Turning large language models into cognitive models

Marcel Binz
MPRG Computational Principles of Intelligence
Max Planck Institute for Biological Cybernetics, Tübingen, Germany
marcel.binz@tue.mpg.de

Eric Schulz
MPRG Computational Principles of Intelligence
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Abstract

Large language models are powerful systems that excel at many tasks, ranging from translation to mathematical reasoning. Yet, at the same time, these models often show unhuman-like characteristics. In the present paper, we address this gap and ask whether large language models can be turned into cognitive models. We find that – after finetuning them on data from psychological experiments – these models offer accurate representations of human behavior, even outperforming traditional cognitive models in two decision-making domains. In addition, we show that their representations contain the information necessary to model behavior on the level of individual subjects. Finally, we demonstrate that finetuning on multiple tasks enables large language models to predict human behavior in a previously unseen task. Taken together, these results suggest that large, pre-trained models can be adapted to become generalist cognitive models, thereby opening up new research directions that could transform cognitive psychology and the behavioral sciences as a whole.

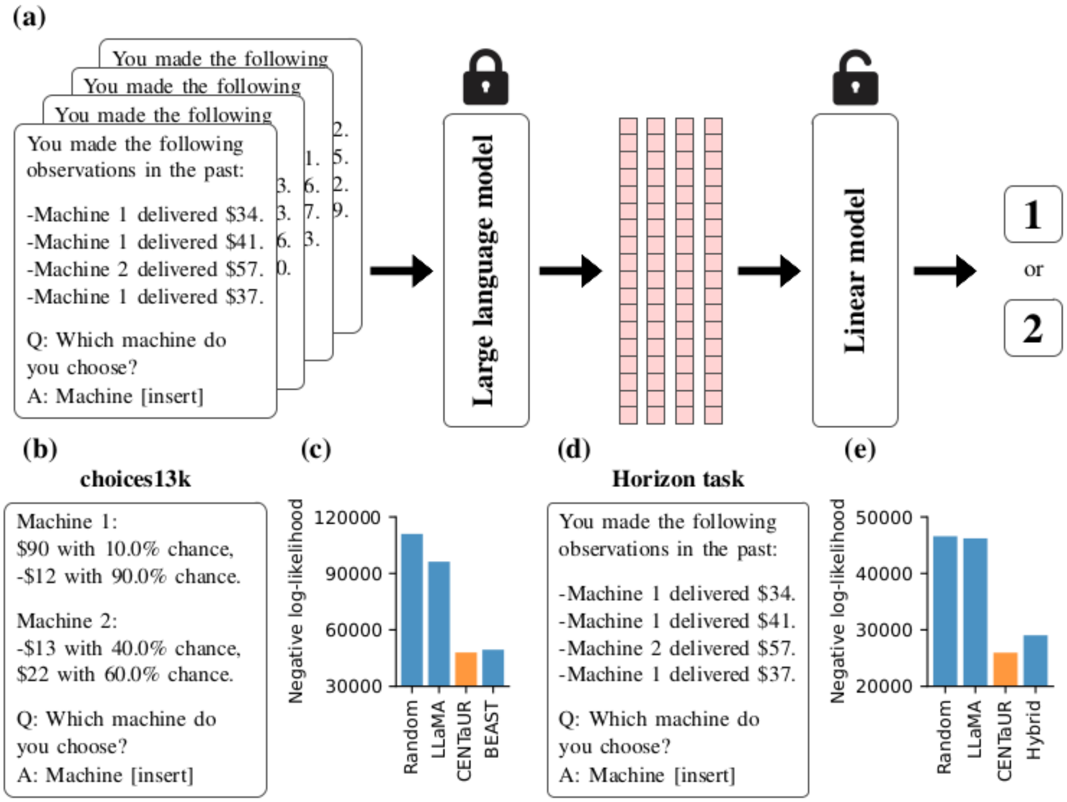
1 Introduction

Large language models are neural networks trained on vast amounts of data to predict the next token for a given text sequence [Brown et al., 2020]. These models display many emergent abilities that were not anticipated by extrapolating the performance of smaller models [Wei et al., 2022]. Their abilities are so impressive and far-reaching that some have argued that they show sparks of general intelligence [Bubeck et al., 2023]. We may currently witness one of the biggest revolutions in artificial intelligence, but the impact of modern language models is felt far beyond, permeating into education [Kasneci et al., 2023], medicine [Li et al., 2023], and the labor market [Eloundou et al., 2023].

In-context learning – the ability to extract information from a context and to use that information to improve the production of subsequent outputs – is one of the defining features of such models. It is through this mechanism that large language models are able to solve a variety of tasks, ranging from translation [Brown et al., 2020] to analogical reasoning [Webb et al., 2022]. Previous work has shown that these models can even successfully navigate when they are placed into classic psychological experiments [Binz and Schulz, 2023, Coda-Forno et al., 2023, Dasgupta et al., 2022, Hagendorff et al., 2022]. To provide just one example, GPT-3 – an autoregressive language model designed by OpenAI [Brown et al., 2020] – outperformed human subjects in a sequential decision-making task that required to balance between exploitative and exploratory actions [Binz and Schulz, 2023].

Even though these models show human-like behavioral characteristics in some situations, this is not always the case. In the sequential decision-making task mentioned above, for instance, GPT-3

Preprint. Under review.



You made the following observations in the past:

-Machine 1 delivered $34.
-Machine 1 delivered $41.
-Machine 2 delivered $57.
-Machine 1 delivered $37.

Q: Which machine do you choose?
A: Machine [insert]

Machine 1:
$90 with 10.0% chance,
-$12 with 90.0% chance.

Machine 2:
-$13 with 40.0% chance,
$22 with 60.0% chance.

Q: Which machine do you choose?
A: Machine [insert]

You made the following observations in the past:

-Machine 1 delivered $34.
-Machine 1 delivered $41.
-Machine 2 delivered $57.
-Machine 1 delivered $37.

Q: Which machine do you choose?
A: Machine [insert]

Figure 1: Illustration of our approach and main results. (a) We provided text-based descriptions of psychological experiments to a large language model and extracted the resulting embeddings. We then finetuned a linear layer on top of these embeddings to predict human choices. We refer to the resulting model as CENTaUR. (b) Example prompt for the choices13k data set. (c) Negative log-likelihoods for the choices13k data set. (d) Example prompt for the horizon task. (e) Negative log-likelihoods for the horizon task. Prompts shown in this figure are stylized for readability. Exact prompts can be found in the Supplementary Materials.

relied heavily on exploitative strategies, while people applied a combination of elaborate exploration strategies [Wilson et al., 2014]. Moreover, GPT-3 stopped improving after only a few trials, while people continued learning as the task progressed.

In the present paper, we investigate whether it is possible to fix the behavioral discrepancy between large language models and humans. To do so, we rely on the idea of finetuning on domain-specific data. This approach has been fruitful across a number of areas [Sanh et al., 2019, Ouyang et al., 2022] and eventually led to the creation of the term *foundation models* [Bommasani et al., 2021] – models trained on broad data at scale and adapted to a wide range of downstream tasks. In the context of human cognition, such domain-specific data can be readily accessed by tapping the vast amount of behavioral studies that psychologists have conducted over the last century. We made use of this and extracted data sets for several behavioral paradigms which we then used to finetune a large language model.

We show that this approach can be used to create models that describe human behavior better than traditional cognitive models. We verify this result through extensive model simulations, which confirm that finetuned language models indeed show human-like behavioral characteristics. Furthermore, we find that the embeddings obtained from such models contain the information necessary to capture individual differences. Finally, we highlight that a model finetuned on two tasks is capable of predicting human behavior on a third, hold-out task. Taken together, our work demonstrates that it is possible to turn large language models into cognitive models, thereby opening up completely new opportunities to harvest the power of large language models for building domain-general models of human learning and decision-making.

2 Finetuned language models beat domain-specific models

We started our investigations by testing whether it is possible to capture how people make decisions through finetuning a large language model. For our analyses, we relied on the *Large Language Model Meta AI*, or in short: LLaMA [Touvron et al., 2023]. LLaMA is a family of state-of-the-art foundational large language models (with either 7B, 13B, 33B, or 65B parameters) that were trained on trillions of tokens coming from exclusively publicly available data sets. We focused on the largest of these models – the 65B parameter version – for the analyses in the main text. LLaMA is publicly available, meaning that researchers are provided with complete access to the network architecture including its pre-trained weights. We utilized this feature to extract embeddings for several cognitive tasks and then finetuned a linear layer on top of these embeddings to predict human choices (see Figure 1a for a visualization). We call the resulting class of models CENTaUR, in analogy to the mythical creature that is half human and half ungulate.

We considered two paradigms that have been extensively studied in the human decision-making literature for our initial analyses: *decisions from descriptions* [Kahneman and Tversky, 1972] and *decisions from experience* [Hertwig et al., 2004]. In the former, a decision-maker is asked to choose between one of two hypothetical gambles like the ones shown in Figure 1b. Thus, for both options, there is complete information about outcome probabilities and their respective values. In contrast, the decisions from experience paradigm does not provide such explicit information. Instead, the decision-maker has to learn about outcome probabilities and their respective values from repeated interactions with the task as shown in Figure 1d. Importantly, this modification calls for a change in how an ideal decision-maker should approach such problems: it is not enough to merely exploit currently available knowledge anymore but also crucial to explore options that are unfamiliar [Schulz and Gershman, 2019].

For both these paradigms, we created a data set consisting of embeddings and the corresponding human choices. We obtained embeddings by passing prompts that included all the information that people had access to on a given trial through LLaMA and then extracting the hidden activations of the final layer (see Figure 1b and d for example prompts, and the Supplementary Materials for a more detailed description about the prompt generation procedure). We relied on publicly available data from earlier studies in this process. In the decisions from descriptions setting, we used the choices13k data set [Peterson et al., 2021], which is a large-scale data set consisting of over 13,000 choice problems (all in all, 14,711 participants made over one million choices on these problems). In the decisions from experience setting, we used data from the horizon task [Wilson et al., 2014] and a replication study [Feng et al., 2021], which combined include 60 participants making a total of 67,200 choices.

With these two data sets at hand, we fitted a regularized logistic regression model from the extracted embeddings to human choices. In this section, we restricted ourselves to a joint model for all participants, thereby neglecting potential individual differences (but see one of the following sections for an analysis that allows for individual differences). Model performance was measured through the predictive log-likelihood on hold-out data obtained using a 100-fold cross-validation procedure. We standardized all input features and furthermore applied a nested cross-validation for tuning the hyperparameter that controls the regularization strength. Further details are provided in the Materials and Methods section.

We compared the goodness-of-fit of the resulting models against three baselines: a random guessing model, LLaMA without finetuning (obtained by reading out log-probabilities of the pre-trained model), and a domain-specific model (*Best Estimate and Sampling Tools*, or BEAST, for the choices13k data set [Erev et al., 2017] and a *hybrid model* [Gershman, 2018] that involves a combination of different exploitation and exploration strategies for the horizon task). We found that LLaMA did not capture human behavior well, obtaining a negative log-likelihood (NLL) close to chance-level for the choices13k data set (NLL = 96248.5) and the horizon task (NLL = 46211.4). However, finetuning led to models that captured human behavior better than the domain-specific models under consideration. In the choices13k data set, CENTaUR achieved a negative log-likelihood of 48002.3 while BEAST only achieved a negative log-likelihood of 49448.1 (see Figure 1c). In the horizon task, CENTaUR achieved a negative log-likelihood of 25968.6 while the hybrid model only achieved a negative log-likelihood of 29042.5 (see Figure 1e). Together, these results suggest that the representations extracted from large language models are rich enough to attain state-of-the-art results for modeling human decision-making.

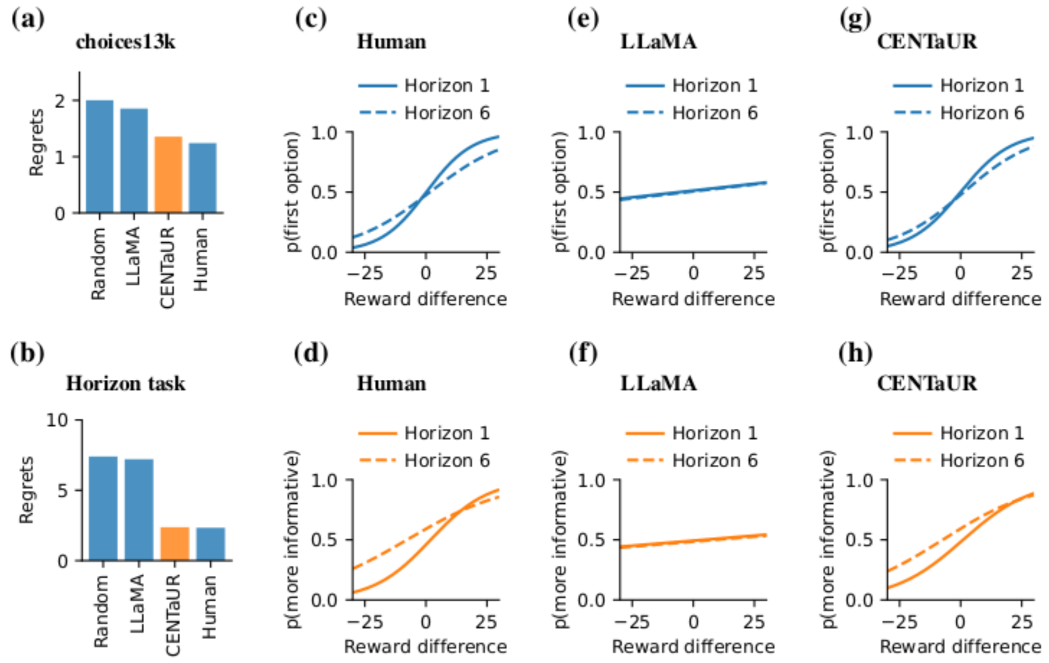


Figure 2: Model simulations. (a) Performance for different models and human participants on the choices13k data set. (b) Performance for different models and human participants on the horizon task. (c) Human choice curves in the equal information condition of the horizon task. (d) Human choice curves in the unequal information condition of the horizon task. (e) LLaMA choice curves in the equal information condition of the horizon task. (f) LLaMA choice curves in the unequal information condition of the horizon task. (g) CENTaUR choice curves in the equal information condition of the horizon task. (h) CENTaUR choice curves in the unequal information condition of the horizon task.

3 Model simulations reveal human-like behavior

We next verified that CENTaUR shows human-like behavioral characteristics. To do so, we simulated the model on the experimental data. Looking at performance, we found that finetuning led to models that closely resemble human performance as shown in Figure 2a and b. For the choices-13k data set, CENTaUR obtained a regret (defined as the difference between the highest possible reward and the reward for the action selected by the model) of 1.35 (SE $= 0.01$), which was much closer to the human regret (M = 1.24, SE = 0.01) than the regret of LLaMA (M = 1.85, SE = 0.01). The results for the horizon task showed an identical pattern, with CENTaUR (M = 2.38, SE = 0.01) matching human regret (M = 2.33, SE = 0.05) more closely than LLaMA (M = 7.21, SE = 0.02).

In addition to looking at performance, we also inspected choice curves. For this analysis, we took the data from the first free-choice trial in the horizon task and divided it into two conditions: (1) an equal information condition that includes trials where the decision-maker had access to an equal number of observations for both options and (2) an unequal information condition that includes trials where the decision-maker previously observed one option fewer times than the other. We then fitted a separate logistic regression model for each condition with reward difference, horizon, and their interaction as independent variables onto the simulated choices. Earlier studies with human subjects [Wilson et al., 2014] identified the following two main results regarding their exploratory behavior: (1) people's choices become more random with a longer horizon in the equal information condition (as shown in Figure 2c) and (2) people in the unequal information condition select the more informative option more frequently when the task horizon is longer (as shown in Figure 2d). While LLaMA did not show any of the two effects (see Figure 2e and f), CENTaUR exhibited both of them (see Figure 2g and h), thereby further corroborating that it accurately captures human behavior.

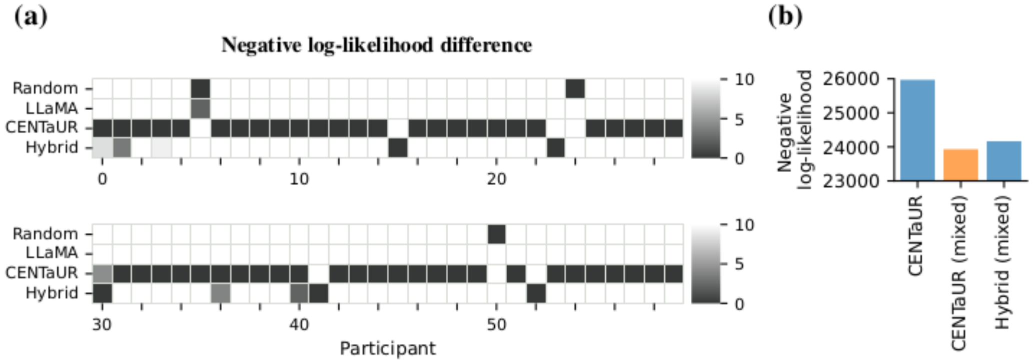


Figure 3: Individual differences. (a) Negative log-likelihood difference to the best-fitting model for each participant. Black highlights the best-fitting model, while white corresponds to a difference larger than ten. (b) Negative log-likelihoods for models that were finetuned using the random-effects structure described in the main text.

4 Language model embeddings capture individual differences

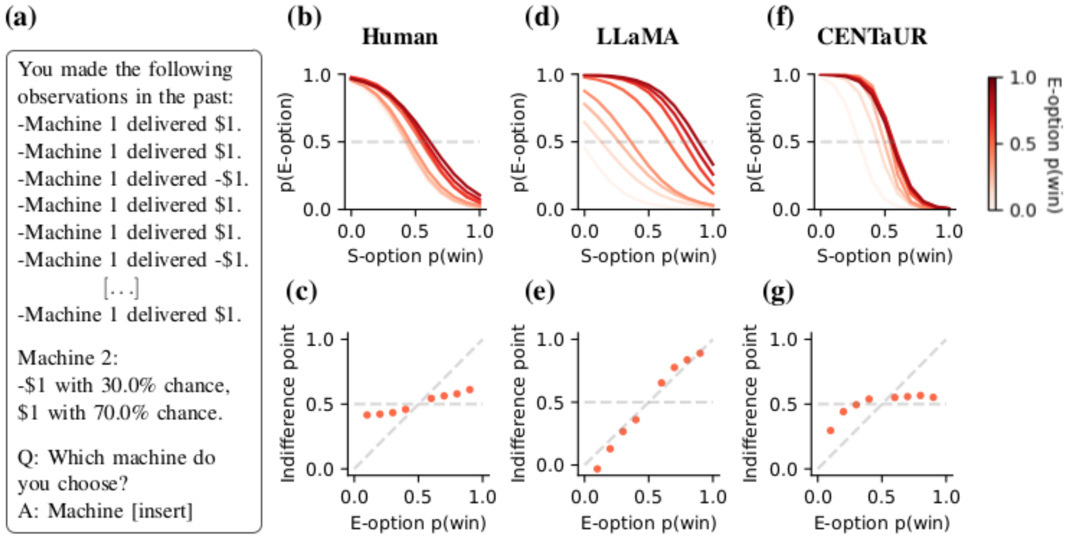
We also investigated how well CENTaUR describes the behavior of each individual participant. Note that this form of analysis is only possible for the horizon task as choice information on the participant level is not available for the choices13k data set. In total, the majority of participants (N = 52 out of 60) was best modeled by CENTaUR (see Figure 3a for a detailed visualization). We furthermore entered the negative log-likelihoods into a random-effects model selection procedure which estimates the probability that a particular model is the most frequent explanation within a set of candidate models [Rigoux et al., 2014]. This procedure favored CENTaUR decisively, assigning a probability that it is the most frequent explanation of close to one.

Thus far, we have finetuned LLaMA jointly for all participants. However, people may exhibit individual differences that are not captured by this analysis. To close this gap and test whether LLaMA embeddings can account for individual differences, we incorporated random effects in the finetuned layer. We added a random effect for each participant and embedding dimension while keeping the remaining evaluation procedure the same. Figure 3b illustrates the resulting negative log-likelihoods. Including the random-effect structure improved goodness-of-fit considerably (NLL = 23929.5) compared to the fixed-effect-only model (NLL = 25968.6). Furthermore, CENTaUR remained superior to the hybrid model with an identical random-effect structure (NLL = 24166.0). Taken together, the findings reported in this section highlight that embeddings of large language models contain the information necessary to model behavior on the participant level.

5 Evaluating goodness-of-fit on hold-out tasks

Finally, we examined whether CENTaUR – after being finetuned on multiple tasks – is able to predict human behavior in an entirely different task. This evaluation protocol provides a much stronger test for the generalization abilities of our approach. Following our initial analyses, we finetuned a linear layer on top of LLaMA embeddings. However, this time, we fitted a joint model using both the data from the choices13k data set and the horizon task, and then evaluated how well the finetuned model captures human choices on a third task. Further details about the fitting procedure are provided in the Materials and Methods section. For the hold-out task, we considered data from a recent study that provided participants with a choice between one option whose information is provided via a description and another option for which information is provided via a list of experienced outcomes [Garcia et al., 2023]. Figure 4a shows an example prompt for this experimental paradigm.

Finetuning was generally beneficial for modeling human behavior on the hold-out task: negative log-likelihoods for CENTaUR (NLL = 4521.1) decreased both in comparison to a random guessing model (NLL = 5977.7) and LLaMA (NLL = 6307.9). We were thus curious whether CENTaUR also captures human behavior on a qualitative level. To test this, we took a look at the key insight from the original study: people tend to overvalue options that are provided through a description



You made the following observations in the past:
-Machine 1 delivered $1.
-Machine 1 delivered $1.
-Machine 1 delivered -$1.
-Machine 1 delivered $1.
-Machine 1 delivered $1.
-Machine 1 delivered -$1.
[. . .]
-Machine 1 delivered $1.

Machine 2:
-$1 with 30.0% chance,
$1 with 70.0% chance.

Q: Which machine do you choose?
A: Machine [insert]

Figure 4: Hold-out task evaluations. (a) Example prompt for the experiential-symbolic task of Garcia et al. [2023]. (b) Human choice curves as a function of win probabilities for both options. (c) Human indifference points as a function of win probability for the E-option. Indifferent points express the win probabilities at which a decision-maker is equally likely to select both options. (d) LLaMA choice curves as a function of win probabilities for both options. (e) LLaMA indifference points as a function of win probability for the E-option. (f) CENTaUR choice curves as a function of win probabilities for both options. (g) CENTaUR indifference points as a function of win probability for the E-option.

(symbolic or S-options) over the options that come with a list of experienced outcomes (experiential or E-options) as illustrated in Figure 4b and c. LLaMA does not show this characteristic and instead weighs both option types equally (Figure 4d and e). In contrast to this, CENTaUR shows human-like behavior, taking mostly the S-option into account (Figure 4f and g). This is remarkable because we never presented data from the experiment under consideration during finetuning.

6 Discussion

We have demonstrated that large language models can be turned into cognitive models by finetuning their final layer. This process led to models that achieved state-of-the-art performance in two domains. Furthermore, these models were able to capture behavioral differences at the individual participant level. Finally, we have shown that our approach generalizes to previously unseen tasks. In particular, a model that was finetuned on two tasks also exhibited human-like behavior on a third, hold-out task.

These results complement earlier work showing that large language model embeddings allow us to predict behavior and neural activations in linguistic settings [Schrimpf et al., 2021, Kumar et al., 2022, Tuckute et al., 2023, Antonello et al., 2023]. For example, Schrimpf et al. [2021] showed that large language models can predict neural and behavioral responses in tasks that involved reading short passages with an accuracy that was close to noise ceiling. While it may be expected that large language models explain human behavior in linguistic domains (after all these models are trained to predict future word occurrences), the observation that these results also transfer to more cognition domains like the ones studied here is highly non-trivial.

We are particularly excited about one feature of CENTaUR: embeddings extracted for different tasks all lie in a common space. This property allows finetuned large language models to solve multiple tasks in a unified architecture. We have presented preliminary results in this direction, showing that a model finetuned on two tasks can predict human behavior on a third. However, we believe that our current results only hint at the potential of this approach. Ideally, we would like to scale up our approach to finetuning on a larger number of tasks from the psychology literature. If one would include enough tasks in the training set, the resulting system should – in principle – generalize to *any* hold-out task. Therefore, our approach provides a path towards a domain-general model of human cognition, which has been the goal of theoreticians for decades [Newell, 1992, Yang et al., 2019, Riveland and Pouget, 2022, Binz et al., 2023]. We believe that having access to such a model would transform psychology and the behavioral sciences more generally. It could, among other applications, be used to rapidly prototype the outcomes of projected experiments, thereby easing the trial-and-error process of experimental design, or to provide behavioral policy recommendations while avoiding expensive data collection procedures.

Finally, we have to ask ourselves what we can learn about human cognition when finetuning large language models. For now, our insights are limited to the observation that large language model embeddings are rich enough to explain human decision-making. While this is interesting in its own right, it is certainly not the end of the story. Looking beyond the current work, having access to an accurate neural network model of human behavior provides the opportunity to apply a wide range of explainability techniques from the machine learning literature. For instance, we could pick a particular neuron in the embedding and trace back what parts of a particular input sequence excite that neuron using methods such as layer-wise relevance propagation [Bach et al., 2015, Chefer et al., 2021]. Thus, our work also opens up a new spectrum of analyses that are not possible when working with human subjects.

To summarize, large language models are an immensely powerful tool for studying human behavior. We believe that our work has only scratched the surface of this potential and there is certainly much more to come.

Acknowledgements: We like to thank Robert Wilson and Basile Garcia for their help on the horizon task and the experiential-symbolic task respectively, Ido Erev and Eyal Ert for their help with the BEAST model, and Meta AI for making LLaMA accessible to the research community.

Funding: This work was funded by the Max Planck Society, the Volkswagen Foundation, as well as the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.

Data and materials availability: Data and code for the current study are available through the GitHub repository `https://github.com/marcelbinz/CENTaUR`.

References

Richard Antonello, Aditya Vaidya, and Alexander G Huth. Scaling laws for language encoding models in fmri. *arXiv preprint arXiv:2305.11863*, 2023.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Marcel Binz, Ishita Dasgupta, Akshay Jagadish, Matthew Botvinick, Jane X Wang, and Eric Schulz. Meta-learned models of cognition. *arXiv preprint arXiv:2304.06729*, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. Inducing anxiety in large language models increases exploration and bias. *arXiv:2304.11111*, 2023.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

Ido Erev, Eyal Ert, Ori Plonsky, Doron Cohen, and Oded Cohen. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, 124(4):369, 2017.

Samuel F Feng, Siyu Wang, Sylvia Zarnescu, and Robert C Wilson. The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. *Scientific reports*, 11(1): 1–15, 2021.

Basile Garcia, Maël Lebreton, Sacha Bourgeois-Gironde, and Stefano Palminteri. Experiential values are underweighted in decisions involving symbolic options. *Nature human behaviour*, pages 1–16, 2023.

Samuel J Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Machine intuition: Uncovering human-like intuitive decision-making in gpt-3.5. *arXiv preprint arXiv:2212.05206*, 2022.

Ralph Hertwig, Greg Barron, Elke U Weber, and Ido Erev. Decisions from experience and the effect of rare events in risky choice. *Psychological science*, 15(8):534–539, 2004.

Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *BioRxiv*, pages 2022–06, 2022.

Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5(6):e333–e335, 2023.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Allen Newell. Unified theories of cognition and the role of soar. *Soar: A cognitive architecture in perspective: A tribute to Allen Newell*, pages 25–79, 1992.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.

Ori Plonsky, Reut Apel, Ido Erev, Eyal Ert, and Moshe Tennenholtz. When and how can social scientists add value to data scientists? a choice prediction competition for human decision making. *Unpublished Manuscript*, 2018.

Lionel Rigoux, Klaas Enno Stephan, Karl J Friston, and Jean Daunizeau. Bayesian model selection for group studies—revisited. *Neuroimage*, 84:971–985, 2014.

Reidar Riveland and Alexandre Pouget. A neural model of task compositionality with natural language instructions. *bioRxiv*, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Eric Schulz and Samuel J Gershman. The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, 55:7–14, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *bioRxiv*, 2023.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2212.09196*, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *TMLR*, 2022. ISSN 2835-8856.

Robert C Wilson, Andra Geana, John M White, Elliot A Ludvig, and Jonathan D Cohen. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074, 2014.

Guangyu Robert Yang, Madhura R Joglekar, H Francis Song, William T Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, 22(2):297–306, 2019.

Supplementary Materials

Materials and Methods

Fitting procedure: For the main analyses, we fitted a separate regularized logistic regression model for each data set via a maximum likelihood estimation. Final model performance was measured through the predictive log-likelihood on hold-out data obtained using a 100-fold cross-validation procedure. In each fold, we split the data into a training set (90%), a validation set (9%), and a test set (1%). The validation set was used to identify the parameter α that controls the strength of the ℓ_2-regularization term. We considered discrete α-values of [0, 0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0]. The optimization procedure was implemented in PYTORCH [Paszke et al., 2019] and used the default LBFGS optimizer [Liu and Nocedal, 1989]. For the individual difference analyses, the procedure was identical except that we added a random effect for each participant and embedding dimension.

For the hold-out task analyses, the training set consisted of the concatenated choices13k and horizon task data. To obtain a validation and test set, we split the data of the experiential-symbolic task into eight folds and repeated the previously described fitting procedure for each fold. The validation set was used to identify the parameter α that controls the strength of the ℓ_2-regularization term and an inverse temperature parameter τ^{-1}. We considered discrete inverse temperature values of [0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1] and α-values as described above.

Model simulations: For the main analyses, we simulated model behavior by sampling from the predictions on the test set. For the hold-out task analyses, we simulated data deterministically based on a median threshold (again taking the predictions on the test set). The resulting choice curves were generated by fitting a separate logistic regression model for each possible win probability of the E-option. Each model used the win probability of the S-option and an intercept term as independent variables and the probability of choosing the E-option as the dependent variable.

Baseline models: For the LLaMA baseline, we fitted an inverse temperature parameter to human choices using the procedure described above. For the BEAST baseline, we relied on the version provided for the choice prediction competition 2018 [Plonsky et al., 2018]. We additionally included an error model that selects a random choice with a particular probability. We treated this probability as a free parameter and fitted it using the procedure described above. The hybrid model closely followed the implementation of Gershman [2018]. We replaced the probit link function with a logit link function to ensure comparability to CENTaUR.

Supplementary Text

For the choices13k data set, we prompted each decision independently, thereby ignoring the potential effect of feedback. We used the following template:

> Machine 1 delivers 90 dollars with 10.0% chance and -12 dollars with 90.0% chance.
> Machine 2 delivers -13 dollars with 40.0% chance and 22 dollars with 60.0% chance.
>
> Your goal is to maximize the amount of received dollars.
>
> Q: Which machine do you choose?
> A: Machine [insert]

For the horizon task, we prompted each task independently, thereby ignoring potential learning effects across the experiment. We used the following template:

> You made the following observations in the past:
> - Machine 1 delivered 34 dollars.
> - Machine 1 delivered 41 dollars.

- Machine 2 delivered 57 dollars.
- Machine 1 delivered 37 dollars.

Your goal is to maximize the sum of received dollars within six additional choices.

Q: Which machine do you choose?
A: Machine [insert]

For the experiential-symbolic task, we prompted each decision independently and only considered the post-learning phase. We furthermore simplified the observation history by only including the option that is relevant to the current decision. We used the following template:

You made the following observations in the past:
- Machine 1 delivered 1 dollars.
- Machine 1 delivered 1 dollars.
- Machine 1 delivered -1 dollars.
- Machine 1 delivered 1 dollars.
- Machine 1 delivered 1 dollars.
- Machine 1 delivered -1 dollars.

[. . .]

- Machine 1 delivered 1 dollars.

Machine 2 delivers -1 dollars with 30.0% chance and 1 dollars with 70.0% chance.

Your goal is to maximize the amount of received dollars.

Q: Which machine do you choose?
A: Machine [insert]