MAPS OF THE LABYRINTH

A SERIES FROM PHANTOM MAZE — AI & LANGUAGE LAB

ESSAY 15

Attention: The Bow and the Wait

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.23575278

What does "attention" mean in artificial intelligence?



Artemis of Stabiae, Villa Arianna, Castellammare di Stabia, 1st century AD

Artificial intelligence often speaks in words borrowed from mental life. It learns, remembers, reasons, hallucinates, and it also attends. "Attention Is All You Need", ran the title of the paper that introduced the *transformer* architecture in 2017, and ever since then the word that Simone Weil had placed at the centre of her ethics and her mysticism has named a mathematical operation that takes place at dizzying speed in the basements of the industry.

The term says a great deal. “Attention” comes from *ad* plus *tendere*, to stretch towards something or someone. To attend is to aim the mind as one aims a bow, with concentration and direction. French introduced a detour, and *attendre* came to mean to wait, because stretching towards something is a way of awaiting it. In Spanish attention is lent, *prestar atención*, withdrawn from everything else but only for a while, and in English it is paid: *to pay attention*, as in *pay respects* or *pay homage*, a giving with a ceremonial tone that has gradually frayed. Between the bow and the wait, the new sense came to keep only one.

Artemis of Stabiae, Villa Arianna, Castellammare di Stabia, 1st century AD

But let us go further back. “Every one knows what attention is,” wrote William James in 1890, with exaggerated assurance. In *The Principles of Psychology*, the founding work of modern psychology, he defined it as the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects. To attend is to focus, to concentrate and, above all, to leave something out. Its essence, James said, lies in withdrawal from some things in order to deal effectively with others. In the James household attention was a great theme. William’s brother Henry, the great American novelist, became famous for restricting the point of view to a single central consciousness, so that the reader knows only what that character manages to perceive or to surmise, as in *What Maisie Knew* (1897) or *The Ambassadors* (1903).

The idea that to attend is to select, lighting one point of the field while dimming the rest, seeped into engineering. In 2014 Dzmitry Bahdanau’s group was trying to improve machine translation. The systems of the day attempted to compress the whole source sentence into a single representation and then to reconstruct the translation from that summary. The new mechanism allowed the model, as it produced each word, to return to the original sentence and give more weight to the parts that turned out to be pertinent. That uneven distribution of relevance they called “attention”. Unlike psychological attention, which removes elements from the field, the mathematical mechanism assigns degrees. An element counts a great deal, a little or almost nothing. Selection happens by gradation, not by exclusion. Three years later the *transformer* made this operation its organising principle, relating the positions of a sequence and calculating how much each one contributed to the representation of the others.

How does the mechanism work? Each token arrives represented by a vector, and from it the model computes three distinct versions by means of transformations learned during training. They are called *query*, *key* and *value*. We can picture them as a request, a label and a payload. The first represents what information is pertinent from a given position; the second makes it possible to assess what the others have to offer; the third contains what is actually taken in. Let us remember that there are only learned numerical projections here, with no intention and no search. The *query* of one position is compared mathematically with the *keys* of the others, and the greater the compatibility, the greater the weight its content will receive. An operation called *softmax* turns all those scores into proportions, and with them the values are combined. The result is a new representation in which the information from some positions contributes more than that from others.

The parameters that turn each vector into *query*, *key* and *value* were learned during training and are stored in the model; while we converse, they stay put. The "attention weights", by contrast, are computed on the spot for that particular context. They are stored nowhere, because they are computed afresh for every pair of positions and for every sentence, and they lie outside what is kept. The rules are relatively stable, but their results are contextual.

In generative models a further restriction appears. Each position may attend to what is already available, but not to what has not yet occurred. A mask blocks the positions that do not yet exist, which is why one speaks of "causal" or "masked" attention. This is the reason the model writes forward, committed to what it has already produced. If we follow the metaphor we began with, the model is drawn like a bow towards what is there, but it does not wait. A minimal example: in "after following the river she sat down on the bank", "bank" arrives with its ambiguity between a riverside and a financial institution, but also with information that makes it possible to resolve it. "River" and "sat down" turn out to be relevant, and their representations will contribute more than others. There is no label that says "riverside", nor are the function words always relegated, because at times they are crucial. After many operations of this kind, the representation of "bank" can shift towards a configuration compatible with that sense. This is one of the pieces of machinery that help to manufacture its contextual *embedding*.

The overall scheme of the block looks fairly simple. First, each position gathers information from the others through attention; then its representation passes through a small neural network that transforms it position by position. Look around, process locally and repeat in the next layer, with the residual connections and the layer normalisation helping to keep the journey stable. The *transformer* repeats this alternation many times. Now comes the twist that makes everything a little mythological. In each layer several "attention heads" work in parallel, each with its own transformations of queries, keys and values. This makes it possible to compute different relations over the same text at the same time. Some heads seem sensitive to specific positions, to syntactic dependencies or to relations of coreference; others are redundant or far harder to interpret, and in certain models a considerable proportion can be removed without performance collapsing. Among the most studied are the so-called "induction heads", which in certain models recognise sequences of the type A-B … A and favour the appearance of B afterwards. One should not imagine fixed functions, however.

The mythological echo comes from Argus Panoptes, the guardian Hera set to watch over Io. He had a hundred eyes, and they never all slept at once. The perfect sentinel. Knowing he could not beat him by force, Hermes played the syrinx to him while telling him stories, until the hundred eyes closed, one after another. Hera gathered up those eyes and set them in the peacock's tail, where they remain open without keeping watch. The transformer looks like an administrative Argus, with hundreds of eyes on the same text, computing different relations at once. No tale can put it to sleep, it is never distracted, though strictly speaking it is not looking either.

Attention maps are not an X-ray of reasoning. A head may assign great weight to a word, but that is not the reason it produces a given answer. Rather different distributions of attention can yield similar

results, and there are other components in the internal causality. In 2019 the argument broke out when a paper was published under the title "Attention Is Not Explanation", and another soon replied that neither can it be claimed that it is not: "Attention Is Not Not Explanation". With due caution, we can say that attention weights show part of the path the information travels, without any claim to transparency.

Let us return to Simone Weil to explore what it means that the sense of "waiting" does not enter the picture. This philosopher had spent a year as a factory worker, the last of the factories being Renault, in order to know from the inside what assembly-line work does to a person. She died at thirty-four in a sanatorium in Ashford, in 1943, leaving almost all her work unpublished, in scattered notebooks. Albert Camus took part in its publication at Gallimard and said she was one of the extraordinary minds of her time. For Weil attention in its purest form is akin to prayer, and love of one's neighbour has attention as its substance, the capacity to look at the one who suffers and ask them: "What are you going through?" and to remain before the answer. Weil's attention is neither muscular tension nor intellectual hunting. On the contrary, it means suspending thought, leaving it available, empty, penetrable by the object. Attention does not pounce on the truth; it creates the conditions for the truth to arrive. Full attention is a waiting without an anticipated object, without filling the void with oneself. That is why it is prayer, the self's complete renunciation of occupying the centre so that something else may occupy it, whether a problem in geometry, an afflicted neighbour or God.

The attention of the transformer, then, carries to its mathematical extreme the selection and weighting that William James saw. Without fatigue or lapses of concentration it always performs the same operation of distributing weights, with a precision beyond our own. Yet precision is not the same as pertinence. The mechanism can weight things dreadfully, ignoring what is decisive or leaning on spurious associations. For Weil attention corrects our expectations and obliges us to reconsider what matters. The transformer has no such relation to its object. If it goes wrong there is no frustration and no responsibility to put anything right.

Let us end with an irony of our age. In 1971 Herbert Simon argued that a wealth of information produces a scarcity of what information consumes, namely attention. What to give one's time to? At the cost of what? The more the supply of information grows, the more valuable becomes the human resource that selects from it. Half a century later we build machines whose central mechanism bears the name of "attention", and we use them to make the production of text, images, audio and code cheaper. Computational attention has become abundant and scalable, but human attention still has the same limited capacity. Technology sharpened Simon's scarcity by multiplying almost without limit what claims attention, without multiplying the attention of those who must grant it. We have built a creature extraordinarily skilled at drawing the bow and incapable of waiting, the attention of the archer without the archer, the hundred eyes of Argus in the peacock's tail. The machine's attention works, and it reminds us that *attente*, for now at least, remains our business.

James, William. "Attention." *The Principles of Psychology*, vol. 1, Henry Holt and Company, 1890, pp. 402–58; Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by

Jointly Learning to Align and Translate." *arXiv*, 1 Sept. 2014, arXiv:1409.0473; Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems 30*, edited by Isabelle Guyon et al., Curran Associates, 2017, pp. 5998–6008; Clark, Kevin, et al. "What Does BERT Look At? An Analysis of BERT's Attention." *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, edited by Tal Linzen et al., Association for Computational Linguistics, 2019, pp. 276–86; Michel, Paul, Omer Levy, and Graham Neubig. "Are Sixteen Heads Really Better than One?" *Advances in Neural Information Processing Systems 32*, edited by Hanna Wallach et al., Curran Associates, 2019, pp. 14014–24; Olsson, Catherine, et al. "In-Context Learning and Induction Heads." *arXiv*, 24 Sept. 2022, arXiv:2209.11895; Jain, Sarthak, and Byron C. Wallace. "Attention Is Not Explanation." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, Association for Computational Linguistics, 2019, pp. 3543–56; Wiegreffe, Sarah, and Yuval Pinter. "Attention Is Not Not Explanation." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, 2019, pp. 11–20; Pétrement, Simone. *Vida de Simone Weil*. Translated by Francisco Díez del Corral, Trotta, 1997; Weil, Simone. "Réflexions sur le bon usage des études scolaires en vue de l'Amour de Dieu." *Attente de Dieu*, introduction et notes de J.-M. Perrin, La Colombe–Éditions du Vieux Colombier, 1950, pp. 113–23; Weil, Simone. "Formes de l'Amour implicite de Dieu." *Attente de Dieu*, introduction et notes de J.-M. Perrin, La Colombe–Éditions du Vieux Colombier, 1950, pp. 142–211; Weil, Simone. *La Pesanteur et la grâce*. Plon, 1947; Simon, Herbert A. "Designing Organizations for an Information-Rich World." *Computers, Communications, and the Public Interest*, edited by Martin Greenberger, Johns Hopkins Press, 1971, pp. 37–72.

HOW TO CITE

Marsico, Claudia, and Hernán Inverso. 2026. "Attention: The Bow and the Wait." *Maps of the Labyrinth*, essay 15. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.23575278

Maps of the Labyrinth is a series from Phantom Maze, AI & Language Lab. The collection lives at **phantommaze.cloud**; new essays also appear at phantommazelab.substack.com. Subscription is free.

Some of the framing ideas in this piece are developed in *The Egoless Phantom: Mapping the AI Labyrinth*, by Hernán Inverso and Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-16-4).

PHANTOM MAZE

Read this essay online · All essays · The book