MIT Press • Open Encyclopedia of Cognitive Science

Transformers

Raphaël Millière[1]

[1]**Macquarie University**

MIT Press

Published on: Jul 17, 2025

URL: https://oecs.mit.edu/pub/ppxhxe2b

Transformers are a type of artificial neural network architecture designed to process sequential data—such as text—by selectively focusing on relationships between different elements in the sequence through a mechanism called attention. Unlike earlier approaches that processed sequences one element at a time, transformers can process multiple elements in parallel and capture both local and long-range dependencies between them. The transformer architecture forms the foundation of modern large language models that can engage in human-like dialogue, translate languages, answer questions, and generate creative text. Although originally developed for natural language processing, transformers have since been adapted for other domains including computer vision and robotics.

History

The transformer architecture emerged from research in neural machine translation, in which the goal was to translate text between languages automatically. Before transformers, the dominant approaches used recurrent neural networks (RNNs; Cho et al., 2014; Elman, 1990) and long short-term memory networks (LSTMs; Hochreiter & Schmidhuber, 1997; Sutskever et al., 2014) [see Recurrent Neural Networks]. These models rely on recurrence, a method in which the network reads a sequence one word at a time, updating an internal memory (or *hidden state*) at each step to carry context forward. Although effective for this kind of sequential processing on short texts, these approaches struggled with longer ones, as they had difficulty maintaining context over long distances.

In 2017, researchers at Google Brain introduced the transformer architecture in their influential paper, “Attention Is All You Need” (Vaswani et al., 2017). The key innovation was replacing recurrence with a mechanism called *self-attention*, which allowed the model to directly compute relationships between all words in a sequence, regardless of their distance from each other. The success of the original transformer quickly led to variations outperforming other architectures on many natural language processing and computer vision tasks.

Core concepts

The defining feature of transformers is the self-attention mechanism, which allows the model to weigh the importance of different elements—such as words—when processing a sequence. Self-attention operates through two main steps (see video):

1. For each element, the model generates three vectors known as the query, the key, and the value. To decide relevance, the query vector of the current element is compared against the key vector of every other element in the sequence. This comparison produces a compatibility score; a high score means the elements are highly relevant to one another. These scores are then normalized into attention weights that determine how much focus to apply.
2. The model uses these attention weights to selectively combine information. Each item’s value vector contributes to the final output in proportion to its attention weight, allowing the model to create a context-

aware representation of each element.

This process happens in parallel across the sequence through multiple *attention heads* that can each capture different types of relationships between elements. For example, in a transformer model processing text data, different heads might learn to focus on syntactic versus semantic relationships, allowing the model to build rich representations of language.

Visit the web version of this article to view interactive content.

The attention mechanism in Transformers (simplified)

Questions, controversies, and new developments

Despite intense competition to develop new architectures, transformers have become remarkably dominant across many areas of machine learning since their introduction. The reasons behind their enduring success remain a matter of debate.

Compared to other neural network architectures, transformers have relatively weak inductive biases—the built-in assumptions a model uses to learn and make predictions. For instance, they lack the inherent sequential processing and recursion built into recurrent architectures like RNNs and LSTMs. These recurrent models are designed to process information one piece at a time in a strict order, updating a single memory state at each step. This assumes a linear structure and creates a memory bottleneck, as all relevant information from the past must be compressed into that single state. Transformers also lack the strong assumptions of convolutional neural networks (CNNs), which are built on locality (important features are found in small, local patches of data, like adjacent pixels in an image) and translation equivariance (a feature is the same regardless of where it appears).

Because transformers are not constrained by strong assumptions about sequential order or locality, they can weigh the influence of every input element against every other, regardless of their position. This does not mean they lack inductive biases altogether; the self-attention mechanism has a bias toward learning relationships between elements based on their content rather than just their sequential position. Although transformers can struggle with tasks requiring precise sequential processing (like counting), they are well equipped to capture long-range dependencies between elements. As such, the architecture provides enough structure to make learning efficient while remaining flexible enough to adapt to diverse tasks.

Another notable ingredient in the success of transformers is their scalability. Because they process all elements in a sequence simultaneously, they are easier to train than RNNs on modern parallel computing hardware like graphics processing units that can perform many calculations at once. In addition, their performance on language modeling appears to improve predictably according to power laws as model size, dataset size, and

computational resources are increased (Kaplan et al., 2020). Although this suggests that the dramatic progress of language models can be largely attributed to the benefits of scale rather than architectural breakthroughs, it should be noted that different architectures exhibit markedly different scaling properties [see Large Language Models]. Indeed, the standard transformer's performance improves more consistently and predictably with increased scale than many of its variants, which can plateau or show diminishing return (Tay et al., 2022).

Broader connections

Unlike CNNs, which were inspired by the architecture of the mammalian primary visual cortex, the transformer architecture was not directly inspired by insights from cognitive science—despite the potentially misleading choice of "attention" as the label for its central mechanism (Lindsay, 2020). Nonetheless, researchers have become increasingly interested in its potential to shed light on aspects of human cognition (Frank, 2023; Millière, 2024). One advantage of modern transformers is their *stimulus computability*: the ability to process the same rich stimuli given to human participants (like text and images) rather than the simplified or schematized representations required by previous models (Frank & Goodman, 2025). This has allowed them to match patterns of human performance on classical psychology tasks at a scale and generality not previously possible (Bhatia & Richie, 2024; Lampinen et al., 2024; Strachan et al., 2024; Webb et al., 2023). In the linguistic domain, they can acquire sophisticated linguistic knowledge and predict neural responses during language comprehension with remarkable accuracy (Millière, 2024; Tuckute et al., 2024). In the visual domain, they exhibit a more human-like shape bias compared to CNNs (Dehghani et al., 2023; Tuli et al., 2021) and predict neural responses to natural scenes (Adeli et al., 2023).

However, these findings should be interpreted with caution. Behaviorally, transformers still exhibit striking failure modes and lack human-like robustness. They struggle with abstract visual reasoning tasks compared to humans (Puebla & Bowers, 2024), may be overly reliant on statistical patterns from their training data (McCoy et al., 2024), and their mastery of formal linguistic patterns may not translate to broader reasoning abilities (Mahowald et al., 2024). The typical learning conditions of transformer-based language models also differ substantially from human learning in terms of both the scope and nature of their training data—although recent work has begun addressing some of these limitations by training smaller models in developmentally plausible learning scenarios (Warstadt et al., 2023). This involves, for example, limiting training data to a human-scale quantity (millions of words, not billions) and using more naturalistic sources like transcribed parent–child conversations instead of curated web text. There is also an ongoing debate about whether transformers' success in predicting neural activity might reflect superficial correlations rather than deep mechanistic similarities (Bowers et al., 2022).

Regardless of whether transformers fulfill their promise as cognitive models, they have gained increased adoption as research tools in cognitive science to facilitate tasks such as data annotation, text analysis, and stimulus design (Abdurahman et al., 2024; Bzdok et al., 2024; Demszky et al., 2023). Conversely, the methods

of cognitive science can be adapted to investigate the capacities and limitations of transformer-based models beyond simple behavioral benchmarks (Frank, 2023).

Further reading

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *arXiv*. https://doi.org/10.48550/arXiv.2010.11929
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 5998–6008). Curran Associates, Inc.

References

- Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, *3*(7), pgae245. https://doi.org/10.1093/pnasnexus/pgae245. ↩
- Adeli, H., Minni, S., & Kriegeskorte, N. (2023). Predicting brain activity using transformers. *bioRxiv*. https://doi.org/10.1101/2023.08.02.551743. ↩
- Bhatia, S., & Richie, R. (2024). Transformer networks of human conceptual knowledge. *Psychological Review*, *131*(1), 271–306. https://doi.org/10.1037/rev0000319. ↩
- Bowers, J. S., Malhotra, G., Dujmović, M., Llera Montero, M., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep problems

↩

- Bzdok, D., Thieme, A., Levkovskyy, O., Wren, P., Ray, T., & Reddy, S. (2024). Data science opportunities of large language models for neuroscience and biomedicine. *Neuron*, *112*(5), 698–717. https://doi.org/10.1016/j.neuron.2024.01.016.

↩

- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1724–1734). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179.

↩

- Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Riquelme, C., Minderer, M., Puigcerver, J., Evci, U., … Houlsby, N. (2023). Scaling vision transformers to 22 billion parameters. https://doi.org/10.48550/arXiv.2302.05442

↩

- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*, 688-701. https://doi.org/10.1038/s44159-023-00241-5.

↩

- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1.

↩

- Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *2*(8), 451–452. https://doi.org/10.1038/s44159-023-00211-x.

↩

- Frank, M. C. (2023). Openly accessible LLMs can help us to understand human cognition. *Nature Human Behaviour*, *7*(11), 1825–1827. https://doi.org/10.1038/s41562-023-01732-4.

↩

- Frank, M. C., & Goodman, N. D. (2025). Cognitive modeling using artificial intelligence. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/wv7mg_v1.

↩

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

↩

- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv*. https://doi.org/10.48550/arXiv.2001.08361.

↩

- Lampinen, A. K, Dasgupta, I., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, *3*(7), pgae233. https://doi.org/10.1093/pnasnexus/pgae233.

↩

- Lindsay, G. W. (2020). Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, *14*. https://doi.org/10.3389/fncom.2020.00029.

↩

- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011.

↩

- McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), e2322420121. https://doi.org/10.1073/pnas.2322420121.

↩

- Millière, R. (2024). Language models as models of language. *arXiv*. https://doi.org/10.48550/arXiv.2408.07144

↩

- Millière, R. (2024). Philosophy of cognitive science in the age of deep learning. *WIREs Cognitive Science*, *15*(5), e1684. https://doi.org/10.1002/wcs.1684.

↩

- Puebla, G., & Bowers, J. S. (2024). Visual reasoning in object-centric deep neural networks: A comparative cognition approach. *arXiv*. https://doi.org/10.48550/arXiv.2402.12675.

↩

- 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., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285–1295. https://doi.org/10.1038/s41562-024-01882-z.

  ↩

- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *arXiv*. https://doi.org/10.48550/arXiv.1409.3215

  ↩

- Tay, Y., Dehghani, M., Abnar, S., Chung, H. W., Fedus, W., Rao, J., Narang, S., Tran, V. Q., Yogatama, D., & Metzler, D. (2022). Scaling laws vs model architectures: How does inductive bias influence scaling? *arXiv*. https://doi.org/10.48550/arXiv.2207.10551.

  ↩

- Tuckute, G., Kanwisher, N., & Fedorenko, E. (2024). Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47, 277–301. https://doi.org/10.1146/annurev-neuro-120623-101142.

  ↩

- Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision? *arXiv*. https://doi.org/10.48550/arXiv.2105.07197.

  ↩

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 5998–6008). Curran Associates, Inc.

  ↩

- Warstadt, W., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In W. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, & R. Cotterell (Eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (pp. 1–6). Association for Computational Linguistics.

  ↩

- Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*, 1526-1541. https://doi.org/10.1038/s41562-023-01659-w.

↩