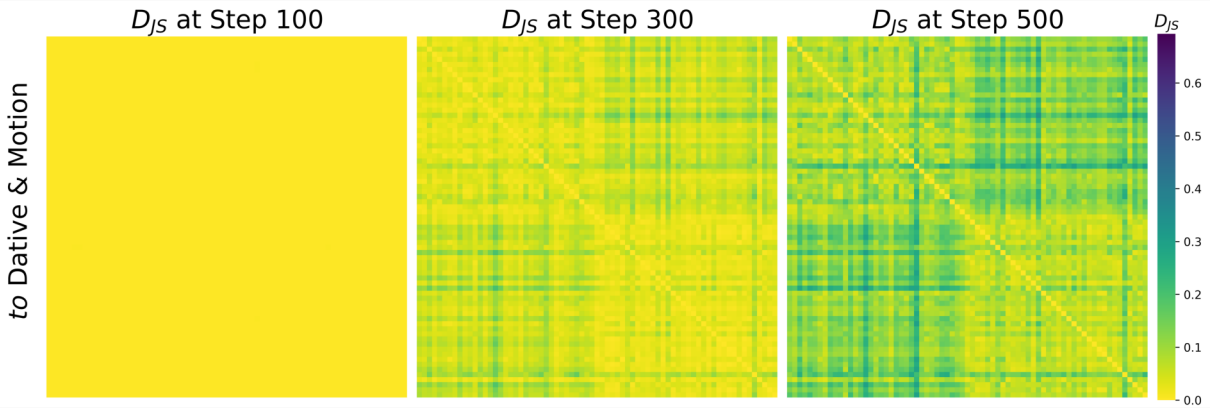
- small divergence: two contexts generate similar predictions;
- large divergence: the LM treats the contexts differently.

Two quantities matter:

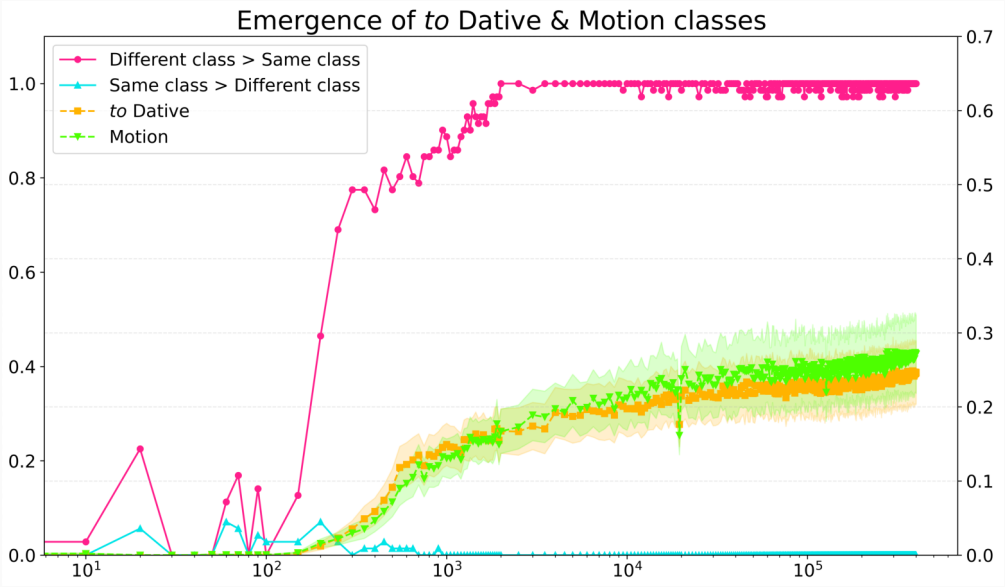
- within-class divergence: differences among individual verbs;
- between-class divergence: differences between verb classes.

If between-class structure appears early, that is evidence for abstraction-first learning.

Experiment 1: Visualizing Jensen–Shannon Divergence



Experiment 1: Difference between Classes vs. between Verbs



Result: Abstraction Comes First

The central result of Experiment 1 is: class-level differentiation appears before substantial item-specific differentiation.

As soon as GPT-2 begins to distinguish relevant continuations:

- verbs from different classes behave differently;
- verbs from the same class still behave relatively similarly;
- idiosyncratic knowledge about individual verbs develops later.

$$\boxed{\text{class-level learning}} \longrightarrow \boxed{\text{item-level refinement}}$$

This matters because both theories predict similar final behavior, but different learning trajectories. GPT-2's trajectory is consistent with abstraction-first learning.

Experiment 2: Other Linguistic Constructions

Experiment 2: Other Linguistic Constructions

Do the same learning dynamics occur beyond semantic verb classes?

Jian and Manning test two syntactic phenomena.

Transitivity

(3) a. The truck hadn't astounded ___
b. The truck hadn't fallen ___

Relative clauses/filler–gap dependencies

(4) a. The person that they sold the land to ___
b. The person thinks that they sold the land to ___

Now the continuation depends on the syntactic construction.

Question: do predictions first differentiate constructions or individual lexical items?

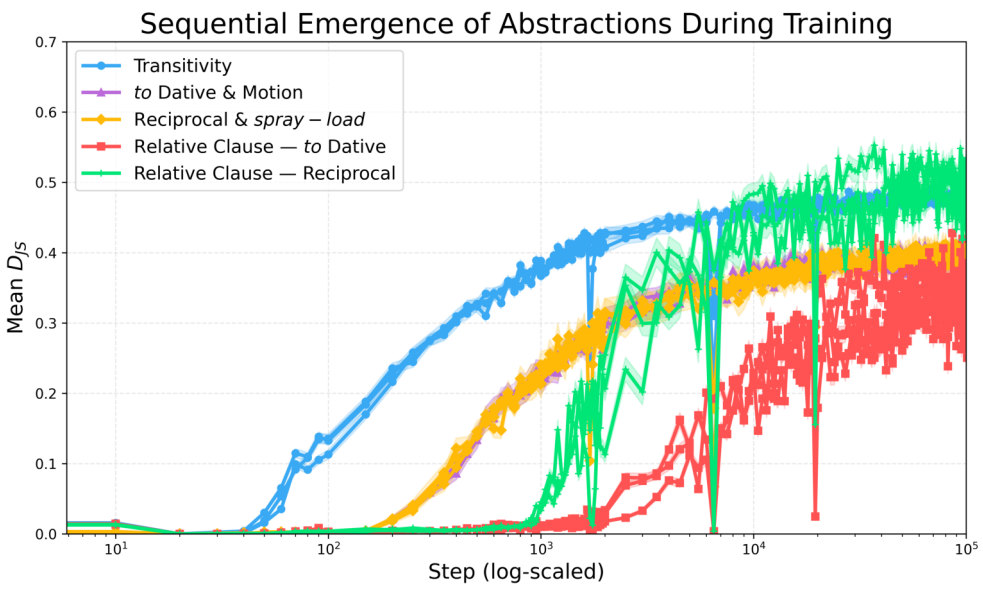
Result: Abstraction Generalizes Across Constructions

Across the phenomena tested:

- class-general behavior emerges as soon as item-level learning is detectable;
- constructions do not appear to be learned one lexical item at a time and generalized only later;
- the results therefore again favor an abstraction-first description of GPT-2 learning.

But there is another striking pattern: different linguistic abstractions emerge at different points in training.

Abstractions Emerge Sequentially



Connection to Human Language Acquisition

The result resembles an abstraction-first view of child language learning: structural regularities emerge early, lexical knowledge is refined subsequently.

Some abstraction-first theories propose that linguistic structure is innately specified.

But GPT-2 has no built-in linguistic categories. It therefore provides an existence proof:

- class-level behavior can emerge before item-specific knowledge;
- this can happen through distributional learning.

This does not show that children learn like GPT-2! It just shows that this type of learning is possible in principle.

Time for a short quiz on WooClap!



LLMs as Human Proxies

From Cognitive Models to Human Proxies

So far we have treated large models as tools for understanding human cognition.

But LLMs are increasingly used for a broader purpose: standing in for humans.

Examples include:

- characters in simulations;
- agents performing human tasks;
- artificial participants in experiments;
- synthetic respondents in surveys.

Karetnikov et al. (2026) argue that calling all of these systems "human-like" obscures crucial differences.

Four Roles for Human Proxies

Karetnikov et al. (2026) distinguish four roles:

a Believable agents



Non-playable characters in games
- Can LLMs simulate artificial society that reflects believable human behavior?



Synthetic social platforms
- Can LLMs simulate the interactions of an online community before it exists?

b Task agents



Software development
- Can LLM agents design, implement and test software through collaboration?



Scientific discovery
- Can LLM agents generate research hypotheses that experts validate in real laboratories?

c Experimental subjects



Alignment with moral norms
- Do LLMs show amplified cognitive biases in moral decision-making?



Capacity for theory of mind
- Can LLMs attribute and reason about mental states as humans do?

d Silicon samples



Prediction of human choices
- Can LLMs fine-tuned on behavioral data predict a person's next choice in an experiment?



Prediction of voting behavior
- Can LLMs reproduce the vote choices of demographic groups?

Each role makes a different claim about similarity to humans.

Four Roles for Human Proxies

Believable agents

- Goal: behave like a convincing character or social actor.
- Relevant comparison: whether people accept the behavior as plausible or believable.

Task agents

- Goal: perform work normally done by humans.
- Relevant comparison: whether the system performs the task competently.

Neither role requires the model to use human-like cognitive mechanisms.

A chess program can outperform humans without playing chess like a human; the same distinction applies to LLM agents.

Experimental Subjects and Silicon Samples

Experimental subjects

- The LLM itself is the object of study.
- We can administer experiments or psychometric tests to compare machine and human behavior.
- The inference is primarily about the model.

Silicon samples

- The model is intended to represent particular humans or a human population.
- Researchers may use generated responses instead of recruiting human respondents.
- The intended inference is about humans.

That second claim requires substantially stronger validation.

Validity Depends on the Role

The key message of Karetnikov et al. (2026): "Human-like" is not a single property.

Different uses require different human reference points.

Role	What needs to match?
Believable agent	perceived behavior
Task agent	task competence
Experimental subject	relevant machine–human effects
Silicon sample	human response distributions

A model may therefore be valid for one role and invalid for another.

Define the scientific claim first; then choose the validation test.

Summary

- Large models can contribute to cognitive science in ways that go beyond predicting human responses.
- Jian and Manning (2026) use learning trajectories to distinguish competing theories of category acquisition.
- GPT-2 shows an abstraction-first trajectory: class-level behavior emerges before extensive item-specific refinement.
- This provides an existence proof, not evidence that humans and GPT-2 use the same learning mechanism.
- Karetnikov et al. (2026) distinguish four roles in which LLMs can stand in for humans; each requires a different form of validation.

The scientific value of a large model depends on the claim we ask it to support.

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.

Frank, Michael C. and Noah D. Goodman. 2026. Cognitive modeling using artificial intelligence. *Annual Review of Psychology* 77(1):543–566.

Jian, Jasper and Christopher D. Manning. 2026. Humans and transformer LMs: Abstraction drives language learning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Rabat, Morocco, pages 752–765.

Karetnikov, Nikita, Iyad Rahwan, and Davor Svetinovic. 2026. Large language models as human proxies. *Nature Computational Science* 6(10):1057–1066.

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