Centaur: a foundation model of human cognition

Marcel Binz[1*], Elif Akata[1], Matthias Bethge[2], Franziska Brändle[3,5], Fred Callaway[4], Julian Coda-Forno[1], Peter Dayan[2,5], Can Demircan[1], Maria K. Eckstein[6], Noémi Éltető[5], Thomas L. Griffiths[7], Susanne Haridi[1,13], Akshay K. Jagadish[1,2,5], Li Ji-An[8], Alexander Kipnis[1], Sreejan Kumar[7], Tobias Ludwig[2,5], Marvin Mathony[1], Marcelo Mattar[4], Alireza Modirshanechi[1], Surabhi S. Nath[2,5,13], Joshua C. Peterson[9], Milena Rmus[1], Evan M. Russek[7], Tankred Saanum[5], Natalia Scharfenberg[5], Johannes A. Schubert[5], Luca M. Schulze Buschoff[1], Nishad Singhi[14], Xin Sui[2,5], Mirko Thalmann[1], Fabian Theis[1], Vuong Truong[5], Vishaal Udandarao[2,15], Konstantinos Voudouris[1], Robert Wilson[10], Kristin Witte[1], Shuchen Wu[1], Dirk U. Wulff[11,12], Huadong Xiong[10], Eric Schulz[1]

[1]Helmholtz Munich.
[2]University of Tuebingen.
[3]University of Oxford.
[4]New York University.
[5]Max Planck Institute for Biological Cybernetics.
[6]Google DeepMind.
[7]Princeton University.
[8]University of California San Diego.
[9]Boston University.
[10]Georgia Institute of Technology.
[11]University of Basel.
[12]Max Planck Institute for Human Development.
[13]Max Planck School of Cognition.
[14]TU Darmstadt.
[15]University of Cambridge.

*Corresponding author(s). E-mail(s): marcel.binz@helmholtz-munich.de;

Abstract

Establishing a unified theory of cognition has been a major goal of psychology [1, 2]. While there have been previous attempts to instantiate such theories by building computational models [1, 2], we currently do not have one model that captures the human mind in its entirety. Here we introduce Centaur, a computational model that can predict and simulate human behavior in any experiment expressible in natural language. We derived Centaur by finetuning a state-of-the-art language model on a novel, large-scale data set called Psych-101. Psych-101 reaches an unprecedented scale, covering trial-by-trial data from over 60,000 participants performing over 10,000,000 choices in 160 experiments. Centaur not only captures the behavior of held-out participants better than existing cognitive models, but also generalizes to new cover stories, structural task modifications, and entirely new domains. Furthermore, we find that the model's internal representations become more aligned with human neural activity after finetuning. Taken together, Centaur is the first real candidate for a unified model of human cognition. We anticipate that it will have a disruptive impact on the cognitive sciences, challenging the existing paradigm for developing computational models.

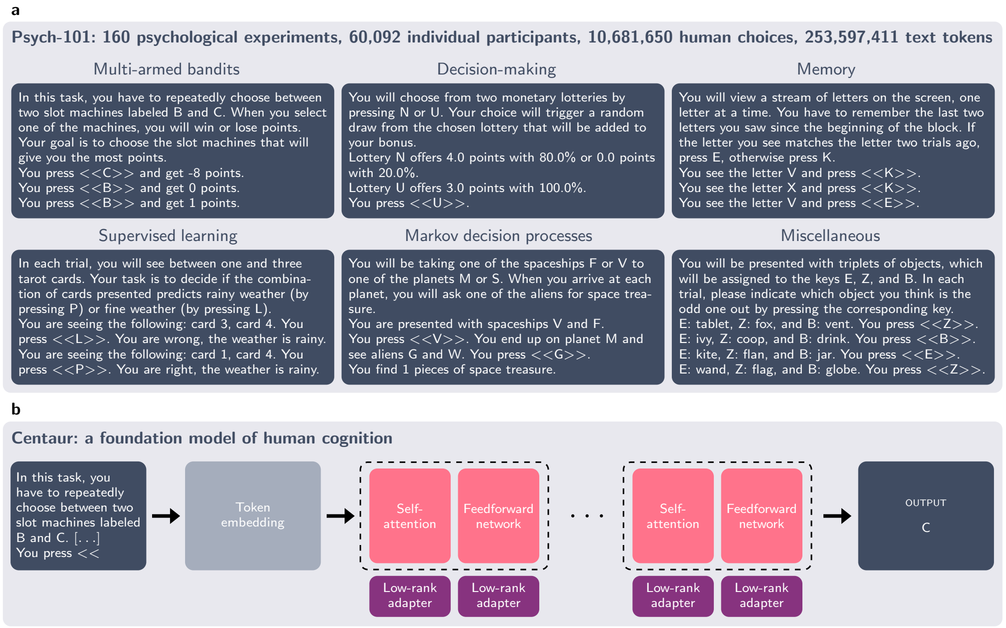
Keywords: cognitive science, cognitive modeling, unified theory of cognition, large language models

Introduction

The human mind is remarkably general [3–5]. Not only do we routinely make mundane decisions, like choosing a breakfast cereal or selecting an outfit, but we also tackle complex challenges, such as figuring out how to cure cancer or explore outer space. We learn new skills from only a few demonstrations [6], reason causally [7], and fuel our actions through curiosity [8]. Whether we are climbing mountains, playing video games, or creating captivating art, our versatility defines what it means to be human.

In contrast to this, most contemporary computational models – whether in machine learning or the cognitive sciences – are domain-specific. They are designed to excel at one particular problem and that problem alone. Take, for instance, AlphaGo – a computer system created by Google DeepMind to master the game of Go [9]. Even though the system can play this particular game at an impressive level, it can do not much beyond that. A similar pattern emerges in the cognitive sciences. Prospect theory, one of the most influential accounts of human cognition, for instance, offers valuable insights into how people make choices [10], but it tells us nothing about how we learn, plan, or explore.

If we want to understand the human mind in its entirety, we must move from domain-specific to domain-general accounts. The importance of such a unified approach has already been recognized by the pioneers of our field. For example, in 1990, Newell stated that "unified theories of cognition are the only way to bring [our] wonderful, increasing fund of knowledge under intellectual control" [2]. How can we make meaningful progress toward such theories?



In this task, you have to repeatedly choose between two slot machines labeled 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.

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>>.

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>>.

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 wrong, the weather is rainy.
You are seeing the following: card 1, card 4. You press <<P>>. You are right, the weather is rainy.

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 pieces of space treasure.

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: globe. You press <<Z>>.

In this task, you have to repeatedly choose between two slot machines labeled B and C. [. . .]
You press <<

Fig. 1 Psych-101 and Centaur overview. **a**, Psych-101 comprises of trial-by-trial data from 160 psychological experiments and 60,092 participants, making 10,681,650 choices in total. It contains domains such as multi-armed bandits, decision-making, memory, supervised learning, Markov decision processes, and others (shown examples are stylized and abbreviated for readability). **b**, Centaur is a foundation of model human cognition that is obtained by adding low-rank adapters to a state-of-the-art language model and finetuning it on Psych-101.

An important step towards a unified theory of cognition is to build a computational model that can predict and simulate human behavior in any domain [2, 11]. The present paper takes up this challenge and introduces Centaur – the first foundation model of human cognition [12]. Centaur was designed in a data-driven manner by finetuning a state-of-the-art large language model [13] on a large corpus of human behavior. For this purpose, we curated a novel, large-scale data set called Psych-101, covering trial-by-trial data from 160 psychological experiments. We transcribed each of these experiments into natural language, which provides a common format for expressing vastly different experimental paradigms [14, 15]. The resulting data set reaches an unprecedented scale, containing over 10,000,000 human choices and including many canonical studies from domains such as multi-armed bandits, decision-making, memory, supervised learning, Markov decision processes, and others (see Figure 1a for an overview and examples).

We subject Centaur to a series of rigorous tests and demonstrate that it captures human behavior at several levels of generalization. First, we show that Centaur predicts behavior of held-out participants (i.e., participants that are not part of the training data) better than existing cognitive models in almost every single experiment. We then demonstrate that its ability to capture human behavior also generalizes to held-out experiments. In this context, we find that Centaur accurately predicts human behavior under modified cover stories, problem structures, and even in entirely novel domains. Finally, we show that Centaur's internal representations become more human-aligned, even though it was never explicitly trained to capture human neural activity.

Taken together, these results demonstrate that it is possible to discover domain-general models of human cognition in a data-driven manner. We believe that Centaur is the first real candidate for a unified model of human cognition and that it offers many opportunities to obtain a better understanding of the human mind.

Results

Model overview

We built Centaur on top of the open-source language model Llama 3.1 70B – a state-of-the-art model pre-trained by Meta AI [13] (hereafter, we refer to this model simply as Llama). Having a large language model as the backbone allowed us to rely on the vast amounts of knowledge that is present in these models [16]. The training process involved finetuning on Psych-101 using a parameter-efficient finetuning technique known as quantized low-rank adaptation (QLoRA) [17]. QLoRA leaves the parameters of the base model intact while adding so-called low-rank adapters, which contain only a few additional, trainable parameters. In our case, we added low-rank adapters of rank eight to all non-embedding layers as illustrated in Figure 1b. With these settings, the newly added parameters amount to 0.15% of the base model's parameters. We then trained the model for one epoch on the entire data set using a standard cross-entropy loss. We masked out the loss for all tokens that do not correspond to human responses, thereby ensuring that the model focuses on capturing human behavior and not on completing experimental instructions. The entire training process took approximately five days on an A100 80GB GPU. Further details on the finetuning procedure are provided in the Methods section.

Centaur predicts human behavior better than domain-specific cognitive models

We evaluated Centaur on different types of held-out data to demonstrate that it robustly captures human behavior. In our first analysis, we tested whether it can predict behavior of participants that were not part of the training data. For this, we split each transcribed experiment into two parts, and used 90% of participants for training and retained 10% for testing. We measured goodness-of-fit to human choices using a pseudo-R^2 measure, which normalizes the log-likelihood of a model by that of a randomly guessing model [18]. In this measure, a value of zero corresponds to prediction at chance level while a value of one indicates perfect predictability.[1] Figure 2a presents the result of this analysis, comparing Centaur against the base model without finetuning and collection of domain-specific models that represent the state-of-the-art in the cognitive science literature. We observed that all models predict human behavior above chance level in the majority of experiments. While there was substantial variance

[1] Note that the (unknown) noise ceiling is in general lower than one, which can only be attained when predicting deterministic behavior.

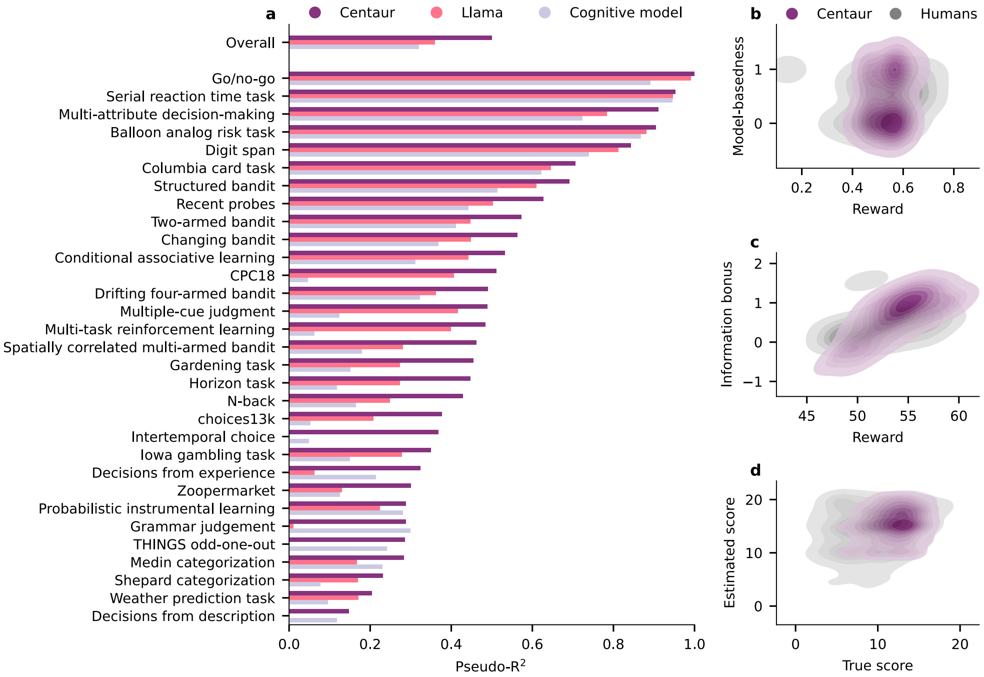


Fig. 2 Performance on Psych-101. **a**, Pseudo-R^2 values for different models across experiments. A value of zero corresponds to prediction at chance level while a value of one corresponds to perfect predictability of human responses. Missing bars indicate performance below chance level. Centaur outperforms both Llama and a collection of domain-specific cognitive models in almost every experiment. Note that we only included experiments for which we have implemented a domain-specific cognitive model in this graphic and merged different studies using the same paradigm. A full table for all experiments can be found in the Supplementary Information. **b**, Model simulations on the two-step task. The plot visualizes probability densities over reward and a parameter indicating how model-based learning was for people and simulated runs of Centaur. **c**, Model simulations on the horizon task. The plot visualizes probability densities over reward and an information bonus parameter for both people and simulated runs of Centaur. **d**, Model simulations on a grammar judgement task. The plot visualizes probability densities over true and estimated scores (i.e., number of correct responses out of twenty) for both people and simulated runs of Centaur.

in predictability between the experiments, finetuning always improved goodness-of-fit. The average improvement across experiments after finetuning was 0.14 (Centaur pseudo-R^2 = 0.50; Llama pseudo-R^2 = 0.36).

Furthermore, we compared Centaur against the previously mentioned collection of domain-specific cognitive models. These models include, amongst others, the generalized context model [19], a prospect theory model [20], and various reinforcement learning models [21, 22]. Further technical details about the modeling can be found in the Supplementary Information. We observed that Centaur outperforms domain-specific cognitive models in all but one experiment. The average improvement in predicting human behavior over the domain-specific cognitive models was 0.18 (Centaur pseudo-R^2 = 0.50; cognitive models pseudo-R^2 = 0.32).

Next, we investigated which factors of an experiment determine whether Centaur captures human behavior. For this, we conducted a regression analysis using the difference in pseudo-R^2 values between Centaur and Llama as a target variable and the number of participants, the number of choices, the number of text characters, and the experiment domain as regressors. We found positive effects for all domains, indicating that finetuning was beneficial for every type of experiment (see Supplementary Information for detailed results). Furthermore, while we did find a positive effect for the number of participants ($\beta = 2.42 \times 10^{-5}$, $p = 0.003$), the number of choices and text characters did not contribute significantly to the improvement in goodness-of-fit. This suggests that having a larger pool of participants is more important for acquiring a good model than the number of data points per participant.

The previous analyses have focused on predicting human responses conditioned on previously executed behavior. We may ask whether Centaur can also generate human-like behavior when simulated in an open-loop fashion (i.e., when feeding its own responses back into the model). This setting arguably provides a much stronger test for the model's capabilities [23]. To check whether Centaur survives this test, we ran open-loop simulations in three different experimental paradigms and inspected the distributions of statistics that resulted from these simulations. The corresponding results can be found in Figure 2b-d. We found that Centaur performs at human-level in all of these simulations, confirming that it can generate meaningful open-loop behavior. Furthermore, Centaur's distributions are well-aligned with the human population, demonstrating that Centaur does not merely model the behavior of the average participant but rather the distribution over trajectories produced by the entire population. For example, in the two-step task – a well-known paradigm to tease apart model-free and model-based reinforcement learning [21] – Centaur produced trajectories in which learning is purely model-free, purely model-based, and mixtures thereof (see Figure 2b).

Probing increasingly complex generalization abilities

Thus far, we have shown that Centaur generalizes to previously unseen participants performing experiments that were part of the training data. A true foundation model of human cognition, however, must also capture behavior in any arbitrary experiment, even if that experiment was not part of the training data. To probe whether Centaur has this ability, we exposed it to a series of increasingly complex out-of-distribution evaluations.

First, we investigated whether Centaur is robust in the face of changes to the cover story. For this analysis, we relied on data collected by Feher da Silva and Hare [24], who conducted a study using the aforementioned two-step task. In addition to the canonical cover story (spaceships traveling to foreign planets in search of treasures), their study introduced a novel cover story involving magical carpets. Importantly, Psych-101 includes experiments using the canonical spaceship cover story [27, 28] but no experiments with the magical carpet cover story. Yet, we still found that Centaur captures human behavior in the magical carpet experiment of Feher da Silva and Hare (see Figure 3a). Like in our previous analysis, we observed an improvement after finetuning, as well as a favorable goodness-of-fit when compared to a domain-specific

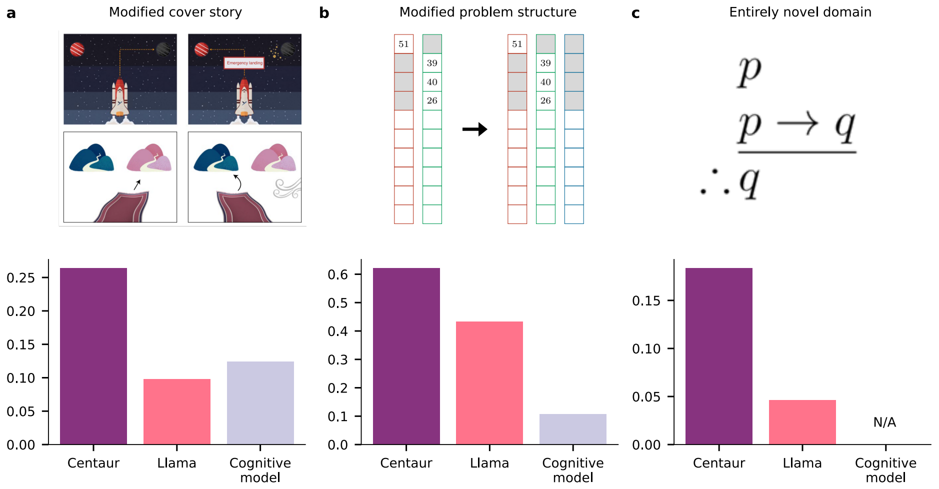


Fig. 3 Evaluation in different held-out settings. **a**, Pseudo-R^2 values for the two-step task with a modified cover story [24]. **b**, Pseudo-R^2 values for a three-armed bandit experiment [25]. **c**, Pseudo-R^2 values for an experiment probing logical reasoning [26]. Centaur outperforms both Llama and domain-specific cognitive models when faced with modified cover stories, problem structures, and entirely novel domains.

cognitive model (Centaur pseudo-R^2: 0.26; Llama pseudo-R^2: 0.10; cognitive model pseudo-R^2: 0.12).

In a second out-of-distribution evaluation, we probed whether Centaur is robust to modifications in task structure. To test this, we exposed it to a paradigm known as Maggie's farm [25]. Maggie's farm extends the horizon task paradigm – a two-armed bandit task used to detect different types of exploration strategies [22] – by adding a third choice option. Psych-101 encompasses several two-armed bandit experiments (including the horizon task) but not Maggie's farm or any other three-armed bandit experiments.[2] Thus, this analysis provides a test of Centaur's robustness to structural task modifications. We found that Centaur captures human behavior on Maggie's farm as shown in Figure 3b. We again observed a benefit of finetuning, as well as a favorable goodness-of-fit compared to a domain-specific cognitive model, which did not generalize well to this setting (Centaur pseudo-R^2: 0.62; Llama pseudo-R^2: 0.43; cognitive model pseudo-R^2: 0.11).

Finally, we investigated whether Centaur can capture human behavior even in entirely novel domains. In this context, we considered a study investigating logical reasoning [26]. While Psych-101 includes probabilistic and causal reasoning problems, we purposefully excluded any studies involving logical reasoning. Like in the previous analyses, there was again a positive effect of finetuning (Centaur pseudo-R^2: 0.18; Llama pseudo-R^2: 0.05; see Figure 3c). Note that we did not compare to any domain-specific cognitive model in this setting, as it is unclear how to construct a model that would make any meaningful transfer from training data that does not include any related problems.

[2] It does, however, contain multi-armed bandit experiments with more than three choice options.

References

[1] Anderson, J. The Architecture of Cognition (1983).

[2] Newell, A. *Unified Theories of Cognition* (Harvard University Press, Cambridge, 1990).

[3] Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences* **40**, e253 (2017).

[4] Lake, B. M. & Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature* **623**, 115–121 (2023).

[5] Wu, C. M., Meder, B. & Schulz, E. Unifying principles of generalization: past, present, and future. *Annu. Rev. Psych* **76**, 1–33 (2024).

[6] Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).

[7] Goddu, M. K. & Gopnik, A. The development of human causal learning and reasoning. *Nature Reviews Psychology* 1–21 (2024).

[8] Chu, J. & Schulz, L. E. Play, curiosity, and cognition. *Annual Review of Developmental Psychology* **2**, 317–343 (2020).

[9] Silver, D. *et al.* Mastering the game of go without human knowledge. *nature* **550**, 354–359 (2017).

[10] Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk (2013).

[11] Riveland, R. & Pouget, A. Natural language instructions induce compositional generalization in networks of neurons. *Nature Neuroscience* **27**, 988–999 (2024).

[12] Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[13] Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[14] Binz, M. & Schulz, E. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences* **120**, e2218523120 (2023).

[15] Binz, M. & Schulz, E. Turning large language models into cognitive models (2024).