MAPS OF THE LABYRINTH
A SERIES FROM PHANTOM MAZE — AI & LANGUAGE LAB

ESSAY 14

Pre-training: Betty Shannon's After-Dinner Game

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.23575274

Why does it matter so much to guess what comes next?



F. Porter, The Living Room (1952)

On January 23, 1913, before the Imperial Academy of Sciences in Saint Petersburg, a fifty-six-year-old mathematician, famous for his rigour and his bad temper, presented the results of an extravagant experiment. Andrei Markov had taken the first twenty thousand letters of Pushkin's *Eugene Onegin*, stripped out the spaces and the punctuation and arranged them in two hundred grids of ten by ten. Then, counting by hand, he determined how many were vowels and how many consonants, and how many times a vowel followed a vowel, a consonant a consonant, or the two alternated. Behind that

labour lay a quarrel that was a good deal less literary. The mathematician Pavel Nekrasov had maintained that the law of large numbers, according to which frequencies tend to stabilise as the number of observations grows enormously, required independent events, and had gone so far as to link that independence with free will. Markov had spent years insisting that he was wrong. A sequence of events that depend on one another can also show stable statistical regularities. Pushkin offered him a perfect laboratory, since each letter changes the probability of what comes next and yet, over thousands of letters, the frequencies settle. Markov chains were born in that climate. Today's artificial intelligence descends from those grids scribbled in Saint Petersburg, because the task in the pre-training of models, simple and mysterious at once, is to guess the word that comes next.

The vocabulary of anticipation can help orient us. *To predict*, pre-dict, is in general "to say before". It may involve "divining" (*divinare*), which is to do what a god does, arrogating to oneself the knowledge of what is to come. Divination comes in inspired versions and in others that are almost technical. The Roman augur who traced a rectangle in the sky with his staff and read the future in how the birds crossed it was using *conjecture*, "to throw together", like someone who gathers the signs to interpret what follows in that rather chaotic ensemble. When Cicero wrote his treatise against divination, he partly spared this variant, which confines itself to inferring frequencies from signals. A language model, too, "pre-dicts", without access to what is coming and without divine illumination, conjecturing from what has been thrown together, that is, the context and the corpus, to bet on the next symbol. The model is a statistical augur watching the words go by like birds.

Thirty-eight years after Markov, the scene was repeated in a house in New Jersey, this time as an after-dinner game. Claude Shannon, who had just founded information theory, wanted to measure how much redundancy English has, that is, how much of a text is, so to speak, already said by the text before it. His wife became the measuring instrument. Betty Shannon was given a passage from a novel, covered over, and had to guess it letter by letter. If she got it right, they moved on; if not, Claude told her the letter and they continued. By counting how many attempts were needed at each position, Shannon calculated that printed English carries about one bit of information per letter. That means that in an alphabet of twenty-seven symbols, the next letter is so conditioned by the ones before it that guessing it is far easier than it seems. The conclusion, published in 1951 under the title "Prediction and Entropy of Printed English", is the bridge between Pushkin and the chatbots. Predicting a text and compressing it are the same operation seen from two sides, and the rate of correct prediction makes it possible to measure how much is known about the language. Playing after dinner, Betty was carrying out by hand the work that data centres do today.

F. Porter, The Living Room (1952)

Shannon had also performed the reverse experiment. In his founding article of 1948 he manufactured series of "approximations to English", generated by choosing each symbol according to the statistics of the one before. First single letters at random, then pairs, then words chained by their frequencies. The series starts out as typographic soup and gradually falls into order. In the approximation by pairs of words, phrases appear such as "the head and in frontal attack on an english writer that the

character of this point is therefore another method", which say nothing and yet *sound* like English, like a badly tuned radio on which we can more or less recognise a language but not the message. In 1948 Shannon was generating text with a Markov chain. Between that badly tuned radio and the prose of a chatbot lie seventy years of improvement.

To train a machine, corrected examples are needed. Photos of dogs labelled "dog" by someone, X-rays annotated by doctors, translations validated by translators and texts that a native speaker confirms sound right. That human labelling is slow and costly. Next-word prediction, however, sidesteps the problem, because the text brings its own answer sheet. Take any written sentence, for instance "The cat climbed onto the roof", cover the last word, ask the model for its conjecture, and the correction lies in the covered word without anyone having to annotate anything. This procedure is called self-supervision, and it turns a corpus of trillions of words into a gigantic exercise book. In a sense, training, with its covering, conjecturing, comparing and adjusting the weights over and over, seems monotonous, but it is not quite so. To complete "the cat climbed onto the..." it is enough to apply statistics to neighbourhoods in order to choose among roof, tree and armchair, but completing "the capital of Argentina is..." involves a reference to the world, completing "one hundred and seventeen times three is..." requires calculation, and closing the final sentence of a detective novel, something like "the murderer is...", requires having followed alibis and motives, that is, having really read.

The seemingly humble task of predicting well the next word of any possible text would demand modelling what the texts talk about, including grammar, facts, regularities of reasoning, styles and a long et cetera, if the model is to get better at the guessing game. The opposing camp is sceptical. What if they are only stochastic parrots, that is, if they utter things at random? Combining pieces according to probabilities, after all, is not understanding. The discussion remains open because it touches on questions we have always argued about without consensus. What is understanding? Can it occur without a body, without a world and without communicative intention? We will deal with this later, but for now let us simply note the tension between prediction as a form of understanding and as an imitation whose fluency simulates that understanding. It is a very old problem, and in the meantime the model keeps improving its conjecture of the word to come.

When the model finishes training, prediction turns into generation. Instead of merely conjecturing the next word, it writes it and incorporates it as part of the context from which the following one is conjectured. Word by word, the prophecy feeds on itself. In technical jargon this is called autoregression, but it is not far from the games of the surrealist avant-garde. In 1925, in a house on the rue du Château in Paris, a colourful group that included Breton, Prévert and Tanguy wrote sentences in a chain. Each one wrote down his part seeing only the end of what came before, folded the paper and passed it on. The first sentence became the name of the game that rummaged in the collective unconscious: "The exquisite corpse shall drink the new wine". The same mechanism of continuing what has barely been seen entertains a solitary player who passes the paper to himself at enormous speed. If Tzara's bag resembles the machine without order, the exquisite corpse is the machine in operation, a kind of writing in which each step conjectures from what it manages to see, and yet out comes prose that makes sense and code that compiles. The model writes forward and

without a draft, conjecturing word by word, bound by the previous ones, like a chained improviser who begins a long sentence aloud and can only press on.

Among all the definitions of intelligence that philosophy and science have tried out, many speculate about solving problems, adapting, abstracting or creating, but none mentioned guessing the next word. That was at most a parlour game for a patient wife after dinner or for a stubborn mathematician, and it turned out to be the key. It was trainable, with free answers in infinite quantity and the possibility of absorbing everything else that lives in a text. It is very curious that Markov should have counted vowels to win a dispute over free will, and that a century later the grandchildren of his chains should discuss free will with anyone who cares to.

While engineering was discovering how much conjecture yields, an influential current in the neurosciences had been proposing that the human brain is, itself too and above all, a prediction machine: an organ that ceaselessly anticipates the next signal. The next syllable of the person we are talking to, the next step of the staircase, do they not involve learning the difference between what was expected and what arrived? The hypothesis of predictive processing, which Andy Clark summed up in a perfect title, *Surfing Uncertainty*, is debatable, but it invites us to think about what lies behind anticipation as a mechanism associated with intelligence.

HOW TO CITE

Marsico, Claudia, and Hernán Inverso. 2026. "Pre-training: Betty Shannon's After-Dinner Game." *Maps of the Labyrinth*, essay 14. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.23575274

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

Sources. Claude E. Shannon, "Prediction and Entropy of Printed English", *The Bell System Technical Journal* 30, no. 1 (1951), pp. 50–64; Claude E. Shannon, "A Mathematical Theory of Communication", *The Bell System Technical Journal* 27, no. 3 (1948), pp. 379–423, and 27, no. 4 (1948), pp. 623–656; Andy Clark, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford University Press, 2016; Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training", OpenAI, 2018; Yoshua Bengio, Réjean Ducharme, Pascal Vincent and Christian Jauvin, "A Neural Probabilistic Language Model", *Journal of Machine Learning Research* 3 (2003), pp. 1137–1155; Emily M. Bender, Timnit Gebru, Angelina McMillan-Major and Margaret Mitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, Association for Computing Machinery, 2021, pp. 610–623.