From Prediction to Understanding: Will Al Foundation Models Transform Brain Science?

Thomas Serre and Ellie Pavlick

Departments of Cognitive & Psychological Sciences and Computer Science

Carney Center for Computational Brain Science

Brown University

Abstract

Generative pretraining (the "GPT" in ChatGPT) enables language models to learn from vast amounts of internet text without human supervision. This approach has driven breakthroughs across AI by allowing deep neural networks to learn from massive, unstructured datasets. We use the term foundation models to refer to large pretrained systems that can be adapted to a wide range of tasks within and across domains, and these models are increasingly applied beyond language to the brain sciences. These models achieve strong predictive accuracy, raising hopes that they might illuminate computational principles. But predictive success alone does not guarantee scientific understanding.

Here, we outline how foundation models can be productively integrated into the brain sciences, highlighting both their promise and their limitations. The central challenge is to move from prediction to explanation: linking model computations to mechanisms underlying neural activity and cognition.

Main text

Over the past year, two foundation models published back-to-back in Nature have drawn significant attention from brain scientists. The first is a neural foundation model trained on large-scale calcium imaging data from the mouse visual cortex¹. The second, Centaur, is a behavioral foundation model trained to predict human decision-making across hundreds of psychology experiments². Both achieve impressively predictive accuracy—generalizing across experimental tasks, subjects, and stimulus domains. These case studies sharpen a familiar question: what mechanisms, if any, do such predictors capture? We briefly review their training paradigms and then ask what predictive success can—and cannot—reveal about underlying mechanisms.

The Pretrain-Finetune Recipe. At the core of large language model (LLM) capabilities is self-supervised learning (SSL)—a family of approaches in which models learn by predicting missing or masked parts of their input without external labels. The most widely used form in LLMs is generative pretraining³ (the "GPT" in ChatGPT), where models are trained autoregressively to predict the next token in large, unstructured text datasets, thereby acquiring broad linguistic and world knowledge. See Box 1 for additional forms of SSL and architectural details. A subsequent finetuning phase then adapts the model to specific domains and

applications. In many cases, this involves supervised training on labeled datasets—for example, to classify clinical notes⁴, analyze legal documents^{5,6}, or model political decision-making^{7,8}.

For open-ended text generation as in ChatGPT, models are first finetuned on human-written demonstrations—for example, prompts paired with ideal responses that illustrate helpful dialogue—and then further aligned using reinforcement learning. For example, in reinforcement learning from human feedback (RLHF), models optimize their responses to match human preferences rather than merely maximizing the likelihood of training text^{9–12}. This two-step process effectively puts the "Chat" in ChatGPT, enabling models to generate responses that are more closely aligned with human intent. Variants of this RLHF pipeline are now standard across both general-purpose assistants and domain-specific systems, including mental health chatbots¹³.

Box 1 | Self-Supervised Learning

Self-supervised learning (SSL) is a family of methods in which a model creates its own training signal. Rather than relying on external labels, the model withholds or corrupts parts of its input and learns to predict them from the surrounding context. This allows the system to learn directly from large, unlabeled datasets.

A widely used SSL approach is generative pretraining³, introduced in the original GPT paper, where models are trained to predict the next element in an input sequence (e.g., the next word in a sentence). The basic units the model operates on are called tokens, which generalize across data types. In text, tokens may be words, sub-words, or characters; in images, they are patches; in audio or video, short frames or windows; and in neural or behavioral data, they can be time bins or events (e.g., spike counts). This shared tokenization principle allows the same SSL objectives to extend beyond language to images, audio/video, neural recordings, and behavior.

A foundation model refers to a network pretrained at scale (usually with SSL) on broad, heterogeneous data to learn versatile representations. These representations can then be adapted for downstream tasks using simple methods such as linear readouts, prompting, or finetuning.

Today's foundation models are almost all built on the transformer architecture¹⁴, which has largely replaced earlier designs (MLPs, RNNs, CNNs). Transformers use a mechanism called self-attention, where each token's representation is updated based on weighted relations ("attention scores") with other tokens in the input sequence. In decoder-only models (e.g., GPT, Centaur²), attention is restricted to past tokens (causal masking), supporting next-token prediction. In encoder models (e.g., BERT¹⁵ language model, calcium-imaging model¹), attention is bidirectional, supporting masked-token prediction. Despite these differences, the common recipe is to pretrain a transformer on large datasets to learn general-purpose representations, and then adapt it to specific tasks.

For neuroscientists, a useful analogy is the contrast between hand-engineered features (e.g., filtering neural signals by frequency bands, applying PCA/ICA) and learned representations. Traditional workflows depend on manually defined features that are then fed into a decoder (often linear). SSL removes the manual feature-engineering step: the model learns features

automatically from raw or lightly processed data at scale. Downstream analyses then often reduce to training a linear classifier on top of the learned representation.

The remarkable progress of foundation models reflects this shift—from handcrafted features to learned, general-purpose representations—coupled with massive amounts of data and compute. This flexibility enables models to capture complex dependencies in neural, behavioral, and multimodal data, driving strong performance on predictive tasks¹⁶.

SSL extends well beyond language. In vision, models learn by masking or removing image patches and predicting the missing content, enabling applications ranging from object recognition to medical imaging^{17,18}. In speech, related methods reconstruct masked or corrupted audio segments, achieving state-of-the-art recognition with far less transcribed data¹⁹. In discrete domains such as programming, models trained on large code corpora predict missing or next tokens, supporting applications in code generation, translation, and debugging²⁰. Across domains, the recipe is consistent: large-scale pretraining on raw data followed by application-specific finetuning. The same principles extend to multimodal models, which integrate text, images, audio, and even actions into shared representations. We summarize these developments in Box 2.

Box 2 | Foundation Model Scaling

Recent progress in AI has been driven by empirical scaling laws, which show that model performance improves predictably as data, compute, and parameters increase. These laws have guided the design of ever-larger foundation models pretrained on broad datasets and adaptable to diverse downstream tasks.

The same pretraining–finetuning paradigm now enables multimodal learning across text, images, audio, and other data types. For example, Vision-language models^{21,22} (VLMs) trained on large image–text datasets can support clinical clinical tasks, while vision–language–action²³ (VLA) models enable robots to follow natural-language commands.

Frontier models—such as GPT-4o, Claude Sonnet 4, and Gemini 2.0—represent multimodal systems trained at unprecedented scales. Yet while scaling delivers impressive predictive gains, it does not guarantee that models uncover the causal mechanisms underlying the systems they aim to represent.

The same pretrain–finetune recipe is now being applied across the sciences, where foundation models are accelerating discovery in diverse domains. In biology, protein and genomics foundation models can predict 3D protein structures, design enzymes, and suggest functional mutations^{24–26}—capabilities exemplified by AlphaFold²⁵, whose lead developers, Demis Hassabis and John Jumper, shared half of the 2024 Nobel Prize in Chemistry, with the other half awarded to David Baker for computational protein design. In climate science, Al-based weather

models can forecast storms days in advance with accuracy that rivals the best physics-based systems^{27,28}. In materials science, AI-based systems have proposed millions of candidate materials, many of which have already been synthesized and tested²⁹.

Powerful Predictive Models for Neuroscience. Just as AlphaFold transformed drug discovery by overcoming the long-standing bottleneck of protein structure prediction, neural foundation models promise to transform areas of neuroscience where accurate prediction is critical—for example, enabling adaptive deep brain stimulation for Parkinson's disease³⁰, identifying early behavioral or biological markers in depression or obsessive-compulsive disorders^{31,32}, or advancing neural prosthetics to restore walking after spinal cord injury³³.

A recent study introduced a foundation model of the mouse visual cortex trained on large-scale calcium imaging data¹. The model not only predicts responses across new stimulus domains and individual animals with high accuracy, but also captures information about neuronal cell types, dendritic morphology, and connectivity. Similar progress is underway in human neuroimaging, where large fMRI foundation models predict brain states and clinical variables, while generalizing to new cohorts^{34,35}.

In cognitive science, Centaur is the first large-scale behavioral foundation model trained to predict human decision-making across a broad range of tasks and experimental settings². The model encodes both natural-language task descriptions and prior participant choices as tokens, and predicts the participant's next choice from this sequence. Centaur outperforms classical cognitive models on held-out participants and generalizes to previously unseen task variations.

As neuroscience datasets expand, so too does the prospect of foundation model–based digital twins. For instance, high-density probes can now record from thousands of neurons across long timescales^{36,37} and single-cell atlases aggregate tens of millions of cells with spatial context³⁸. In principle, if a generative foundation model could produce neural or behavioral time series that are empirically indistinguishable (under rich tests) from those of an individual or cohort, it would function as a neurally realistic simulator. Such a simulator would form the core of a digital twin, enabling in silico experimentation, personalized medicine, and neurotechnology³⁹. Yet indistinguishability is only a predictive criterion; it does not, by itself, provide a mechanistic account. Clarifying how digital twins might move beyond predictive mimicry toward mechanistic fidelity will determine whether they can advance neuroscience rather than remain another black box.

These examples sharpen the overarching question: do foundation models uncover causal mechanisms, or merely exploit statistical regularities? In AI, the issue is whether models recover any plausible generative process consistent with the data, or simply rely on pattern matching. In neuroscience and cognitive science, the challenge is sharper: do such models capture the specific data-generating process underlying neural activity and cognition, rather than just one of many fits? This distinction between prediction and explanation is long-standing, but the extraordinary predictive power of today's models makes it easy to mistake fit for understanding. In what follows, we review recent work that begins to address this central question in AI.

Prediction is not Explanation. There is growing optimism that, as foundation models improve, they may transition from capturing correlations to uncovering the generative processes underlying their data. If a model speaks like a human, perhaps it has internalized grammar; if it can play chess or Othello from game records alone^{40–42}, perhaps it has inferred the rules of the game. Such hopes are grounded in empirical scaling laws, which show that performance improves predictably with increasing model size, data, and compute.

Indeed, there is evidence that these hopes sometimes bear out. For example, while debates persist about the details, it is generally accepted that language models represent syntactic categories and relations, such as parts of speech and syntactic dependencies⁴³, and that these representations play a measurable causal role in the model behavior⁴⁴. Recent work also suggests that transformer networks may capture learning dynamics that parallel those observed in humans. For instance, one study shows that the interplay between in-context and in-weight learning in neural networks mirrors dual-process theories in cognitive science, reproducing trade-offs between flexibility and retention and between blocked and interleaved curricula⁴⁵. Together, these findings offer cautious optimism that foundation models might not only achieve predictive accuracy but also provide insights into the cognitive and neural mechanisms underlying language and learning.

At the same time, models can achieve strong performance not by capturing mechanisms but by exploiting spurious correlations—a phenomenon known as shortcut learning⁴⁶. But even perfect predictive accuracy does not guarantee explanatory value: Ptolemy's epicycles, for example, offered accurate predictions of planetary motion while reflecting a false theory of celestial mechanics. Predictive abstraction can hold independently of mechanistic truth. It remains debated, for example, whether superhuman-level game-playing models trained purely on move sequences truly encode the game's rules^{40–42,47,48}, or whether models that solve analogies at a human level possess reasoning mechanisms like those of humans⁴⁹. In both cases, studies show that performance degrades in situations not encountered during training.

A recent illustration comes from foundation models trained on synthetic orbital trajectories generated by Newtonian mechanics⁵⁰. Although these models achieved high predictive accuracy, they failed to generalize to related physics tasks—suggesting that they had not internalized the underlying physical laws. This example highlights a key limitation: even when the true generative process is simple, well-defined, and embedded in the training data, a model may still fail to recover it. What evidence is there that current neural and behavioral foundation models capture genuine mechanisms of brain and cognition?

So far, the evidence remains limited. For example, the calcium-imaging foundation model¹ learns weight structures that organize information in ways partially consistent with anatomy—for instance, neuron types, dendritic morphology, and connectivity. Yet these correspondences fall short of revealing the circuit mechanisms by which cortex computes. Likewise, after finetuning, the Centaur behavioral model² shows increased alignment with human fMRI, linking its internal states to brain activity. However, critiques highlight that Centaur's predictions often diverge from human behavior in well-known psychology experiments⁵¹. In some cases, it achieves accurate performance even without access to task information, relying instead on statistical regularities in

choice sequences—strategies fundamentally incompatible with human decision-making⁵². These limitations highlight the need for stronger criteria to assess when alignment between models and human data reflects genuine mechanism rather than surface-level fit.

These findings underscore that predictive alignment, on its own, is insufficient for mechanistic explanation. Without evidence that models have discovered genuine computational mechanisms of brain and cognition, we risk replacing one black box (the brain) with another (a deep neural network) of comparable complexity. Realizing the promise of foundation models for understanding cognition will require grounding them in established neuroscience and psychological theories, and identifying the specific mechanisms that underlie their predictive power. Only through such mechanistic understanding can foundation models generate testable hypotheses that genuinely advance our knowledge of human cognition.

Toward Mechanistic Understanding. This raises a practical question: how can foundation models be grounded in the theoretical frameworks of brain science? Conventional wisdom has long held that neural networks are "black boxes," antithetical to the kinds of explanations sought in neuroscience and psychology. Yet recent advances in interpretability are beginning to reveal computational structure at multiple levels.

The emerging field of mechanistic interpretability^{53,54} seeks to map functional subcircuits that reproducibly implement specific computations. Analyses of attention weights and hidden-layer activations reveal specialized components—for example, circuits that extend sequences or suppress irrelevant inputs⁵⁵. These elementary functions illustrate how complex behaviors emerge from combinations of simpler building blocks, which can be assembled into multi-step circuits for tasks like arithmetic⁵⁶ and factual recall⁵⁷.

Although these algorithmic "circuits" differ from biological ones, their compositional structure echoes theories of canonical microcircuits in neuroscience, where simple motifs are reused across the cortex⁵⁸. Recent theoretical work⁵⁹ characterizes transformers in terms of primitive operations, offering the beginnings of a computational theory that predicts both the tasks LLMs can solve and the neural mechanisms that might support them.

At the representational level, research has uncovered individual model units with selectivity for particular concepts, echoing long-standing debates in neuroscience about "grandmother cells" and the merits of interpreting neural codes at the level of single neurons versus distributed populations. Early interpretability studies reported specialized units, such as "cat neurons" in large computer vision models⁶⁰, "gender neurons" in early language models⁶¹, and more recently, a "Golden Gate Bridge feature"⁶² (a distinctive learned pattern in the model's internal representation) in LLMs. This line of work parallels findings in human neuroscience, where single-unit recordings have revealed "concept cells" in the medial temporal lobe⁶³—including neurons that respond selectively to landmarks such as the Sydney Opera House or to specific individuals such as Jennifer Aniston.

Many computations, however, are best viewed as distributed representations. Many LLM behaviors reduce to simple linear operations in high-dimensional activation spaces. Researchers can now "steer" language models through targeted interventions^{64,65}. For example,

given a neutral prompt such as the word "good" with no additional cues, the model can be directed to produce the antonym ("bad") by adding a vector pointing in the "antonym" direction, or to produce the Spanish translation ("bueno") by adding a vector pointing in the "Spanish" direction⁶⁶. These steering directions generalize across contexts, suggesting that semantic relationships are encoded as consistent geometric patterns.

This points toward testable neuroscience hypotheses. If human conceptual knowledge follows similar organizational principles, then semantic relations should be recoverable through linear operations on neural activity patterns—a prediction testable with neuroimaging or electrophysiology. Importantly, this "linear algebra" hypothesis is only one among the earliest testable predictions to emerge from interpretability research. As interpretability research expands and our understanding of artificial neural representations deepens, many more concrete, experimentally testable hypotheses about brain function are likely to follow.

Limitations and Opportunities. Should we expect foundation models to converge on true brain mechanisms? In AI, scaling—training larger models on bigger datasets with greater compute—drives reliable predictive gains captured by empirical scaling laws. In biology, by contrast, increases in representational capacity emerge through evolutionary and developmental constraints. This asymmetry helps explain why empirical scaling laws may yield predictive gains without converging on the path-dependent mechanisms of biology.

Approaches that embed evolutionary- and development-like constraints offer one path to bridge the gap between prediction and mechanistic explanation. For example, researchers can begin with biologically grounded architectures and test how optimization yields explanatory insights⁶⁷, or apply deep learning to identify developmental principles shaping visual representations⁶⁸. We develop this argument in more detail elsewhere, highlighting how the absence of evolutionary constraints may limit the explanatory value of current Al models⁶⁹. Such strategies may point the way toward foundation models that are not only predictive but also mechanistically explanatory.

But even if the mechanisms discovered in foundation models differ from those in the brain, interpretability can yield theoretical insights. Identifying computational mechanisms within these models and generating testable hypotheses fuels the crucial feedback loop between theory and experiment, driving scientific progress. Whether or not the mechanisms are "correct," this process advances our understanding of both artificial and biological intelligence.

Conclusion. Foundation models alone will not transform neuroscience. Their scientific value depends on moving beyond prediction to interpretation—understanding their computations, grounding them in brain sciences theory, and designing experiments to test those links. Digital twins may become powerful tools, but their value hinges on converting predictive structures into mechanisms grounded in brain science. As Ptolemy's epicycles remind us, even perfect prediction is no substitute for mechanistic explanation. The future of neuroscience with foundation models depends on whether we can transform data-fitting machines into theory-bearing scientific instruments—tools that reveal not only what intelligence can accomplish, but also how it works.

Acknowledgements

We would like to thank Wael Assad, David Badre, John Davenport, Alexander Fleischman, Mikey Lepori, Drew Linsley, Maria Grazia Ruocco, and Gretchen Schrafft for feedback on the manuscript. T.S. was supported by the Office of Naval Research (N00014-24-1-2026 and REPRISM MURI N00014-24-1-2603), the National Science Foundation (IIS-2402875 and EAR-1925481), and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004). E.P. was supported by a Young Faculty Award from the Defense Advanced Research Projects Agency (Grant #D24AP00261).

Declaration of Interest

E.P. is a paid consultant for Google DeepMind. The content of this article does not necessarily reflect that of the US Government or of Google, and no official endorsement of this work should be inferred.

References

- 1. Wang, E.Y., Fahey, P.G., Ding, Z., Papadopoulos, S., Ponder, K., Weis, M.A., Chang, A., Muhammad, T., Patel, S., Ding, Z., et al. (2025). Foundation model of neural activity predicts response to new stimulus types. Nature *640*, 470–477. https://doi.org/10.1038/s41586-025-08829-y.
- 2. Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M.K., Éltető, N., et al. (2025). A foundation model to predict and capture human cognition. Nature, 1–8. https://doi.org/10.1038/s41586-025-09215-4.
- 3. Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. Preprint.
- 4. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., and Ting, D.S.W. (2023). Large language models in medicine. Nat. Med. 29, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. In Findings of the Association for Computational Linguistics: EMNLP 2020, T. Cohn, Y. He, and Y. Liu, eds. (Association for Computational Linguistics), pp. 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261.
- Narendra, S., Shetty, K., and Ratnaparkhi, A. (2024). Enhancing Contract Negotiations with LLM-Based Legal Document Comparison. In Proceedings of the Natural Legal Language Processing Workshop 2024, N. Aletras, I. Chalkidis, L. Barrett, C. Goanţă, D. Preoţiuc-Pietro, and G. Spanakis, eds. (Association for Computational Linguistics), pp. 143–153. https://doi.org/10.18653/v1/2024.nllp-1.11.
- 7. Li, L., Li, J., Chen, C., Gui, F., Yang, H., Yu, C., Wang, Z., Cai, J., Zhou, J.A., Shen, B., et al.

- (2024). Political-LLM: Large Language Models in Political Science. Preprint at arXiv, https://doi.org/10.48550/arXiv.2412.06864 https://doi.org/10.48550/arXiv.2412.06864.
- 8. Yu, C., Ye, J., Li, Y., Li, Z., Ferrara, E., Hu, X., and Zhao, Y. (2025). A Large-Scale Simulation on Large Language Models for Decision-Making in Political Science. Preprint at arXiv, https://doi.org/10.48550/arXiv.2412.15291 https://doi.org/10.48550/arXiv.2412.15291.
- 9. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms. Preprint at arXiv, https://doi.org/10.48550/arXiv.1707.06347 https://doi.org/10.48550/arXiv.1707.06347.
- 10. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. In Advances in Neural Information Processing Systems (Curran Associates, Inc.).
- 11. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., and Finn, C. (2023). Direct preference optimization: your language model is secretly a reward model. In Proceedings of the 37th International Conference on Neural Information Processing Systems NIPS '23. (Curran Associates Inc.), pp. 53728–53741.
- 12. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., et al. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. Preprint at arXiv, https://doi.org/10.48550/arXiv.2402.03300 https://doi.org/10.48550/arXiv.2402.03300.
- 13. Heinz, M.V., Mackin, D.M., Trudeau, B.M., Bhattacharya, S., Wang, Y., Banta, H.A., Jewett, A.D., Salzhauer, A.J., Griffin, T.Z., and Jacobson, N.C. (2025). Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. NEJM AI 2, Aloa2400802. https://doi.org/10.1056/Aloa2400802.
- 14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17. (Curran Associates Inc.), pp. 6000–6010.
- 15. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, eds. (Association for Computational Linguistics), pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
- 16. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2022). On the Opportunities and Risks of Foundation Models. Preprint at arXiv, https://doi.org/10.48550/arXiv.2108.07258 https://doi.org/10.48550/arXiv.2108.07258.
- 17. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. In, pp. 9650–9660.
- 18. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A Simple Framework for

- Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning (PMLR), pp. 1597–1607.
- 19. Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Advances in Neural Information Processing Systems (Curran Associates, Inc.), pp. 12449–12460.
- 20. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating Large Language Models Trained on Code. Preprint at arXiv, https://doi.org/10.48550/arXiv.2107.03374 https://doi.org/10.48550/arXiv.2107.03374.
- 21. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. Preprint at arXiv, https://doi.org/10.48550/arXiv.2204.14198 https://doi.org/10.48550/arXiv.2204.14198.
- 22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. Preprint at arXiv, https://doi.org/10.48550/arXiv.2103.00020 https://doi.org/10.48550/arXiv.2103.00020.
- 23. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al. (2023). RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. Preprint at arXiv, https://doi.org/10.48550/arXiv.2307.15818 https://doi.org/10.48550/arXiv.2307.15818.
- 24. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. *118*, e2016239118. https://doi.org/10.1073/pnas.2016239118.
- 25. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature *630*, 493–500. https://doi.org/10.1038/s41586-024-07487-w.
- 27. Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al. (2025). Probabilistic weather forecasting with machine learning. Nature *637*, 84–90. https://doi.org/10.1038/s41586-024-08252-9.
- 28. Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. (2023). Learning skillful medium-range global weather forecasting. Science *382*, 1416–1421. https://doi.org/10.1126/science.adi2336.
- 29. Merchant, A., Batzner, S., Schoenholz, S.S., Aykol, M., Cheon, G., and Cubuk, E.D. (2023).

- Scaling deep learning for materials discovery. Nature *624*, 80–85. https://doi.org/10.1038/s41586-023-06735-9.
- 30. Oehrn, C.R., Cernera, S., Hammer, L.H., Shcherbakova, M., Yao, J., Hahn, A., Wang, S., Ostrem, J.L., Little, S., and Starr, P.A. (2024). Chronic adaptive deep brain stimulation versus conventional stimulation in Parkinson's disease: a blinded randomized feasibility trial. Nat. Med. *30*, 3345–3356. https://doi.org/10.1038/s41591-024-03196-z.
- 31. Abd-Alrazaq, A., AlSaad, R., Shuweihdi, F., Ahmed, A., Aziz, S., and Sheikh, J. (2023). Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. Npj Digit. Med. *6*, 84. https://doi.org/10.1038/s41746-023-00828-5.
- 32. Bruin, W.B., Abe, Y., Alonso, P., Anticevic, A., Backhausen, L.L., Balachander, S., Bargallo, N., Batistuzzo, M.C., Benedetti, F., Bertolin Triquell, S., et al. (2023). The functional connectome in obsessive-compulsive disorder: resting-state mega-analysis and machine learning classification for the ENIGMA-OCD consortium. Mol. Psychiatry *28*, 4307–4319. https://doi.org/10.1038/s41380-023-02077-0.
- 33. Lorach, H., Galvez, A., Spagnolo, V., Martel, F., Karakas, S., Intering, N., Vat, M., Faivre, O., Harte, C., Komi, S., et al. (2023). Walking naturally after spinal cord injury using a brain–spine interface. Nature *618*, 126–133. https://doi.org/10.1038/s41586-023-06094-5.
- 34. Caro, J.O., Fonseca, A.H. de O., Rizvi, S.A., Rosati, M., Averill, C., Cross, J.L., Mittal, P., Zappala, E., Dhodapkar, R.M., Abdallah, C., et al. (2024). BrainLM: A foundation model for brain activity recordings. In.
- 35. Wang, C., Jiang, Y., Peng, Z., Li, C., Bang, C., Zhao, L., Lv, J., Sepulcre, J., Yang, C., He, L., et al. (2025). Towards a general-purpose foundation model for fMRI analysis. Preprint at arXiv, https://doi.org/10.48550/arXiv.2506.11167 https://doi.org/10.48550/arXiv.2506.11167.
- 36. Vogt, N. (2025). Neuropixels for nonhuman primates. Nat. Methods *22*, 1622–1622. https://doi.org/10.1038/s41592-025-02791-3.
- 37. Steinmetz, N.A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., et al. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. Science *372*, eabf4588. https://doi.org/10.1126/science.abf4588.
- 38. Yao, Z., van Velthoven, C.T.J., Kunst, M., Zhang, M., McMillen, D., Lee, C., Jung, W., Goldy, J., Abdelhak, A., Baker, P., et al. (2023). A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. bioRxiv, 2023.03.06.531121. https://doi.org/10.1101/2023.03.06.531121.
- 39. Sandrone, S. (2024). Digital Twins in Neuroscience. J. Neurosci. 44. https://doi.org/10.1523/JNEUROSCI.0932-24.2024.
- 40. Li, K., Hopkins, A.K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. (2022). Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In.
- 41. Nanda, N. (2023). Actually, Othello-GPT Has A Linear Emergent World Representation.

- https://www.neelnanda.io/mechanistic-interpretability/othello.
- 42. Karvonen, A. (2024). Emergent World Models and Latent Variable Estimation in Chess-Playing Language Models. Preprint at arXiv, https://doi.org/10.48550/arXiv.2403.15498 https://doi.org/10.48550/arXiv.2403.15498.
- 43. Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. (2021). Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. Trans. Assoc. Comput. Linguist. 9, 160–175. https://doi.org/10.1162/tacl a 00359.
- 44. Tucker, M., Qian, P., and Levy, R. (2021). What if This Modified That? Syntactic Interventions with Counterfactual Embeddings. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, C. Zong, F. Xia, W. Li, and R. Navigli, eds. (Association for Computational Linguistics), pp. 862–875. https://doi.org/10.18653/v1/2021.findings-acl.76.
- 45. Russin, J., Pavlick, E., and Frank, M.J. (2025). Parallel trade-offs in human cognition and neural networks: The dynamic interplay between in-context and in-weight learning. Proc. Natl. Acad. Sci. *122*, e2510270122. https://doi.org/10.1073/pnas.2510270122.
- 46. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F.A. (2020). Shortcut learning in deep neural networks. Nat. Mach. Intell. 2, 665–673. https://doi.org/10.1038/s42256-020-00257-z.
- 47. jylin04, JackS, Karvonen, A., and Can (2024). OthelloGPT learned a bag of heuristics. https://www.lesswrong.com/posts/gcpNuEZnxAPayaKBY/othellogpt-learned-a-bag-of-heuristics-1.
- 48. Mitchell, M. (2025). LLMs and World Models, Part 2. Al Guide Think. Hum. https://aiguide.substack.com/p/llms-and-world-models-part-2.
- 49. Musker, S., Duchnowski, A., Millière, R., and Pavlick, E. (2025). LLMs as models for analogical reasoning. J. Mem. Lang. *145*, 104676. https://doi.org/10.1016/j.jml.2025.104676.
- 50. Vafa, K., Chang, P.G., Rambachan, A., and Mullainathan, S. (2025). What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models. In https://doi.org/10.48550/arXiv.2507.06952.
- 51. Namazova, S., Brondetta, A., Strittmatter, Y., Nassar, M., and Musslick, S. (2025). Not Yet AlphaFold for the Mind: Evaluating Centaur as a Synthetic Participant. Preprint at arXiv, https://doi.org/10.48550/arXiv.2508.07887 https://doi.org/10.48550/arXiv.2508.07887.
- 52. Bowers, J., Puebla, G., Thorat, S., Tsetsos, K., and Ludwig, C. (2025). Centaur: A model without a theory. Preprint at OSF, https://doi.org/10.31234/osf.io/v9w37_v3 https://doi.org/10.31234/osf.io/v9w37_v3.
- 53. Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020). Zoom In: An Introduction to Circuits. Distill *5*, 10.23915/distill.00024.001. https://doi.org/10.23915/distill.00024.001.

- 54. Saphra, N., and Wiegreffe, S. (2024). Mechanistic? Preprint at arXiv, https://doi.org/10.48550/arXiv.2410.09087 https://doi.org/10.48550/arXiv.2410.09087.
- 55. Olsson*, C., Elhage*, N., Nanda*, N., and et al (2022). In-context Learning and Induction Heads. Distill. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.
- 56. Quirke, P., and Barez, F. (2023). Understanding Addition in Transformers. In.
- 57. Geva, M., Bastings, J., Filippova, K., and Globerson, A. (2023). Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, H. Bouamor, J. Pino, and K. Bali, eds. (Association for Computational Linguistics), pp. 12216–12235. https://doi.org/10.18653/v1/2023.emnlp-main.751.
- 58. Douglas, R.J., and Martin, K.A.C. (2004). Neuronal circuits of the neocortex. Annu. Rev. Neurosci. 27, 419–451. https://doi.org/10.1146/annurev.neuro.27.070203.144152.
- 59. Weiss, G., Goldberg, Y., and Yahav, E. (2021). Thinking Like Transformers. In Proceedings of the 38th International Conference on Machine Learning (PMLR), pp. 11080–11090.
- 60. Le, Q.V. (2013). Building high-level features using large scale unsupervised learning. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE), pp. 8595–8598. https://doi.org/10.1109/ICASSP.2013.6639343.
- 61. Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In Advances in Neural Information Processing Systems (Curran Associates, Inc.), pp. 12388–12401.
- 62. Golden Gate Claude https://www.anthropic.com/news/golden-gate-claude.
- 63. Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. Nature *435*, 1102–1107. https://doi.org/10.1038/nature03687.
- 64. Turner, A.M., Thiergart, L., Leech, G., Udell, D., Vazquez, J.J., Mini, U., and MacDiarmid, M. (2024). Steering Language Models With Activation Engineering. Preprint at arXiv, https://doi.org/10.48550/arXiv.2308.10248 https://doi.org/10.48550/arXiv.2308.10248.
- 65. Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A.M. (2024). Steering Llama 2 via Contrastive Activation Addition. Preprint at arXiv, https://doi.org/10.48550/arXiv.2312.06681 https://doi.org/10.48550/arXiv.2312.06681.
- 66. Todd, E., Li, M., Sharma, A.S., Mueller, A., Wallace, B.C., and Bau, D. (2023). Function Vectors in Large Language Models. In.
- 67. Linsley, D., Kim, J., Ashok, A., and Serre, T. (2020). RECURRENT NEURAL CIRCUITS FOR CONTOUR DETECTION.
- 68. Sheybani, S., Hansaria, H., Wood, J.N., Smith, L.B., and Tigani, Z. Curriculum Learning with

Infant Egocentric Videos.

69. Linsley, D., Feng, P., and Serre, T. (2025). Better artificial intelligence does not mean better models of biology. Preprint at arXiv, https://doi.org/10.48550/arXiv.2504.16940 https://doi.org/10.48550/arXiv.2504.16940.