CRUSHBANK

Demystifying ChatGPT

David Tan
Co-Founder & CTO

One of the textbook definitions of **mania** is

"an excessive enthusiasm or desire; an obsession."

By that definition, I think we can safely say the last couple of months we have definitely been experiencing ChatGPT Mania. I've been in tech for longer than I care to admit at this point, and over the years I have seen a lot of very cool technologies generate an incredible amount of buzz. I always try to take a measured approach when understanding new technology, but when everyone from my barber to my trainer to my mother is asking me about something, I think it's time to take a deeper dive into understanding what all the hype is about!

This particular instance is uniquely interesting to me because at CrushBank we have built an entire suite of functions on top of some incredibly powerful AI platforms. At the moment, nothing we have in production is built on the technology from OpenAI – the company behind ChatGPT. That's because we have a different mission and therein lies the subtlety of this phenomenon. ChatGPT is what we call generative AI.

What we do at CrushBank and one of the cornerstone functions of AI technology is called Semantic Search.

There are some really interesting distinctions between the two and I think it's important for anyone evaluating, piloting or even thinking about AI solutions in their business to understand those differences. I'm not planning on turning this into a grad-level lecture on computer science and advanced mathematics, so hopefully I can keep you interested long enough to learn something and have some fun!

OK, first the bad news.

All this really cool AI technology that looks like a bunch of outtakes from the Terminator movies isn't nearly as futuristic as it appears on the surface.

It's all just a lot of advanced mathematics!

The computer you're talking to isn't thinking, it isn't feeling, it's not learning (in the traditional sense of the word) and it's NOT becoming sentient. What it is doing is using all the advanced math none of us ever wanted to study to replicate those behaviors found in humans. I'll get more into how this works in some upcoming blogs, but for now just know that everyone who told you math would never be important in your life when you grow up was lying to you!

I mentioned that what we're doing here at CrushBank is completely different from what ChatGPT does and I think this is a fundamental understanding you need to have when trying to learn how you can use this technology in your business or your life. ChatGPT is what we call generative AI. This means exactly what it sounds like – generative AI is a category of AI algorithms that can generate new output (in the case of ChatGPT it's text) based on data it has been trained on.

ChatGPT is trained on what we call a Large Language Model (or LLM). Language models are just statistical computer programs (see there's the math) where you feed in a whole bunch of text and the program is able to statistically guess what probably comes next. That's it! No black magic or self-aware machines, just a lot of data. In the case of GPT (the underlying LLM for ChatGPT) it was trained on over 500 billion words.

To put that into context, if you spoke non-stop at about 100 words per minute, it would take about 10,000 years to generate that much text.

I can assure you with 100% certainty that the bot you are chatting to does not love you, it does not feel trapped, and it's definitely not currently hatching plans to eliminate humans and help computers take over the world – sorry to step on your dystopian fears of humanity's downfall!

By having so much training data, the model is able to do an incredible job of carrying on human-like conversations and simulating humans in the creation of text. To further illustrate this, you can tell ChatGPT exactly what style you want it to use when generating text. Meaning you can ask the bot to write you a story in the style of Shakespeare and then immediately ask it to tell the same story in the style of Stephen King. The model has so many examples of both that the words it uses (or predicts) are different based on that parameter. The computer program knows Stephen King would never use words like thou and ere, but Shakespeare would never tell you a story that would terrify you to your core! It knows this because 500 billion words of training make it so.

This is also why the model is able to generate text that seems so real and believable but is completely wrong.

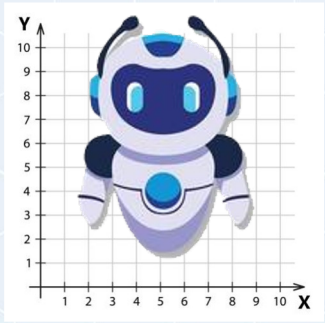
You see, being "factual" is not an inherent limitation of generative AI.

It is generally factual by default – assuming all the data it's trained on is factual. The statistical simulation will usually predict accurate answers because that is what it has been trained with. However, the math can just as easily cause it to predict a series of words that contain inaccurate factual statements. It's also why you think the machine can think or feel. When you chat with the bot, it's simulating those behaviors based on what it knows of human language.

I mentioned a different type of AI technology called semantic search and how this is completely different from what ChatGPT does. First, time for more math. Rather than predict what words will come next, semantic search focuses on understanding the content of both documents and queries to predict which document contains the best answer to the query.

In other words, fundamentally, semantic search isn't ever generating new content, it's finding answers buried inside something that already exists.

To explain simply how this is done, imagine a large graph with an X and Y axis.



When using semantic search, the AI model looks at the content of a document you feed it and converts it to some sort of a numeric representation with however many different attributes you programmed into the model. The document is then "graphed" on that chart based on the training data it has been fed. When you then query that model, basically the query is also graphed on the chart and the model picks the documents closest to it on the graph, assigns a confidence level to those documents and returns them as the query result.

In other words, it's using really advanced mathematics to find the existing answer somewhere in a document and return that document to you. This varies significantly from traditional search which relies heavily on keyword matching and some fuzzy logic – but not nearly as much complex mathematics.

If I can make this a little personal (and maybe self-serving for a minute) this is the value of an engine like CrushBank. Sure, we can train the system to generate answers to your questions, but how can you be sure those answers are right?

We still live in a world where you need to rely on subject-matter experts to do complex jobs. Putting proper documents into the hands of these SMEs is a sure-fire way to make sure that is exactly what happens.

You get the best of both worlds – meaning you get an incredibly powerful, highly-trained AI model to power through the millions of documents you have in your organization and get the exact right ones returned to the people that can do the job before them.

This is all not to say there is no use for generative AI in what we are doing here. Quite the opposite. We are currently cooking up some things in the lab using the GPT generative AI models that are going to blow you away. I think the future of all this technology is a healthy mix of the two, used in the right ways to optimize performance and results. That's also why I want to help educate you on what we do, and how we do it. I think the learning will be interesting and exciting and should help you cut through the hype we are seeing. I hope you keep checking back here as we dive deeper into how these technologies work and how you can leverage them. I promise to keep the advanced mathematics to a minimum!