Lost in Translation: The Algorithmic Gap Between LMs and the Brain

Tommaso Tosato [1] **Pascal Jr Tikeng Notsawo** [2 3] **Saskia Helbling** [4] **Irina Rish** [2 3] **Guillaume Dumas** [1 2]

Abstract

Language Models (LMs) have achieved impressive performance on various linguistic tasks, but their relationship to human language processing in the brain remains unclear. This paper examines the gaps and overlaps between LMs and the brain at different levels of analysis, emphasizing the importance of looking beyond input-output behavior to examine and compare the internal processes of these systems. We discuss how insights from neuroscience, such as sparsity, modularity, internal states, and interactive learning, can inform the development of more biologically plausible language models. Furthermore, we explore the role of scaling laws in bridging the gap between LMs and human cognition, highlighting the need for efficiency constraints analogous to those in biological systems. By developing LMs that more closely mimic brain function, we aim to advance both artificial intelligence and our understanding of human cognition.

1. Introduction

Large Language Models (LMs) have achieved remarkable performance on a wide range of language tasks, from machine translation to question answering (Brown et al., 2020). This success has led to speculation that LMs may provide valid models of human intelligence. However, comparisons between LMs and the brain are often limited to the input-output level, which can be misleading.

It would be incorrect to assume that LMs, because they produce human-like linguistic outputs, must process information similarly to the human brain. David Marr's levels of analysis framework (Marr & Vision, 1982) provides a valuable lens for comparing LMs and human language processing. Marr proposed that information processing systems can be understood at three distinct levels:

- **Computational level:** What is the goal of the computation? This level focuses on the abstract function being computed, independent of how it is implemented. In this context, it refers to tasks like language understanding, question answering, and translating between languages. Comparisons of input-output behaviors pertain to this level.
- **Algorithmic level:** How are the computational goals achieved? This level deals with the specific methods and strategies employed to transform inputs into outputs. In this context, it examines the mechanisms used to parse sentences, represent semantic information, and generate coherent answers. Analyzing which operations intermediate representations undergo pertains to this level.
- **Implementational level:** How are these mechanisms physically realized? This level concerns the actual hardware (biological or artificial) that carries out the computations. For example, the brain uses spiking neurons and electrochemical signaling, while LMs run on silicon hardware with floating-point arithmetic.

Clarifying what aspects of cognition LMs are actually modeling is crucial. Mahowald et al. (2024) argue that while LMs exhibit remarkable *formal* linguistic competence (the ability to produce fluent and grammatical text), they still lack robust *functional* competence (the capacity to use language to reason and achieve real-world goals). However, their performance has led to the idea that, beyond modeling language perception, they may also model more complex cognitive abilities such as reasoning (Jones, 2024), theory of mind (Strachan et al., 2024), and building world models (Li et al., 2023).

Mitchell & Krakauer (2023) contend that despite their seemingly intelligent behavior, LMs do not understand the data they process in the same way humans do. The brittleness and lack of robust generalization in these models are key indicators of their lack of true understanding. LMs are fundamentally retrieval systems that generate outputs by recognizing and interpolating patterns in their training data

[1]CHU Sainte Justine Research Center, Department of Psychiatry, Université de Montréal, Montreal, QC, Canada [2]Mila, Montreal, QC, Canada [3]Department of Computer Science and Operation Research, Université de Montréal, Montréal, Canada [4]Ernst Strüngmann Institute for Neuroscience in Cooperation with Max Planck Society, Frankfurt am Main, Germany. Correspondence to: Tommaso Tosato <tosato.tommaso.office@gmail.com>.

Workshop on Large Language Models and Cognition, 41st International Conference on Machine Learning, Vienna, Austria.

(i.e., *approximate retrieval*) (Kambhampati, 2024). This leads to high performance for task instances which are well represented in the training data and failures for analogous instances which are less represented (McCoy et al., 2023; Wu et al., 2024; Lewis & Mitchell, 2024). In contrast, human understanding is based on rich, causal mental models of the world, rather than massive statistical correlations learned by LMs. These mental models enable humans to make robust predictions, generalizations, and analogies, to reason compositionally and counterfactually, and to intervene in the world to test hypotheses actively.

A key challenge in understanding the true relationship between LMs and the brain is to move beyond comparisons of input-output behavior and intermediate representations (i.e., the *computational level*), and focus on probing the internal processes of these systems (i.e., the *algorithmic level*) using causal interventions and mechanistic interpretability. This argument is developed in the first part of the paper. We then proceed to identify key properties in the architectures and training procedures of LMs which differ from the brain and propose possibilities to reconcile them. By doing so, we aim to develop LMs that better capture human cognitive processing, and as such, could serve as more faithful models of the brain.

2. Probing Representations and Processes

To determine whether LMs process information in ways analogous to the human brain, it is essential to look beyond surface-level behavior. While LLMs have recently been claimed to pass the Turing test (Jones & Bergen, 2024), Dumas et al. (2014) argue that the classic Turing test, which focuses solely on external behavior, is insufficient for investigating the genuineness of the internal mechanisms of machine intelligence. This is especially true for language, where the ELIZA effect can lead people to attribute human-like qualities to any system capable of generating coherent text (Weizenbaum, 1976). Moreover, performance comparisons between LMs and humans at the behavioral level are subject to several pitfalls. These include models trained specifically to solve benchmarks that claim to quantify general cognitive abilities but do so in arbitrary and limited ways (Liu et al., 2024), and the risk of data contamination, where test data may have been included in the model's training set. This section outlines possible approaches for comparing LMs and the brain beyond the input-output level:

Hierarchical Correspondences of Representations. Research in computer vision has shown that representations learned by Convolutional Neural Networks across layers correlate with brain activity in different regions of the visual hierarchy (Yamins & DiCarlo, 2016; Kriegeskorte, 2015; Khaligh-Razavi & Kriegeskorte, 2014). This was demonstrated by using techniques like representational similarity analysis (RSA) (Kriegeskorte et al., 2008) to compare the geometry of neural representations across models and brain regions. Similarly, in the language domain, studies have found correspondences between LM activations across layers and brain activity in different parts of the language network (Caucheteux et al., 2023; Millet et al., 2023). Shared representational spaces have been uncovered through *encoding models*, which produce simulated brain activity from LM embeddings through linear regression and correlate the simulated activity with recorded neural data (Goldstein et al., 2022). The correlation strength is reflected in an index called *brain score* (Schrimpf et al., 2021). However, it is important to note that both the visual system and the language network are not simple feedforward hierarchies but rather highly recurrent networks with pervasive feedback connections and extensive cross-talk between regions (Felleman & Van Essen, 1991; Pessoa, 2023). Moreover, it is crucial to understand that representational similarity between LMs and brains pertains primarily to the computational level, as it captures the input-output mappings of intermediate computational steps, but does not necessarily imply that similar operations or algorithms are used to transform those representations (Schyns et al., 2022; Antonello & Huth, 2024; Tuckute et al., 2024). Two systems can exhibit similar representations while differing substantially in their algorithms. This is evidenced by the fact that not only causal transformers (e.g., GPT), but also masked transformers (e.g., BERT), LSTMs and RNNs generate internal representations that correlate with brain activity, and there is no clear agreement on which one correlates best (Pasquiou et al., 2022; Anderson et al., 2021; Toneva & Wehbe, 2019; Oota et al., 2022). To make claims about algorithmic similarities, we need evaluation approaches that can probe the internal dynamics and causal structure of LMs (Belinkov, 2022) and compare them with results obtained using similar approaches in the brain.

Mechanistic Interpretability. This approach involves analyzing and understanding the internal components and processes of a machine learning model to explain how it transforms inputs into outputs at a granular, algorithmic level (Elhage et al., 2021). It can be used to find units or circuits within LMs that encode particular syntactic or semantic properties. However, individual neurons in LMs respond to a variety of seemingly unrelated and heterogeneous features, posing a challenge for interpreting the network's behavior. Recent work addresses this issue by using sparse autoencoders to extract monosemantic features from transformer language models (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024), including abstract concepts that generalize across modalities and languages. These learned features are significantly more

interpretable than the model's original neurons and exhibit interesting properties such as universality across different random seeds and feature splitting as the number of learned features increases. Comparing these features to those recorded in brain networks (Jamali et al., 2024) could provide a more accurate correspondence between LMs and the brain. Notably, neurons in the brain are also semantically heterogeneous to a certain degree (Leonard et al., 2024; Tye et al., 2024), suggesting that some degree of polysemanticity may be a general property of artificial and biological neural systems.

Ablation and Stimulation Studies. These techniques can provide causal insights in both the brain and LMs. Lesions (Broca et al., 1861; Dronkers et al., 2004) and electric stimulation (Doron & Brecht, 2015; Rofes et al., 2019) studies have allowed mapping of the functional organization of language in the human brain. More recently, optogenetics has enabled targeting and re-activating specific neuronal ensembles, which were previously activated by certain stimuli (Liu et al., 2012). Analogous experiments in LMs, such as pruning, editing or fixing specific weights, can reveal how different parts of the network contribute to linguistic behavior (Meng et al., 2023; Vig et al., 2020). Additionally, activating or suppressing specific extracted features can influence the model's output, and allow steering it in specific directions (Templeton et al., 2024; Marks et al., 2024). For instance, clamping certain features to high values can induce models to generate specific types of content or alter their behavior in predictable ways. Comparable effects of perturbations in LMs and the brain would suggest algorithmic equivalence.

Controlled Linguistic Probes and Interventions. These can be combined with interpretability techniques to gain a more mechanistic understanding of language processing in LMs (Marvin & Linzen, 2018; Geiger et al., 2021; Zhou et al., 2024). By systematically perturbing inputs or fine-tuning LMs on specific tasks, we can observe how internal representations change in response to these interventions. Similar approaches in neuroscience can provide complementary insights and offer ground for comparison.

Meta-representations as Representation of Processes . Proposed by Kanai et al. (2024), this framework offers a promising approach for comparing the internal workings of LMs and the brain at the algorithmic level by focusing on the processes that generate representations, rather than just the representations themselves. In the case of LMs, we can use the patterns of weights across different layers and attention heads as a way to represent the processes that generate the model's outputs. We can then compare the two at the algorithmic level by relating these weight patterns to process models of the brain (Dohmatob et al., 2020), based on, e.g., structural and functional connectivity.

3. Toward Brain-Inspired Language Models

While Language Models (LMs) have achieved human-like performance on several tasks, they still differ significantly from the human brain in the ways they achieve these results. As discussed in the previous section, better methods to test for these differences at the algorithmic level are needed. In this section, we explore how LMs and the brain differ in terms of architectures and learning dynamics, and how these differences lead to divergences at the algorithmic level. By examining these disparities, we can identify key areas where LMs could be modified to more closely resemble the brain's information processing mechanisms. The motivation for reconciling these differences is twofold: 1) By developing LMs that more closely mimic the brain's architecture and dynamics, we can provide neuroscientists with more accurate computational models of language processing in the brain. This can lead to new hypotheses and insights in neuroscience (Jain et al., 2024), fostering a deeper understanding of human cognition. 2) Incorporating brain-inspired features into LMs may result in artificial systems that exhibit more human-like intelligence. This could lead to AI that is more flexible, efficient, and capable of handling reasoning and planning in ways that current models struggle with.

Sparse Connectivity and Modularity. The brain is characterized by sparse connectivity and modularity, with specialized networks for different cognitive functions (Bassett et al., 2018; Seguin et al., 2023). In contrast, transformer models have dense, all-to-all connectivity. During development, the brain undergoes synaptic pruning, resulting in sparser, more efficient networks (Paolicelli et al., 2011). Similar pruning emerges in biologically inspired spiking neural networks (Volzhenin et al., 2022). To incorporate some level of modularity in LMs, Mixture-of-experts (MoE) architectures (Xue et al., 2024) can be employed, leading to improved computational efficiency and more functionally specialized sub-networks. However, the modules in MoE are not interconnected, hierarchically organized, or able to learn to attend to specific parts of the inputs. The Recurrent Independent Mechanisms (RIMs) architecture (Goyal et al., 2019) addresses these limitations by introducing sparse interactions among functionally specialized modules that interact sparsely through attention. This enables these modules to learn to attend to specific input parts and allows for hierarchical stacking. Sparsity can be promoted through adjustment of attention mechanisms (e.g., limiting the number of tokens each token attends to (Child et al., 2019)), choosing regularization methods alternative to L2 (e.g., L1 regularization and elastic net), and

applying dropout and post-training pruning.

Internal States. The human brain maintains and constantly updates rich internal representations over time, whereas LMs have a fixed context window, and only feedforward connections. Incorporating internal states into LMs could change the way they maintain and use context. Recent work on adding recurrence to transformer models (Hwang et al., 2024) is a step in this direction. State space models (Gu & Dao, 2023), which explicitly model the dynamics of latent states, are another promising approach. Models that learn to update and maintain internal representations, may also dynamically encode and integrate information over time in a brain-like manner (Hasson et al., 2015). Moreover, recurrent processing appears to be critical for consciousness (Chalmers, 2023).

Learning Dynamics. The brain's learning process is characterized by continual, incremental acquisition of knowledge over a lifetime, utilizing moderate resources efficiently. A key feature of brain development is the presence of critical periods, during which the brain is especially receptive to specific types of input and learning experiences (Hensch, 2005). Moreover, brain learning dynamics maintain a delicate balance between acquiring new information and selectively forgetting unnecessary details. In contrast, current LMs are typically trained from scratch using random weight initialization on massive corpora, requiring multiple passes over large datasets and consuming enormous computational resources. Additionally, fine-tuning or training LMs on new tasks often leads to catastrophic forgetting (Kirkpatrick et al., 2017), highlighting the lack of balance in retaining old knowledge while acquiring new information. To enable more brain-like learning in LMs, several approaches can be considered: (1) *Continual Learning*: Adapting continual learning techniques such as elastic weight consolidation (Kirkpatrick et al., 2017) and experience replay (Rolnick et al., 2019) could help LMs maintain a balance between learning new information and preserving important past knowledge. (2) *Curriculum Learning:* Gradually increasing the difficulty of training data (Bengio et al., 2009) can enhance learning efficiency. The PAIRED algorithm (Dennis et al., 2020), which uses a multi-agent setup where an adversary generates increasingly challenging environments for the learning agent, is a possible approach to implementing adaptive curriculum learning. (3) *Critical Periods:* Implementing a system of "sensitive periods" in LM training could mimic the brain's critical periods. This could involve dynamically adjusting learning rates for different parts of the network during training. For instance, early in training, the model could have higher plasticity in lower layers to learn basic linguistic features, gradually shifting focus to higher-level abstractions in later stages. Interestingly, current artificial neural networks already show some similarities to biological learning, exhibiting progressive learning of features of increasing complexity (Nakkiran et al., 2019; Mangalam & Prabhu, 2019).

Embodiment, Grounding, and Active Learning. Humans learn language through situated interactions with the physical and social world, grounding linguistic meanings in perceptual, motor, and affective experiences (Bisk et al., 2020; Barsalou, 2008; Di Paolo et al., 2018). LMs, in contrast, are typically trained on disembodied text data with a fixed training objective, limiting their ability to capture the full depth of linguistic meaning. To address this limitation, multimodal models that integrate information from vision, audition, and other sensory modalities are a promising approach (Team, 2024). However, true embodiment may require more than just passive perception—it may need active interaction and exploration (Engel et al., 2013). Coupling LMs with external execution modules that can perform actions or specialized reasoning tasks could provide an action space. However, just adding an action space when the model's weights are frozen won't generate active learning. For true active learning, the model needs to be able to select the training data through actions. Integrating LMs into robotic systems (Tellex et al., 2020; Collaboration et al., 2024) is another avenue for grounding language in embodied experience.

Social Skills and Interactive Learning. The brain's language system is fundamentally geared towards communication and social interaction, with dedicated neural mechanisms for cognitive skills like pragmatics, social reasoning, and theory of mind (Scott, 2019; Hagoort & Indefrey, 2014). While recent studies have shown that LMs are starting to exhibit some of these abilities (Strachan et al., 2024), it is unclear if they are achieving these capabilities in a way related to humans. To improve LMs' social abilities, they could be trained on socially interactive data, such as multi-turn dialogues, incorporating explicit representations of communicative intentions and the mental states of interlocutors. Training objectives that maximize relevance for a communicative context or reward successful and attuned communication in interactive settings (Jaques et al., 2019) could lead to more human-like language use. Multi-agent setups where LMs interact with each other or humans in goal-directed dialogues are promising (Bolotta & Dumas, 2022). By engaging in cooperative or competitive language games, LMs could learn to reason about the beliefs and intentions of other agents. Inverse reinforcement learning techniques (Hadfield-Menell et al., 2016) could help the LMs to infer and align with the communicative goals of human interlocutors.

Reasoning and Compositional Representations. LMs' ability to perform reasoning and planning remains limited

compared to humans (Wu et al., 2023; Valmeekam et al., 2023; Zhou et al., 2023). Techniques like chain-of-thought prompting (Wei et al., 2023) and the use of scratchpads (Nye et al., 2021) to generate step-by-step reasoning traces, have shown promise in improving reasoning abilities. Using reasoning traces as fine-tuning data has also been shown to enhance performance (Zelikman et al., 2022). However, the brain's ability to perform logical reasoning likely relies on more structured, compositional representations of meaning (Dehaene et al., 2015; Lake et al., 2017). GFlowNets (Bengio et al., 2021) are generative models that can sample from complex joint distributions, such as the distribution over sentences and their meanings. GFlowNets could potentially capture the brain's ability to generate diverse outputs and learn structured, compositional representations. Initial applications of GFlowNets to LMs are already showing promising results (Hu et al., 2024). Another idea is to move beyond next-word prediction as the primary training objective for LMs. The brain does not simply predict the next word in a sequence but constructs rich, hierarchical meaning representations that span multiple words and sentences (Heilbron et al., 2022; Lerner et al., 2011). Training objectives that encourage LMs to predict larger chunks of text, such as entire sentences or paragraphs, could lead to more structured and compositional representations (Gloeckle et al., 2024).

Oscillatory Dynamics. One key feature of language processing in the brain is the presence of oscillatory neural dynamics at multiple timescales, with different frequency bands appearing to track the structure of language at different levels, from phonemes to words to phrases (Giraud & Poeppel, 2012; Ding et al., 2016; Gross et al., 2013). These oscillations are thought to play a crucial role in segmenting and chunking continuous speech into discrete units. Incorporating oscillatory mechanisms in deep learning architectures, especially when the input is raw auditory signals, could potentially improve their efficiency and lead to more human-like speech processing, given the existing gaps (Tuckute et al., 2023). Simulation work has been exploring substituting standard neural network nodes with damped harmonic oscillators (Rusch & Mishra, 2020; Effenberger et al., 2022), demonstrating the potential use of oscillatory dynamics for computation.

Allometry and Scaling Laws. Allometry studies the relationship between the size and shape of biological systems. Scaling laws in the human brain, such as the relationship between brain size and neuron count, play a crucial role in the evolution of cognitive abilities, including language (Changeux et al., 2021). Recently, the field of neural scaling laws has similarly started to chart the relationships between the number of parameters, dataset size, computing cost, and performance of machine learning models (Kaplan et al., 2020). These scaling relationships can be seen as global signatures of complex systems (West, 2018), providing a quantitative way to compare brains and LMs. However, LMs benefit from virtually unlimited computational resources, leading to overparameterization without the same efficiency constraints as biological systems (West, 1999). To promote algorithmic similarity, it may be essential to introduce analogous pressures on LMs. Implementing such constraints can help bridge the gap between the computational prowess of LMs and the adaptive efficiency of biological neural systems, ensuring more frugal and meaningful application of scaling laws in LM development.

4. Conclusion

The rapid advancement of Language Models (LMs) has led to impressive performance on various linguistic tasks, prompting comparisons with human cognition. However, this paper argues that despite achieving human-level performance in several tasks (computational level equivalence), LMs exhibit significant divergence from human cognition in how these performances are achieved (algorithmic level discrepancy). We contend that the intelligence emerging from current LMs is fundamentally different from human cognition. As models continue to scale, this divergence may widen, potentially resulting in AI systems that excel at specific tasks but lack the hallmark characteristics of human intelligence: flexibility, robustness, and generalization capabilities.

In this paper, we first examined approaches for comparing LMs and the brain beyond input-output level assessments. We highlighted the limitations of solely analyzing intermediate representations and emphasized the need to understand the processes driving transformations between these representations. We then explored major differences in architecture and training dynamics between LMs and biological neural networks that likely contribute to algorithmic-level divergence, and we proposed several neuroscience-inspired approaches to narrow this divide.

While these biologically-inspired approaches may not immediately enhance computational efficiency or task performance, they are crucial for developing LMs that can serve as more accurate models of human cognition. Such models would offer in silico representations of cognitive processes to neuroscientists and psychologists, illuminating key mechanisms underlying human-like intelligence. Additionally, by constraining the space of possible algorithms to those that are more biologically plausible, we may discover novel architectures and learning paradigms that capture the key principles of human cognition, and lead to AI systems that exhibit the adaptability and generalization capabilities characteristic of human intelligence.

References

Anderson, A. J., Kiela, D., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D., Grimm, S., and Lalor, E. C. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience*, 41(18):4100–4119, 2021.

Antonello, R. and Huth, A. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1): 64–79, 2024.

Barsalou, L. W. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008.

Bassett, D. S., Zurn, P., and Gold, J. I. On the nature and use of models in network neuroscience. *Nature Reviews Neuroscience*, 19(9):566–578, 2018.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Bengio, Y., Jain, M., Korablyov, M., Precup, D., and Bengio, S. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., and Turian, J. Experience grounds language, 2020.

Bolotta, S. and Dumas, G. Social neuro ai: Social interaction as the 'dark matter' of ai. *Frontiers in Computer Science*, 4, 2022.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, Oct 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Broca, P. et al. Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bulletin de la Société Anatomique*, 6: 330–357, 1861.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Caucheteux, C., Gramfort, A., and King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441, 2023.

Chalmers, D. J. Could a large language model be conscious?, 2023. URL https://arxiv.org/abs/2303.07103.

Changeux, J.-P., Goulas, A., and Hilgetag, C. C. A connectomic hypothesis for the hominization of the brain. *Cerebral cortex*, 31(5):2425–2449, 2021.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Collaboration, E., O'Neill, A., Rehman, A., Gupta, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., Tung, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Gupta, A., Wang, A., Kolobov, A., Singh, A., Garg, A., Kembhavi, A., Xie, A., Brohan, A., Raffin, A., Sharma, A., Yavary, A., Jain, A., Balakrishna, A., Wahid, A., Burgess-Limerick, B., Kim, B., Schölkopf, B., Wulfe, B., Ichter, B., Lu, C., Xu, C., Le, C., Finn, C., Wang, C., Xu, C., Chi, C., Huang, C., Chan, C., Agia, C., Pan, C., Fu, C., Devin, C., Xu, D., Morton, D., Driess, D., Chen, D., Pathak, D., Shah, D., Büchler, D., Jayaraman, D., Kalashnikov, D., Sadigh, D., Johns, E., Foster, E., Liu, F., Ceola, F., Xia, F., Zhao, F., Frujeri, F. V., Stulp, F., Zhou, G., Sukhatme, G. S., Salhotra, G., Yan, G., Feng, G., Schiavi, G., Berseth, G., Kahn, G., Yang, G., Wang, G., Su, H., Fang, H.-S., Shi, H., Bao, H., Amor, H. B., Christensen, H. I., Furuta, H., Bharadhwaj, H., Walke, H., Fang, H., Ha, H., Mordatch, I., Radosavovic, I., Leal, I., Liang, J., Abou-Chakra, J., Kim, J., Drake, J., Peters, J., Schneider, J., Hsu, J., Vakil, J., Bohg, J., Bingham, J., Wu, J., Gao, J., Hu, J., Wu, J., Wu, J., Sun, J., Luo, J., Gu, J., Tan, J., Oh, J., Wu, J., Lu, J., Yang, J., Malik, J., Silvério, J., Hejna, J., Booher, J., Tompson, J., Yang, J., Salvador, J., Lim, J. J., Han, J., Wang, K., Rao, K., Pertsch, K., Hausman, K., Go, K., Gopalakrishnan, K., Goldberg, K., Byrne, K., Oslund, K., Kawaharazuka, K., Black, K., Lin, K., Zhang, K., Ehsani, K., Lekkala, K., Ellis, K., Rana, K., Srinivasan, K., Fang, K., Singh,

K. P., Zeng, K.-H., Hatch, K., Hsu, K., Itti, L., Chen, L. Y., Pinto, L., Fei-Fei, L., Tan, L., Fan, L. J., Ott, L., Lee, L., Weihs, L., Chen, M., Lepert, M., Memmel, M., Tomizuka, M., Itkina, M., Castro, M. G., Spero, M., Du, M., Ahn, M., Yip, M. C., Zhang, M., Ding, M., Heo, M., Srirama, M. K., Sharma, M., Kim, M. J., Kanazawa, N., Hansen, N., Heess, N., Joshi, N. J., Suenderhauf, N., Liu, N., Palo, N. D., Shafiullah, N. M. M., Mees, O., Kroemer, O., Bastani, O., Sanketi, P. R., Miller, P. T., Yin, P., Wohlhart, P., Xu, P., Fagan, P. D., Mitrano, P., Sermanet, P., Abbeel, P., Sundaresan, P., Chen, Q., Vuong, Q., Rafailov, R., Tian, R., Doshi, R., Mart'in-Mart'in, R., Baijal, R., Scalise, R., Hendrix, R., Lin, R., Qian, R., Zhang, R., Mendonca, R., Shah, R., Hoque, R., Julian, R., Bustamante, S., Kirmani, S., Levine, S., Lin, S., Moore, S., Bahl, S., Dass, S., Sonawani, S., Tulsiani, S., Song, S., Xu, S., Haldar, S., Karamcheti, S., Adebola, S., Guist, S., Nasiriany, S., Schaal, S., Welker, S., Tian, S., Ramamoorthy, S., Dasari, S., Belkhale, S., Park, S., Nair, S., Mirchandani, S., Osa, T., Gupta, T., Harada, T., Matsushima, T., Xiao, T., Kollar, T., Yu, T., Ding, T., Davchev, T., Zhao, T. Z., Armstrong, T., Darrell, T., Chung, T., Jain, V., Kumar, V., Vanhoucke, V., Zhan, W., Zhou, W., Burgard, W., Chen, X., Chen, X., Wang, X., Zhu, X., Geng, X., Liu, X., Liangwei, X., Li, X., Pang, Y., Lu, Y., Ma, Y. J., Kim, Y., Chebotar, Y., Zhou, Y., Zhu, Y., Wu, Y., Xu, Y., Wang, Y., Bisk, Y., Dou, Y., Cho, Y., Lee, Y., Cui, Y., Cao, Y., Wu, Y.-H., Tang, Y., Zhu, Y., Zhang, Y., Jiang, Y., Li, Y., Li, Y., Iwasawa, Y., Matsuo, Y., Ma, Z., Xu, Z., Cui, Z. J., Zhang, Z., Fu, Z., and Lin, Z. Open x-embodiment: Robotic learning datasets and rt-x models, 2024. URL https://arxiv.org/abs/2310.08864.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, 2023.

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., and Pallier, C. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron*, 88(1):2–19, 2015.

Dennis, M., Jaques, N., Vinitsky, E., Bayen, A., Russell, S., Critch, A., and Levine, S. Emergent complexity and zero-shot transfer via unsupervised environment design. *Advances in neural information processing systems*, 33: 13049–13061, 2020.

Di Paolo, E. A., Cuffari, E. C., and De Jaegher, H. *Linguistic bodies: The continuity between life and language*. MIT press, 2018.

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1):158–164, 2016.

Dohmatob, E., Dumas, G., and Bzdok, D. Dark control: The default mode network as a reinforcement learning agent. *Human brain mapping*, 41(12):3318–3341, 2020.

Doron, G. and Brecht, M. What single-cell stimulation has told us about neural coding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1677): 20140204, 2015.

Dronkers, N. F., Wilkins, D. P., Van Valin Jr, R. D., Redfern, B. B., and Jaeger, J. J. Lesion analysis of the brain areas involved in language comprehension. *Cognition*, 92(1-2):145–177, 2004.

Dumas, G., de Guzman, G. C., Tognoli, E., and Kelso, J. S. The human dynamic clamp as a paradigm for social interaction. *Proceedings of the National Academy of Sciences*, 111(35):E3726–E3734, 2014.

Effenberger, F., Carvalho, P., Dubinin, I., and Singer, W. The functional role of oscillatory dynamics in neocortical circuits: a computational perspective. *bioRxiv*, pp. 2022–11, 2022.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Engel, A. K., Maye, A., Kurthen, M., and König, P. Where's the action? the pragmatic turn in cognitive science. *Trends in cognitive sciences*, 17(5):202–209, 2013.

Felleman, D. J. and Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Giraud, A.-L. and Poeppel, D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, 15(4):511–517, 2012.

Gloeckle, F., Idrissi, B. Y., Rozière, B., Lopez-Paz, D., and Synnaeve, G. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv: 2404.19737*, 2024.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Co-

hen, A., et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., and Garrod, S. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS biology*, 11(12):e1001752, 2013.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Hagoort, P. and Indefrey, P. The neurobiology of language beyond single words. *Annual review of neuroscience*, 37: 347–362, 2014.

Hasson, U., Chen, J., and Honey, C. J. Hierarchical process memory: memory as an integral component of information processing. *Trends in cognitive sciences*, 19 (6):304–313, 2015.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and De Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32): e2201968119, 2022.

Hensch, T. K. Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience*, 6(11):877–888, 2005.

Hu, E. J., Jain, M., Elmoznino, E., Kaddar, Y., Lajoie, G., Bengio, Y., and Malkin, N. Amortizing intractable inference in large language models, 2024.

Hwang, D., Wang, W., Huo, Z., Sim, K. C., and Mengibar, P. M. Transformerfam: Feedback attention is working memory, 2024.

Jain, S., Vo, V. A., Wehbe, L., and Huth, A. G. Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, 5(1):80–106, 2024.

Jamali, M., Grannan, B., Cai, J., Khanna, A. R., Muñoz, W., Caprara, I., Paulk, A. C., Cash, S. S., Fedorenko, E., and Williams, Z. M. Semantic encoding during language comprehension at single-cell resolution. *Nature*, pp. 1–7, 2024.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pp. 3040–3049. PMLR, 2019.

Jones, C. R. and Bergen, B. K. People cannot distinguish gpt-4 from a human in a turing test, 2024. URL https://arxiv.org/abs/2405.08007.

Jones, N. Ai now beats humans at basic tasks—new benchmarks are needed, says major report. *Nature*, 628(8009): 700–701, 2024.

Kambhampati, S. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18, 2024.

Kanai, R., Takatsuki, R., and Fujisawa, I. Meta-representations as representations of processes, Mar 2024. URL osf.io/preprints/psyarxiv/zg27u.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11): e1003915, 2014.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Leonard, M. K., Gwilliams, L., Sellers, K. K., Chung, J. E., Xu, D., Mischler, G., Mesgarani, N., Welkenhuysen, M., Dutta, B., and Chang, E. F. Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature*, 626(7999):593–602, 2024.

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. To-

pographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915, 2011.

Lewis, M. and Mitchell, M. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*, 2024.

Li, J., Karamolegkou, A., Kementchedjhieva, Y., Abdou, M., Lehmann, S., and Søgaard, A. Structural similarities between language models and neural response measurements. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023. URL https://openreview.net/forum?id=ZobkKCTaiY.

Liu, X., Ramirez, S., Pang, P. T., Puryear, C. B., Govindarajan, A., Deisseroth, K., and Tonegawa, S. Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature*, 484:381–385, 2012. doi: 10.1038/nature11028.

Liu, Y. L., Blodgett, S. L., Cheung, J. C. K., Liao, Q. V., Olteanu, A., and Xiao, Z. Ecbd: Evidence-centered benchmark design for nlp. *arXiv preprint arXiv:2406.08723*, 2024.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024.

Mangalam, K. and Prabhu, V. Do deep neural networks learn shallow learnable examples first? In *Proceedings of the Workshop on Identifying and Understanding Deep Learning Phenomena at 36th International Conference on Machine Learning*. PMLR, 2019.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, 2024. URL https://arxiv.org/abs/2403.19647.

Marr, D. and Vision, A. Vision: A computational investigation into the human representation and processing of visual information. 1982.

Marvin, R. and Linzen, T. Targeted syntactic evaluation of language models, 2018.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., and Griffiths, T. L. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*, 2023.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt, 2023.

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., and King, J.-R. Toward a realistic model of speech processing in the brain with self-supervised learning, 2023.

Mitchell, M. and Krakauer, D. C. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13): e2215907120, 2023.

Nakkiran, P., Kaplun, G., Kalimeris, D., Yang, T., Edelman, B. L., Zhang, F., and Barak, B. Sgd on neural networks learns functions of increasing complexity. *arXiv preprint arXiv:1905.11604*, 2019.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Oota, S. R., Alexandre, F., and Hinaut, X. Long-term plausibility of language models and neural dynamics during narrative listening. In *Proceedings of the annual meeting of the cognitive science society*, volume 44, 2022.

Paolicelli, R. C., Bolasco, G., Pagani, F., Maggi, L., Scianni, M., Panzanelli, P., Giustetto, M., Ferreira, T. A., Guiducci, E., Dumas, L., et al. Synaptic pruning by microglia is necessary for normal brain development. *science*, 333(6048):1456–1458, 2011.

Pasquiou, A., Lakretz, Y., Hale, J., Thirion, B., and Pallier, C. Neural language models are not born equal to fit brain data, but training helps. *arXiv preprint arXiv:2207.03380*, 2022.

Pessoa, L. The entangled brain. *Journal of cognitive neuroscience*, 35(3):349–360, 2023.

Rofes, A., Mandonnet, E., de Aguiar, V., Rapp, B., Tsapkini, K., and Miceli, G. Language processing from the perspective of electrical stimulation mapping. *Cognitive neuropsychology*, 36(3-4):117–139, 2019.

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Rusch, T. K. and Mishra, S. Coupled oscillatory recurrent neural network (cornn): An accurate and (gradient) stable architecture for learning long time dependencies. *arXiv preprint arXiv:2010.00951*, 2020.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., and Fedorenko, E. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):

e2105646118, 2021.

Schyns, P. G., Snoek, L., and Daube, C. Degrees of algorithmic equivalence between the brain and its dnn models. *Trends in Cognitive Sciences*, 26(12):1090–1102, 2022.

Scott, S. K. From speech and talkers to the social world: The neural processing of human spoken language. *Science*, 366(6461):58–62, 2019.

Seguin, C., Sporns, O., and Zalesky, A. Brain network communication: concepts, models and applications. *Nature reviews neuroscience*, 24(9):557–574, 2023.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., and Becchio, C. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 2024. doi: 10.1038/s41562-024-01882-z.

Team, C. Chameleon: Mixed-modal early-fusion foundation models, 2024.

Tellex, S., Gopalan, N., Kress-Gazit, H., and Matuszek, C. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55, 2020.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Toneva, M. and Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32, 2019.

Tuckute, G., Feather, J., Boebinger, D., and McDermott, J. H. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21 (12):e3002366, 2023.

Tuckute, G., Kanwisher, N., and Fedorenko, E. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47, 2024.

Tye, K. M., Miller, E. K., Taschbach, F. H., Benna, M. K., Rigotti, M., and Fusi, S. Mixed selectivity: Cellular computations for complexity. *Neuron*, 2024.

Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S., and Kambhampati, S. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change, 2023.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33: 12388–12401, 2020.

Volzhenin, K., Changeux, J.-P., and Dumas, G. Multilevel development of cognitive abilities in an artificial neural network. *Proceedings of the National Academy of Sciences*, 119(39):e2201304119, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Weizenbaum, J. Computer power and human reason: From judgment to calculation. 1976.

West, G. *Scale: The universal laws of life, growth, and death in organisms, cities, and companies*. Penguin, 2018.

West, G. B. The origin of universal scaling laws in biology. *Physica A: Statistical Mechanics and its Applications*, 263(1-4):104–113, 1999.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*, 2023.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks, 2024. URL https://arxiv.org/abs/2307.02477.

Xue, F., Zheng, Z., Fu, Y., Ni, J., Zheng, Z., Zhou, W., and You, Y. Openmoe: An early effort on open mixture-of-experts language models, 2024.

Yamins, D. L. and DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. Star: Bootstrapping reasoning with reasoning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2

Zhou, H., Bradley, A., Littwin, E., Razin, N., Saremi, O., Susskind, J., Bengio, S., and Nakkiran, P. What algo-

rithms can transformers learn? a study in length generalization. *arXiv preprint arXiv:2310.16028*, 2023.

Zhou, S., Weissweiler, L., He, T., Schütze, H., Mortensen, D. R., and Levin, L. Constructions are so difficult that even large language models get them right for the wrong reasons. *arXiv preprint arXiv:2403.17760*, 2024.