MAPS OF THE LABYRINTH

A SERIES FROM PHANTOM MAZE — AI & LANGUAGE LAB

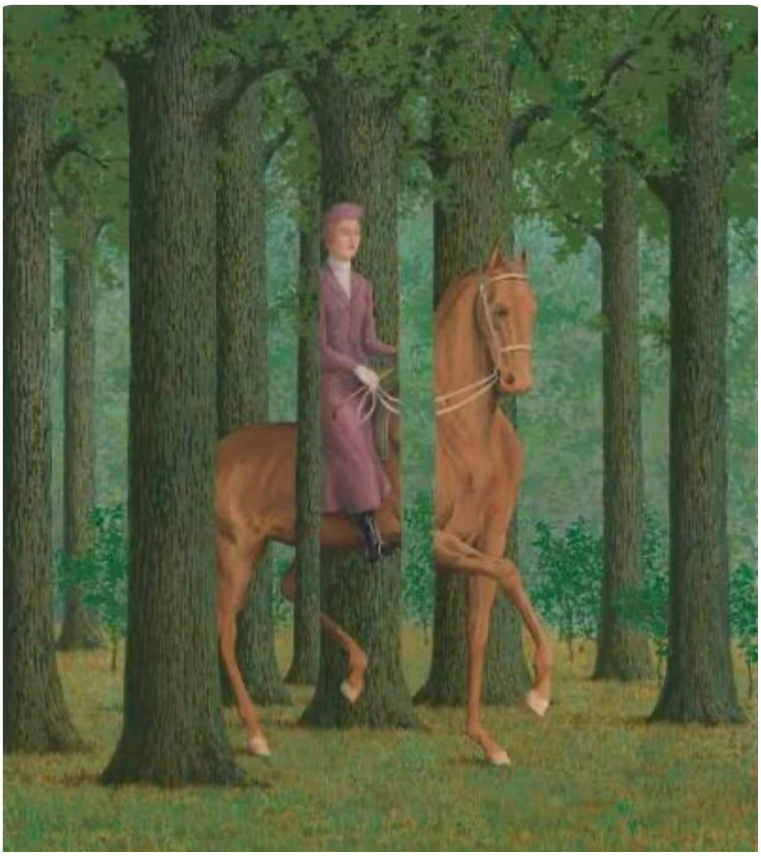
ESSAY 19

The Extraordinary Bag of Tristan Tzara

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.23575221

Positional encoding in AI models, or how far the order of words matters



R. Magritte, Le Blanc-Seing (1965)

In 1920, at the height of the Dada scandal, Tristan Tzara published, first in *Littérature* and some months later in the collection *Manifesto on Feeble Love and Bitter Love*, a recipe for making poems:

Take a newspaper.
Take some scissors.
Choose from this paper an article the length you want to make your poem.
Cut out the article.

Next carefully cut out each of the words that make up this article and put them all in a bag.
Shake gently.
Next take out each cutting one after the other.
Copy conscientiously
in the order in which they left the bag.
The poem will resemble you.
And there you are, an infinitely original author of charming sensibility, even though unappreciated by the vulgar herd.

This anti-art mockery of literary solemnity brushes against a curious point. If we take the order away from the words, the meaning does not survive. Tzara's bag contains exactly the same words as the newspaper article and says nothing remotely like it. Let us now look at the transformer, the architecture behind every language model in use today. In its own way it is a reader of Tzara's bags, since its design does not let it see the order. When it processes tokens, the fragments it works with, all the terms of the sentence are handled at once, each consulting the others, and therefore without any order. If we feed the machine "the dog bit the postman" and "the postman bit the dog", the same words in a different order, its internal arithmetic comes out exactly the same. Mathematicians say, with some elegance, that attention is permutation-invariant. Tzara might have said that the model reads whatever comes out of the bag, so that it needs someone to give the order back to it, because it has no ambition to be a Dadaist. How does one teach order to something that has none?

R. Magritte, Le Blanc-Seing (1965)

To put things in "order" was originally to arrange the threads of a weave, to set up the warp, which makes the word a close relative of "text" and its reference to weaving. But "order" leapt from the loom to the whole world, organizing it as if it were a cloth woven out of religious orders and orders of chivalry, the ordination of priests, ordinal numbers, orders of the day and, in French, the *ordinateur*, the computer. The order of words is the frame on which the text is stretched, so that Tzara's bag unweaves without cutting a single thread.

Strictly speaking, there are languages that lived with the bag half open. Classical Latin and Greek had a very free word order. The *Aeneid* opens with *arma virumque cano*, "Arms and the man I sing". No Roman got lost. Poetry pushed this to its limit in hyperbaton, the figure that dislocates the expected order. The record for glorious excess belongs to Góngora, who could write in the *Soledades* "pasos de un peregrino son errante", something like "steps of a pilgrim are wandering", instead of "son pasos de un peregrino errante", "they are the steps of a wandering pilgrim", scattering the phrase across the line and trusting that the reader would be able to put it back together. "Patience you must have", "much to learn you still have", says Master Yoda in the *Star Wars* saga, in pure hyperbaton.

The more the meaning depends on morphology rather than on position, the more the order can be altered. Latin declines, that is, it gives each word an ending that says what role it plays in the sentence. *Canis virum mordet* and *virum canis mordet* say exactly the same thing (the dog bites the man), because "canis" wears the mark of the subject and "virum" that of the object, wherever each of them

goes. Modern languages, from English to Chinese, chose to weaken or abolish their declensions and surrender to order, and that is why Tzara's bag is lethal for a newspaper of today and would have been almost harmless to Virgil, since every Latin cutting comes out of the bag with its function tattooed on it.

Seen from philology, what the authors of the transformer did in 2017 is astonishingly classical: they turned a language of order into a language of declension. Since their machine could not read in a line, they stamped on each word, in the very body of the word, the mark of its position. Before entering the game, each token arrives converted into a vector, that list of numbers which is its point in the space of meaning. Positional encoding adds to that vector another list of numbers that encodes "I am word number seven", and so the token acquires the mark of its position. The disorder of Tzara's bag no longer matters, because the order is inside each word and that thread is never going to tangle again.

Each position receives its signature made of waves, a collection of sines and cosines of different frequencies, some oscillating fast and others extremely slowly. Imagine a clock with a great many hands, none of which tells the time on its own, but which together identify the instant without ambiguity. Two nearby instants have similar pictures, two distant ones very different pictures, so that the distance between words is also hinted at. The positional signature is added, number by number, to the vector of meaning, so that position and sense end up mixed in the same coordinates, as if the accusative were not attached to the word but dissolved in it. That so simple a sum should be enough to preserve identity and position at the same time was one of those findings in which engineering first showed that the mechanism worked and theory then had to explain why it worked so well.

The copying of books had already faced a similar problem, which we reviewed in connection with the transformer architecture and simultaneous attention, and it is significant here too because of the way one part refers to another. A medieval codex was copied in loose quires, and those independent gatherings then reached the binder. Since he had not read the text and had to sew them in the right order, marks of order were stamped on each piece by means of *signatures*, letters and numbers at the foot of the quires, and the *catchword*, which gave the first word of the following quire. Positional encoding is the token's catchword, that is, the mark that allows blind, simultaneous processing to reconstruct a sequence it never read as such.

The signature of waves of 2017 was only the first way of solving the problem. Shortly afterwards, models such as GPT-2 and BERT began to learn position as well. Instead of receiving a mathematical signature already designed for "position 1", "position 2" or "position 3", the model learns a different vector for each place in the sequence. That gives it more freedom, with the drawback that it knows the positions it was trained on well and has no idea what to do when the text goes beyond them. What does a binder do who learned to number up to a certain number of pages and suddenly receives a far longer volume?

In 2018 came the idea of focusing on the relation between words rather than on their absolute number. What matters is not so much that "dog" sits in position seven as that it comes three places after "the" or two before "barks". The network could learn how far some words were from others.

Strictly speaking, grammar is more a matter of neighbourhoods and distances than of fixed numbers. Out of that idea came RoPE, "Rotary Position Embedding", proposed in 2021, where position is no longer added as a second numerical label but is built in by rotating the representation of each word according to the place it occupies. The word is still represented by a set of numbers, but now rotated in a slightly different way depending on whether it appears first, fifth or hundredth. The difference between those rotations makes it possible to recover information about how far certain words are from one another, as if each word carried a clock hand that tells where it stands with respect to the rest.

An even more austere solution was ALiBi, presented in 2022, which, instead of stamping the position onto the words, acts directly on the attention between them: the farther apart two words are, the more their relation is penalized, so that, other things being equal, nearby words keep more influence than remote ones without the need to learn a specific mark for each position. The simplicity of the procedure had a considerable advantage besides, summed up in the title of the paper, "Train Short, Test Long": a model trained on relatively short sequences could later be extended to longer texts without losing itself in a system of positions it had never learned. This detail ended up becoming a battlefield. The way in which the model represents position determines how far it can follow the thread of a text before running out of useful references. On this depends, in good measure, the "context window" that companies announce as if it were the engine displacement of a car, and which grows with every improvement.

Tzara's recipe ended with an ironic promise: after cutting out the words, putting them in a bag and drawing them out at random, "the poem will resemble you". The language model produces a similar feeling. How many people feel that the model knows them? It is trained to answer in ways we find appealing, and the better it does so, the easier it is to mistake that adaptation for resemblance. The answer can take on our style and anticipate our expectations, and beneath that familiarity there is still Tzara's bag with its discrete pieces, positions, probabilities and numerical relations. The resemblance of the result should not make us forget the differences.

HOW TO CITE

Marsico, Claudia, and Hernán Inverso. 2026. "The Extraordinary Bag of Tristan Tzara." *Maps of the Labyrinth*, essay 19. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.23575221

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. Ashish Vaswani et al., “Attention Is All You Need” (NeurIPS, 2017), section 3.5; Jacob Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers” (2018); Alec Radford et al., “Language Models are Unsupervised Multitask Learners” (2019); Peter Shaw, Jakob Uszkoreit and Ashish Vaswani, “Self-Attention with Relative Position Representations” (NAACL, 2018); Jianlin Su et al., “RoFormer: Enhanced Transformer with Rotary Position Embedding” (2021); Ofir Press, Noah A. Smith and Mike Lewis, “Train Short, Test Long: Attention with Linear Biases Enables Input Length Generalization” (ICLR, 2022).