How did scientists discover the AI species?

By: Lior Messinger, Korra.ai, Founder and CEO

In this article, I'd like to tell you a story of a witness. When one starts a new company, one ventures into a new land. That was what I thought when I founded Korra.ai, a company devoted to understanding and applying large-scale AI models. But what I experienced was a new land that was moving under my feet, continuing to expand as I walked forward. In this article, I'd like to explain the string of discoveries that took place and brought us to where we are today: living and breathing and laughing alongside a new species that was created to serve us.

A few words of explanation are due. The first of them, about the title. One might ask why the term 'discover' is used. Isn't 'invent' more appropriate? Or even 'build'? In my view, the term 'discover' is better suited to the situation science is at right now. When people discovered 6,000 years ago that heating sand makes glass, they used a natural phenomenon without understanding it. What does it mean to "Understand"? It means that you can go at least one level down (to the theory of phase of matter in this case) and be able to make predictions. For example, predicting that cooling air would make it liquid. Well, they couldn't do that back then.

In much the same way, today scientists don't understand LLMs and cannot "explain" them. As we will see, they know how to use them, how to manipulate them, but, alas, a scientific theory is not on the horizon.

A second justification needs to be given to the use of the word "species". Are LLMs a new species? What is a species? One may say it's a species because it can hear, see, talk, and it just can't move (yet). There are several species that were created by evolutionary processes, each presents a subset of these traits. On the other hand, a species can reproduce. Here, it can't (yet!). So, I will leave the species-or-not decision to you, the smart Homo Sapiens reader.

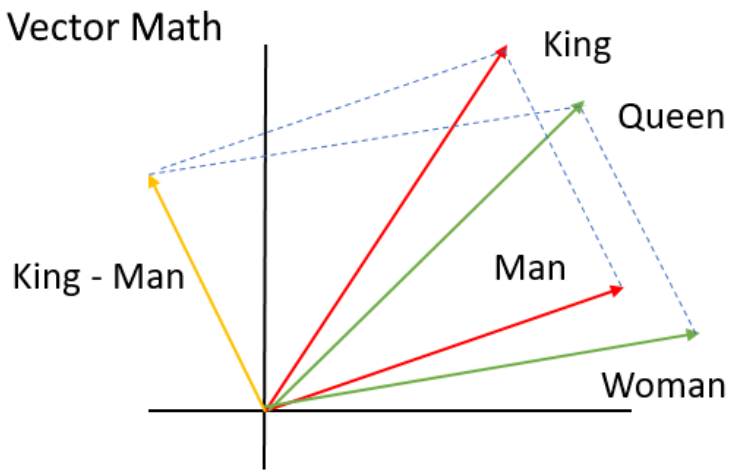
How did it all begin?

Restraint is needed here as we are not going to describe how it *all* began. We would just go back to the renewed rise of deep neural network, to the first noticeable milestone that happened in computer vision, in 2012, when deep learning was used to win the ImageNet competition by a large margin. The AlexNet paper, authored by now-famous Geoffrey Hinton and Ilya Sutskever, showed how to do that. Soon thereafter these deep learning concepts were borrowed into Natural Language Processing (NLP, does anyone still use this term?). Two seminal papers, published in 2013, were Word2Vec (led by a scientist called Mikolov, from Google) and a paper nicknamed "king – man + woman ≈ queen" , by the same Mikolov now joined by – yes, you guessed it, Ilya Sutskever. These teams invented a way to convert words into vectors, list of numbers. Numbers that represented, for the first time, *meaning.*

Meaning

What is meaning? When we are asked for something's meaning, we would describe it using other words that are semantically close to it. What is bread? Something you eat, that is made of wheat. What's a chair? Something you sit on.

So Meaning is proximity. So that if you draw the numbers that represent 'king', on a graph, they will be located near similar words such as 'queen'. And the distance would be the same as the distance between Man and Woman.



That was really the first foundation on which LLMs stand today. When you turn something into numbers, life becomes so much easier. You can use numerical models to try to predict these numbers, and more importantly, you can measure how well the model did. So, scientists began to work on different ways to improve these predictions.

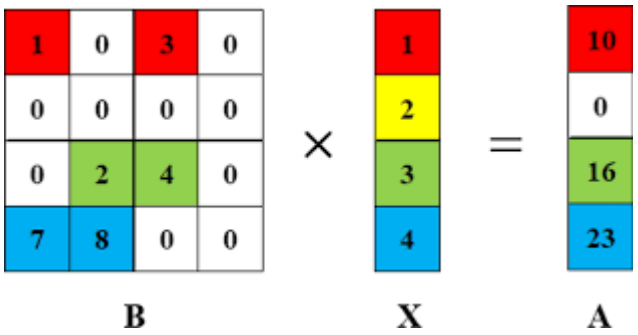
What is a 'model'? Imagine that 39 is the number for horse, 13 represents dog, and you need to solve the equation X - 13 = Y - 39. In other words, you want to find 'hare' and 'bitch', right? On the semantic side, our goal is really to create a model that can be used for the male-female analogies. We want to find such X and Y that would also work with other words like say, lion and lioness, or headmaster and headmistress.

But on the numeric level, how would you "solve" this equation? you'd just try some values, and see if they fit it. For example, 15 and 41 would do. Or 113 and 139. Or infinite amount of others. So what's the big deal?

That's what "training models" mean – find the X and Y that would work *best* for *most* words. So that, if presented to it waiter – waitress = steward – Z? it could predict that Z would be stewardess.

The training starts with random numbers put in X and Y, see how much the error was, and change those numbers in a certain efficient way until it hits as close as possible for all cases. The equations these models need to solve have many more Xs and Ys and results to reach, but the essence is the same.

By the way, the math is also not much more complicated than presented above. It is multiplication and additions – but ones that are made on vectors and matrices – lists of numbers.



In this example, we can see that there's a way to multiply a matrix by a vector, and it would produce another vector. To train mean to change the numbers inside the matrix, so it will fit the meaning of the sentence.

Training

The problem was that to measure the accuracy of a prediction, you need to know what's the right words. And to achieve good models, you need to have a lot of examples. But how do you get those examples?

One way that people used was simply to hire low-salaried people to create them. Amazon even created an app for that - called 'Mechanical Turk' where you could hire students and off-shore freelancers to create and curate those examples. But that only went so far – it's expensive, it takes time, and companies had to build and manage structured sub-organizations to produce quality examples. A breakthrough was badly needed.

NLP science chased its own tale for a few years, until in 2018, Ilya Sutskever co-foudned OpenAI with Musk, Brockman and Altman and took the first step towards what we have today. He used two new ideas. One, was a new architecture called Transformers which was introduced a couple of years earlier. Second, was a method to train models automatically. This method was unbelievably simple. All that is needed was to take an existing sentence from somewhere, give the model a few words, and train it to predict the next word. With no more humans in the loop, the main question left was about the textual material on which you can train models. Where do you find such vast amounts of text? Look no further than here. It's called the World Wide Web. The whole internet is open, public and available for us to use.

Ilya hypothesized that using Transformers, and maybe more importantly, next-word prediction to train models at scale could give language models *understanding*. He was right.

Understanding

In our day-to-day world, what does 'understanding' mean? If we see a woman stumbling and falling to the ground, we understand that she might be aching. If we see someone laugh, we understand he's happy. If we see a bird on the ground, we know it will soon take off. We have knowledge about the world, which was acquired during our lifetime of learning or by billions of years of evolution (some knowledge is innate - for example, interpreting face expressions).

This world knowledge is nicely demonstrated when we conversate. For example, if one would say 'I saw Ben at the class, he was laughing. I'm sure he feels _____ - you could predict the next word to be 'happy' or 'content' or some other positive feeling. You know that because you have knowledge about the world stored in your head somehow.

The ingenuity of Sutskever and friends was to turn this concept on its head. We humans were trained in the world, so we can predict the next word. What if we train a model to predict the next word – would it have world knowledge?

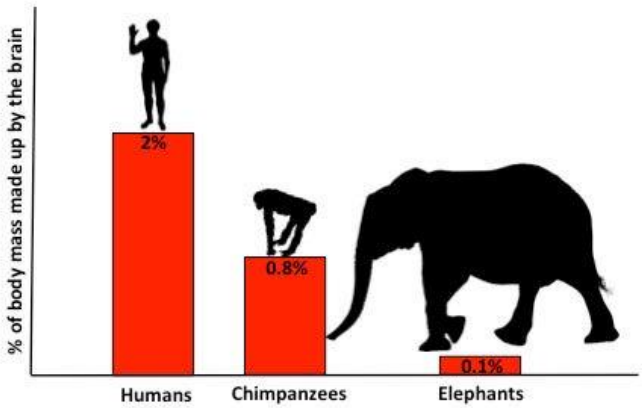
And so they started to train GPT-1. In 2018, around the same time I started to work on Korra, making my first steps in deep learning and totally unaware of the dramatic events around me - they published the paper called 'Improving Language Understanding by Generative Pre-Training'. See the nice usage of the word 'improving'. As it will turn out, that would be one of the biggest understatements in scientific history.

Supersize this brain

The goal was really to create an improved version of the then-top language models. A generic model that could be used as a foundation for various tasks, such as translation, classification, sentiment analysis and others. Interestingly, they didn't come up with new ingenious ways to architect models, no innovative algorithms or anything of that sort. Instead, they changed one parameter: the model size. Just like evolution showed itself that enlarging a creature's brain makes it smarter, they simply made the model bigger. How simple was it? Think about it as if you want to build a bigger building. You can try to invent new architecture, add towers and new structures. Or, you could just add more floors. That's what they did, adding more floors. And more. And many more. That's GPT-1.

What is each floor, you ask? Remember the matrix above? In essence, you add more matrices. And more. And…

You get the idea.



Indeed, performance improved. When measured, translation results came out closer to the correct values, sentiment was identified better, and word classifications were more accurate. The whole thing was not conversational. The tests were done as a completion of a sentence. If given a sentence such as “Good morning, it is still _____”, the model would assign probabilities to the words that could complete it (for example ‘Early’ – 40% ‘Cold’ – 30% ‘Late’ – 0.1% ‘Donkey’ – 0.001%). Or “I hated that film. The film is _____” (‘Bad’ - 45%? ‘Terrible’ – 30% ‘Good’ – 0.004%). Yes, it might be called *understanding,* but such a term still would be viewed as a marketing slogan, too. The model, so it was argued, simply read enough paragraphs to make statistical connections and no more.

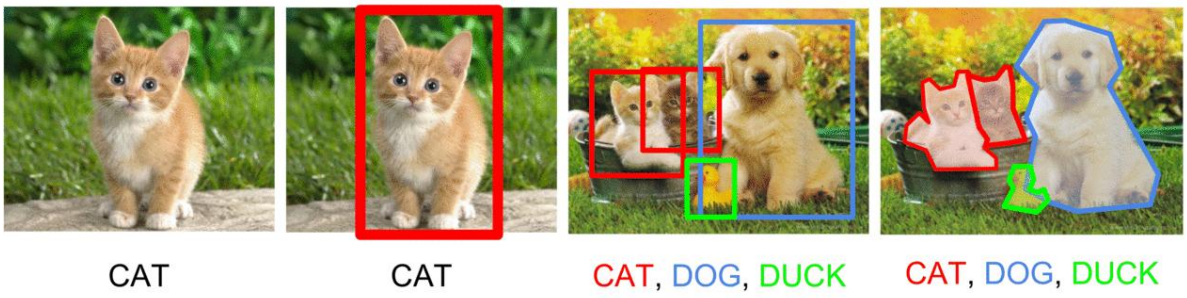
But there was a measurable improvement, and since the only change to prior models was the size, Ilya and friends just continued the trend. What if we made it bigger, they asked? And so they did, and saw that performance improved even more. They enlarged it further, and interestingly, performance kept improving. And again.

That was a very interesting discovery. One would expect that the improvement rate would flatten. The bigger you make a car motor, the faster the car goes, yes. But only up to a certain point. At some point, we should see the returns diminishing – because of different factors, like friction and metal strength. In deep learning models, the performance kept increasing linearly with model size, no end in sight. That was strange. When would improvement slow down?

Emergent Behaviors

As it will turn out, the only limit to increasing performance, and AI abilities, is the energy we have on this planet, that can drive big enough computers that can host ever-increasing model sizes. But at that time, they didn’t know it yet. What they did notice, however, was another strange phenomenon: the model started to display behaviors it wasn’t trained for.

When designing a model, the main question is not how well it does with the examples it was given, but how well it generalizes to new ones. For example, when a model was trained on regular images of cats, how well would it identify cats if we show it a partial picture, but still a picture with a cat in it?



The same generalization tests were done for LLMs. For example, one such test could be to complete the question "*If a remote is on the table, and the table is in the living room, and the living room is on the first floor, where is the remote?* _____". The model should complete it with ‘first floor’.

When tested on smaller models, they would reply ‘table’. Which is still impressive, because there is some story here, a question, and a model that understands it needs to answer – and it doesn’t answer with a random word, like ‘computer’. On the other hand, one could argue that this is not world knowledge. In some sense, the prediction is just parroting what was told to the model. In addition, there’s no alignment with what a person would expect. So this is not the intelligence we seek.

But in 2022, as model sizes increased, models started to output the right answers. They just emerged without any special training. That was complemented with answers to other challenges, too: solving multi-step math problems. Generalizing syntactic rules. The list goes on, and it all points in one direction: models that are intelligent. Models that show deeper conceptual understanding, that goes beyond the text they predict. Models that clearly know *concepts*.

These models answered questions they couldn’t have been trained on. For example, given the question “Where was Plato born?” a model would answer “Greece”, and one would argue that it could have seen it on some web page. But what if we ask, “What is the country code of the place where Plato was born”? Clearly there’s no page that answers this silly question. And models started to answer such questions. That’s called *reasoning*. Or better, *intelligence*.

Not only that these behaviors emerged by increasing size, but there was also another surprising pattern. When measuring model size against the intelligence it demonstrates, scientists saw a jump: up to a certain size, bad grades. Suddenly, a jump. No linearity here - and no explanation, too.

Fast-forward to our days, we can make a list of all those achievements that we see when interactive with ChatGPT, Claude, Llama or any other LLM: LLMs explain jokes, write rhyming poems, build applications and even cheer you up. They generate music, movies, and stories. They have human-like capabilities. And all that *emerged* just by increasing size! Just by adding more floors to the building.

Was it all planned? No.

So, is that a discovery? Yes.

Species Evolution

Truth to be told, it’s not a regular scientific discovery. Yes, it emerged spontaneously without prediction, and it wasn’t engineered with intent. But on the other hand, it’s man-made. So, it was both discovered and created – even if by chance.

Now that this species has been discovered and created, it begs the question: how will it evolve? To what extent will scientists be able to improve it, and will we see the performance curve flattens at some point? Furthermore, scientists use AI capabilities to accelerate scientific progress itself. So, AI improves itself, a clear evolutionary process. To what extent?

As you can see, many questions afloat and prediction is hard, especially about the future. To me, the main question is different. I’m really interested to know when all this will be explainable and predictable. When will we know how to build a new sense of humor, or add a new field of knowledge, just by changing some numbers inside the model? Will it take mankind 6000 years, like it took us to build a theory about melting sand? To be honest, I think it might.