AI and Its Impact on Cybersecurity

Part 2: A Detailed Look at AI

In the first part of our series, we looked at the history of Artificial Intelligence. AI has been around for quite a long time, at least the basics and foundations of what we know as AI today. We looked at how a paper by Google computer scientists was the springboard to the AI explosion we see going on right now. The introduction of the Transformer model, or the ability for AI to pay attention to and understand more completely larger datasets, now allows for faster response time and the inclusion of much more data in the library. This is where we begin, LLMs, or the Large Language Models that AI uses to provide answers to prompts.

What Exactly is AI

Now that we know where AI started and how it evolved, the next logical question is, what exactly is AI? In simple terms, AI is just pattern recognition. On major steroids! Remember in part one we discussed how a dataset of fruit might contain information on an apple and an orange? This is what we are talking about when we say pattern recognition. As a child you were shown photos of an apple and saw it was round and red. If you ate an apple, you knew it was sweet. Over time, after seeing photos of apples, seeing apples in person, eating apples, you then learned what an apple was. Now, not all apples are red nor are all apples sweet. You learned that a green apple was still an apple, but it was sour. You learned apples could be more of a yellow color. Over time, you learned what an apple was, that there are many varieties of apples and a variety of tastes. You began to see patterns emerge. If someone showed you a round red fruit, you could use your pattern recognition skills to say that it is a red apple and it is sweet. This is exactly how AI functions, using pattern recognition.

This pattern recognition took you many years, eating many apples, being introduced to a variety of apples and finally being able to make a pretty good guess as to what type of apple you were looking at. AI, with the faster Transformer model that pays “attention” to an entire description or data string about apples, can learn the differences between every variety of apple, very quickly. As a matter of fact, AI can look at patterns, data, comparisons and differences, compare every possible word and combination for all apples that it has ever seen……all at once!

Look at these comparisons and you’ll see why we say AI does pattern recognition on major steroids.

1. Scale: AI can analyze billions of examples where humans might see a couple hundred
2. Scope: AI can see patterns across thousands of variables at once, humans will max out around 3 or 4.
3. Speed: AI can learn in hours what it would take a human to learn in years.
4. Subtlety: AI can spot correlations so faint or complex that humans would never notice.

The Problem

OK, so this pattern recognition sounds awesome, so what's the catch? Well, it is not perfect. You have probably heard the word hallucination associated with AI models. This means the AI is making connections or seeing patterns that are not there, kind of like when your brain plays tricks on you.

Think about lying out on the beach, enjoying the sun, surf and sand. You are relaxing and watching the clouds go by with friends, trying to outdo each other by finding animals in the shapes of the clouds. "That one looks like a dragon!" "No way, that's totally a giraffe!" The clouds don't match the shape of the animals you are seeing; they just sort of look like them, and maybe you stretch the comparison a bit to get the win. This is similar to what an AI hallucination is. The model is making connections between patterns that are not as strong as it thinks they are, or it's filling in gaps with its best guess when it doesn't have enough clear information.

Here's the thing, remember that LLMs are predicting the most likely next word based on patterns they've seen. Sometimes those predictions lead them astray. Maybe the LLM has seen similar word combinations before, but in a different context. Or maybe it's trying so hard to give you a complete answer that it confidently presents information that sounds right but isn't actually accurate. The model is not lying, it's just doing what it was built to do, finding patterns and making predictions, even when those patterns are weak or misleading.

That is why all AI models come with disclaimers that the information returned may not be accurate. The LLM doesn't know facts the way you and I do. It's making educated guesses based on patterns, and sometimes those guesses are wrong. The better the patterns and the more data it has seen, the better its predictions will be, but there is always a chance it will hallucinate.

Where Does All This Data Come From?

We now know that AI is a super-fast machine that makes comparisons on a very large amount of data. Where does that data come from? The datasets that AI models use for their comparisons come from what are called LLMs, or Large Language Models. Let's look at what these words mean.

Large: This means the system has "read" billions, yes billions, of web pages, books, articles and conversations, a massive number of what humans have posted, written and copied to the internet. Large also means that it has "read" so much that it can see very subtle patterns in human language and how we use it. Think of all those literary terms you learned in English class. The LLM can see those patterns.

Language: The LLM works with words and sentences, with the way we communicate. It uses our language to find and see patterns of how words relate to each other.

Model: An LLM is not like this massive online dictionary and encyclopedia. It does not store all the data in a new location, then searches for answers. LLMs are a giant mathematical map that see how words relate and fit together with one another and then gives weight to each of those relationships. The more words relate to one another, the more weight that relationship carries.

Here is an example. If this sentence is fed into an LLM, "The capital of the United States is ___", the LLM has seen so many instances of the words "capital" "United States" and "Washington DC" relating to one another, that the answer to the question is "Washington DC". The relationship

between those three words is very high mathematically. It is predicting, with a high probability, that the next word in the sentence "The capital of the United States is Washington DC"

Importantly, here is what the LLM is not doing. As we said, the LLM is not storing all this information and looking it up as a fact, it is only looking at patterns. It is not learning like a human would that Washington DC is the capital of the US, it is not retaining facts. It does not even understand what Washington DC really is. And it is not thinking. It is not saying, "AAHH!! I know that one! Ooh, ooh, pick me"

What Does it Look Like Under the Hood?

Let's look at what happens in those few seconds from hitting enter after typing a prompt and the response from an LLM. The first thing the LLM does is "tokenize" each word you typed. Remember LLMs are really complex mathematical computations, so the LLM reads the words as numbers. LLMs are just computers and computers do not understand words, only numbers, so to be able to "understand" and pay "attention" to what you typed, the words need to be converted into numbers. Think of tokenization as translating from one language to another, except here its words into numbers.

Here is an example. When the LLM "reads" the phrase "I love pizza!", it will translate it to something that looks like this [319, 1595, 14403, 0]. Each word and punctuation symbol are translated or converted into a number. Now this is where the real work is done. Earlier when we were defining what LLM meant, we said it was like a huge mathematical map. Each word will have coordinates on that map, words that are closely related will be near one another. Now this map is not flat, this map is 3D and so there are many words that will reside near one another. The LLM learns these positions by seeing what words have appeared in similar contexts millions of times!

Now the secret sauce, the Transformer model! When you hit enter the LLM converts the words to numbers and then asks what words it should pay "attention" to in order to understand each word. So, when an LLM sees a sentence like "I threw the sandwich in the garbage can because I didn't like it", it will look at the word "it" and determine if that word is related to "I", "sandwich" or "garbage can". It then assigns an attention score or weight to finally determine that the word "it" relates to the word "sandwich"! The older RNN models would get the word "it" and be confused as to what "it" was referring to because it forgot most of the other words in the sentence.

These tokens, the numbers assigned to each word, do not just pass through the model once, they get passed through dozens or even hundreds of times going through layers or filters checking relationships, grammar, tone, etc. in order to finally using its mathematical formula to predict what the next word should be, or what the answer to a question is. Our explanation is very basic and there is so much going on in the background and yet these LLMs can return a response in mere seconds. You can see why the development of AI took advancements in processing chips to make what we have today possible.

Now that we have a better, basic understanding of how these LLMs work, we will move on to where they were first being used on a large scale with the general population. In part 3 we will look at email spam filters, search engines and algorithms. Until then, please consider going to our website, www.knowphishing.com and signing up for our free weekly newsletter. Each week our goal is to inform everyday users of technology about what to look out for, the most recent dangers, how

to make sure your technology is as safe as possible to prevent everyone from becoming a statistic in a world filled with phishing, malware and ransomware! Thank you again for your “attention”!!