MAPS OF THE LABYRINTH

A SERIES FROM PHANTOM MAZE — AI & LANGUAGE LAB

ESSAY 18

Transformer: A Handful of Pages That Changed the Shape of Everything

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.23575288

The architecture that made generative AI possible



Wenzel Hablik, Starry Sky, Attempt (1909)

In 2017, eight Google employees published a paper of a few pages. They gave it the title of a song: "Attention Is All You Need" winks at the Beatles' "All You Need Is Love". The internal design document was called "Transformers: Iterative Self-Attention and Processing for Various Tasks", and it included an illustration with six action figures from the toy franchise. As one might easily guess, the team was known as the Transformer Team. The name of the architecture stuck, so the story goes,

because for Jakob Uszkoreit, one of the eight, it stirred childhood memories and sat well with the idea of transforming information. In the paper they noted that the order of the authors was random. Today that casual, playful paper has gathered more than two hundred and fifty thousand citations, ranks among the ten most-cited works of the twenty-first century in any discipline, and the transformer's initial is on everyone's lips through the T in GPT. All eight, incidentally, have left Google, almost all of them to found companies of their own. We are going to explore the history of that architecture and of the idea that made it so successful.

Even in the world of Megatron and Optimus Prime, "transformer" carries the inheritance of Latin and its *transformare* (*trans*, "across, beyond, from one side to the other", and *formare*, "to give shape", from *forma*). To transform is to take something out of one figure and set it in another. Ovid described it two thousand years ago in the opening two lines of the *Metamorphoses*: *In nova fert animus mutatas dicere formas / corpora*, "my spirit moves me to sing of forms changed into new bodies". Then come two hundred and fifty stories of bodies that change shape, Daphne into a laurel, Arachne into a spider, Narcissus into a flower, strung on the thesis that nothing perishes and everything is transformed, that is to say, that form travels from body to body.

Curiously, the name of the architecture turned out to be exact even where its authors were not aiming at it, because a transformer takes a sequence of vectors, the numerical representations of the tokens, those points in the space of meaning by which each word is turned into a position, and transforms it, layer after layer, into another sequence of representations. In its generative descendants, those transformations end up serving to compute which token may come next. Like an Ovidian machine, it changes forms.

To understand what changed in 2017 one has to look at what was done before. Recurrent networks processed a sequence step by step. Each position produced an internal state that depended on the previous one. LSTMs, developed to preserve information better over long distances, greatly improved that memory. Before the transformer, attention had already appeared, allowing the decoder to consult different parts of the input sentence directly. Google's neural translator of 2016 combined precisely those two solutions, i.e., deep LSTM networks and an attention mechanism. There remained, however, a bottleneck that was hard to remove. Recurrence forced the calculations to be done in order. To reach position one hundred, position ninety-nine had to have been computed already. That was especially awkward for GPUs, designed to carry out enormous numbers of simultaneous operations and here left standing in the queue. The transformer made the radical bet of keeping attention and dropping recurrence.

The transformer's insolent idea was, in effect, to do away with sequential reading altogether. No more summaries travelling from word to word. All the words in the sentence look at one another at once, directly, each deciding which others matter to it in order to sharpen its own meaning. In the generative decoder there is one restriction. Each position may consult what came before, but not look at what has not yet happened. That does not prevent training from being parallelized, but it does mean that, when the model generates, the output appears token by token. This mechanism is called

attention. Let us say here only that instead of a reader moving along the line with a finger there is now a round table at which all the terms of the sentence consult one another simultaneously. Jakob Uszkoreit had been suspecting that this alone was enough and that recurrence was a crutch one could throw away. Hence the title "Attention is all you need". The conjecture ran so contrary to the common sense of the field that even his own father, Hans Uszkoreit, a well-known computational linguist, was sceptical. Perhaps Jakob needed some attention too.

The transformer turned reading, which had been a queue, into an assembly ready to exploit the GPU's army of parallel calculators. The original model was trained on eight cards for three and a half days, a pittance by the standards that were to come. It was enough to make history. Of course, not everything is rosy. If every word consults every other, the cost grows with the square of the length of the text, so that doubling the context quadruples the bill, but let us concentrate here on how it works and not on the costs.

The trick of splitting a sequential job so that it can be done simultaneously has a striking medieval precedent. In the thirteenth century, around the universities of Bologna and Paris, the demand for books exploded. Thousands of students needed the same texts of law, theology and philosophy, but a copyist took months to reproduce a single volume, because copying was work done one word at a time, just like reading for a recurrent network. The solution was the *pecia* system, "piece" in Latin. The stationer kept an official exemplar of the text, checked and approved by the university, but divided into numbered gatherings of a few leaves. Each copyist took one piece per week, and so twenty copyists could be working on the same book at once. In that way the complete codex came out in weeks, by moving from the queue to simultaneous attention. It looks a good deal like what the transformer did to the machine's reading. The recurrent network was a solitary copyist chained to the order of the text, while the transformer is the pecia workshop with the text divided into pieces that are all processed at once, with numbering systems in the gatherings and position markers so as not to lose the order, because once the work is done in parallel nobody knows any longer what came before what. In both cases there was great demand and an artisanal method that did not scale.

Yet we must go beyond copyists and networks to illustrate the most interesting step. In Jorge Luis Borges's "The Aleph", the narrator goes down into the cellar of a house on Garay Street and finds a small sphere, two or three centimetres across, in which all the points of the universe are present at once and without overlapping. The problem soon becomes one of writing. He saw simultaneously what he has to write successively, because language has that structure. The transformer lives in part that same disproportion, since it looks at the whole sentence at once, with all the words consulting one another at the same time, and then has to deliver the result the way language delivers it, one word after another. The asymmetry is constitutive, and the privilege of seeing everything together becomes a headache when the time comes to tell it.

Wenzel Hablik, Starry Sky, Attempt (1909)

What followed was a complete colonization. The transformer of 2017 had been born to translate and so it had two parts: an encoder, which received the source sentence and built a representation of it,

and a decoder, which used that representation to produce the translation step by step. Its descendants soon separated the two halves and discovered that they could live very well on their own. BERT, presented by Google in 2018, kept essentially the encoder side. It reads the context in both directions and builds representations that are especially useful for understanding, classifying and searching. The following year Google built it into its search engine to interpret certain queries better. GPT took the complementary path by keeping the decoder and turning it into a machine trained to continue a sequence, predicting a token from the ones before. The generative possibility emerged with Łukasz Kaiser and Noam Shazeer, two of the authors of the original transformer, who shortly afterwards took part in a decoder-only system capable of producing even complete Wikipedia articles from reference documents. The GPT line turned that idea into a programme of ever-growing scale. Its name still preserves the full genealogy: *Generative Pre-trained Transformer*, with the architecture almost as a surname. But it would be a mistake to think the inheritance stayed locked inside language. Soon there were transformers that treated an image as a sequence of patches, others that processed speech and music, and then video, robotics and multimodal systems. An architecture conceived for a fairly concrete task, translating from one language into another, ended up turned into a general mechanism for processing almost anything that can be represented as a sequence of related parts.

It won because it allowed itself to be trained in parallel and made it possible to relate distant positions in a sequence. In the age of massive computation, the architecture that spreads itself best across a thousand calculators takes everything. It is an uncomfortable lesson, and an oft-repeated one. A substantial part of AI is explained more by the cost of scaling than by any abstract notion of intelligence. The machine with a toy's name, the one that changes forms layer after layer like Ovid's poem, ended up starring in the metamorphosis of the entire field. So far it has itself scarcely changed. Today's models have changed many pieces around that architecture, but the fundamental combination of self-attention, feed-forward layers and residual connections is still recognizable. There are pretenders to the throne who promise to dodge its famous quadratic toll. Meanwhile, it is worth remembering that eight pages from 2017 decided the shape of almost everything the machine says, and that forms, as Ovid taught, are only passing through bodies.

HOW TO CITE

Marsico, Claudia, and Hernán Inverso. 2026. "Transformer: A Handful of Pages That Changed the Shape of Everything." *Maps of the Labyrinth*, essay 18. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.23575288

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).

Sources. The founding paper: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin, "Attention Is All You Need" (NeurIPS, 2017), with the note on the random order of the authors and the training details (8 P100 GPUs, 3.5 days for the large model). The colourful details about the Beatles title, the name chosen by Uszkoreit "because he liked the way it sounded", the design document "Transformers: Iterative Self-Attention and Processing for Various Tasks" with the six action figures from the franchise, and Hans Uszkoreit's scepticism, in Steven Levy, "8 Google Employees Invented Modern AI. Here's the Inside Story" (*Wired*, 2024), and in Madhumita Murgia, "Transformers: The Google Scientists Who Pioneered an AI Revolution" (*Financial Times*, 2023), which also documents the departure of the eight authors from Google. The paper's place among the most-cited of the century, in Helen Pearson et al., "Exclusive: The Most-Cited Papers of the Twenty-First Century" (*Nature*, 2025). Ovid, *Metamorphoses*, I, 1-2, and the doctrine of universal transformation in the speech of Pythagoras in book XV. Jorge Luis Borges, "The Aleph", published in *Sur* in 1945 and collected in *The Aleph* (1949): the iridescent sphere of two or three centimetres in the cellar of the house on Garay Street, and the narrator's despair at the simultaneity of what he saw against the succession that language imposes on him. Google's neural translator of 2016: Yonghui Wu et al., "Google's Neural Machine Translation System" (2016). On long short-term memory: Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory" (*Neural Computation*, 1997). On the pecia system: Jean Destrez, *La pecia dans les manuscrits universitaires du XIIIe et du XIVe siècle* (1935), the classic study that identified exemplars and pieces from Bologna, Paris, Oxford and Naples; a summary in the entry "Pecia system" of the *Oxford Companion to the Book* (Oxford University Press, 2010). On the descendants: Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers" (2018); Alec Radford et al., "Improving Language Understanding by Generative Pre-Training" (2018), the first GPT; Alexey Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (2021), the vision transformer.