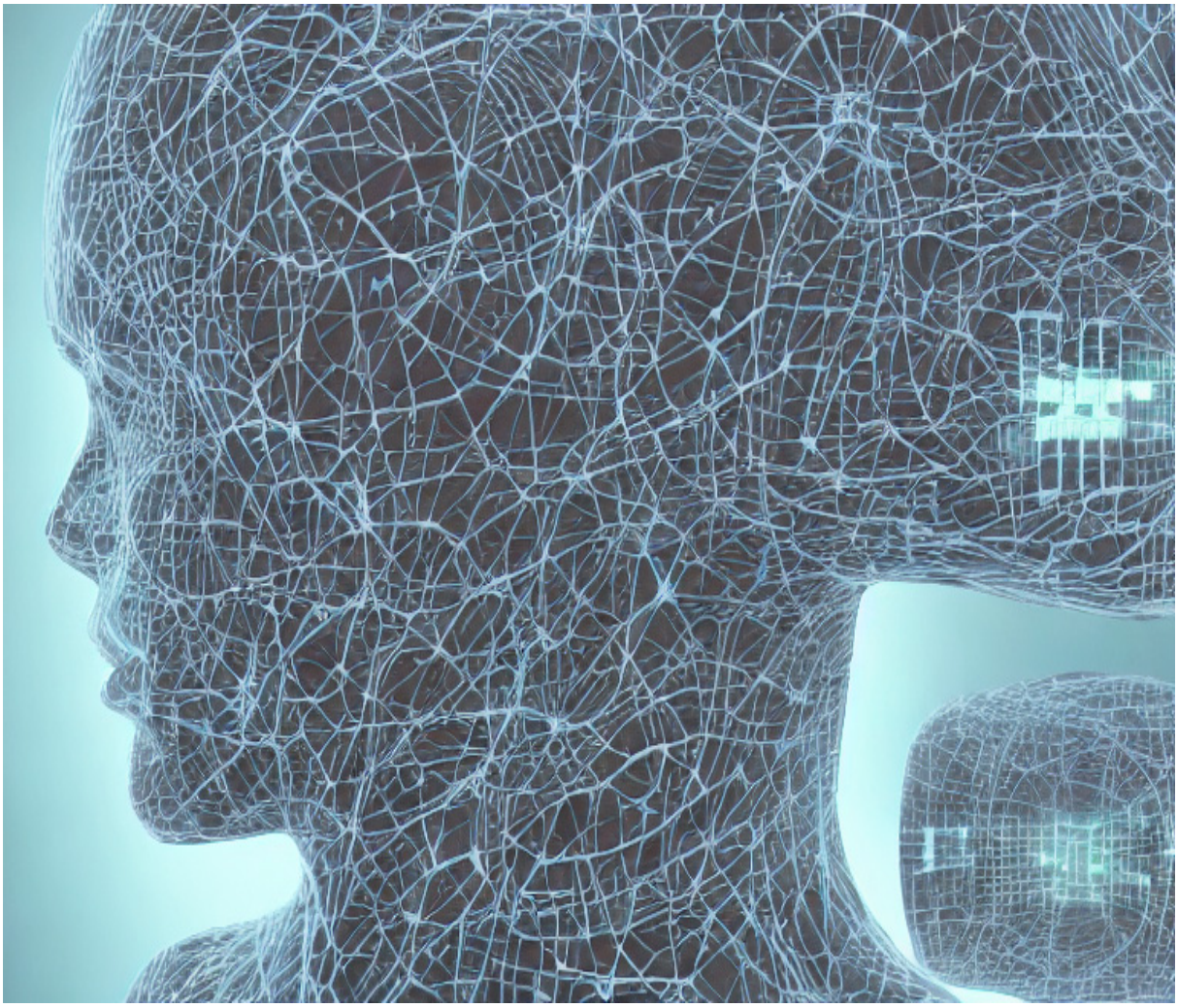
Generative AI: What Will Change in 2023



What is Generative AI? (Explained by Generative AI)

Instead of writing myself a bland, general introductory paragraph about what generative AI and ChatGPT are, I asked ChatGPT to explain that to me. Here it goes the ChatGPT explanation:

Generative AI refers to artificial intelligence systems that are able to generate new content that is similar in style or content to a given input. This can include tasks such as generating text, images, or audio. Generative AI systems are fascinating because they have the potential to create new and creative content that can be used in a variety of applications, such as generating news articles, creating art, or even generating realistic-sounding dialogue for virtual assistants.

One of the main milestones in the development of generative AI has been the development of large language models, such as GPT-3 (short for "Generative Pre-trained Transformer 3"). These models have been trained on vast amounts of data and are able to generate human-like text that is often difficult to distinguish from text written by a person.

ChatGPT is a variant of the GPT-3 language model that has been fine-tuned specifically for the task of generating chatbot responses. It has been trained on a large dataset of conversations, and is able to generate responses that are appropriate for a given conversation context. ChatGPT is an interesting development because it allows chatbots to have more natural and realistic conversations with users.

This is a bland, general introduction to Generative AI and ChatGPT, but you may find (as I do) that the fact that an AI system can do it at my request is close to magic!

What will come after the 2022 LLMs

The first thing to notice about ChatGPT is a distinction that will be the focus of this article:

A genuine commercial product is not the same as a testing platform.

ChatGPT is not a full-fledged product: it's a testing platform.

The difference between an actual product and a testing platform is that products have a well-defined use case and are associated with a way of making a profit, which I'll call a "business case." In contrast, a testing platform provides a general functionality for many potential applications.

ChatGPT doesn't have a specific use case; it can be used for school essays, marketing copywriting, the initial version of blog posts, computer programming, and much more.

ChatGPT is free (for now, at least) because OpenAI wants user feedback -by the way, that's a brilliant move for getting public awareness and, at the same time, debugging their software. ChatGPT is more of a "showcase" than a well-defined product.

And the main component needs to be added, which is the business case. OpenAI could already have a business case for a commercial version of ChatGPT, but to explain it, we need first to take a look at the application architecture Generative AI products will have.

A two-layer stack

Generative AI-based products will have a two-layer architecture:

1. At the bottom level, we find the Large Language Model (LLM), a humungous neural network, as explained at the beginning of this post.
2. At the top level is the application layer, which leverages the LLM to provide the intended value to the user.

ChatGPT has a minimal web interface for users, and it's up to them to build "interesting" prompts. In a commercial product, the LLM layer is accessed through an API (Application Programming Interface) that can be called from the application layer.

Products instead of toy platforms

What will change in 2023 is the development of products instead of testing platforms.

The products I'm talking about won't be offered by OpenAI but by smaller companies with a specialized customer need or by large companies like Microsoft that will enhance their already available products with additional features (think Word with automatic writing capabilities).

The two-layer architecture makes sense for both Ope-

nAI and also for developers leveraging the underlying Generative AI technology: for small developers, all they have to do is to find a compelling use case, build an interface on top of the Generative AI engine, and start selling their products (eventually they'll have to provide additional training and configuration).

Small companies won't have to deal with the astronomical costs of building and testing LLM (in the millions of dollars); this is left to OpenAI and its competitors. Application builders can focus on providing value to their customers.

An example: EnglishVoice

There are infinite possibilities for applying Generative AI to provide value to customers. I have seen many examples, ranging from proposing questions for a guest interview to writing commercial copy on webpages, dating help; you name it. But having many potential applications means, in practice, that there is no intended application or use case at all.

An actual use-case for Generative AI would be, for instance, an English-learning voice tutor that proposes an English lesson to cover and asks questions to the human user, providing feedback accordingly. Let's call it "EnglishVoice." The architecture would include the components:

1. The underlying LLM, like GTP-3, is configured to behave as an English tutor, accessed using API calls.
2. The voice recognition and voice synthesis software packages (Whisper, etc.) and the module for building the API calls to GTP-3 and keeping track of the user progress and challenges.

EnglishVoice wouldn't do "whatever the user comes up with"; it would just give English lessons to people learning to speak it. That is the use case.

But even more importantly, an actual application needs a business case.

The need for a business case

We are used to getting free products, but if you take the point of view of the company offering the service, they have to make money somehow. Google makes (a lot of) money with advertising (I learned from an insider that when Google started public operations, they didn't have a clue about how to make money).

What would be the business model for EnglishVoice, our hypothetical company? It could take a subscription model, charging the final user for the service for some 30 dollars per month. Around 2,600 active subscribers would be needed to become a million-dollar annual revenue company. One of the main expenses would be to pay OpenAI for using their underlying GPT platform through API calls.

Closing thoughts

We'll see many new products coming out here and there leveraging Generative AI, both from small companies with a single product and big companies adding features to their line of products.

I think OpenAI wants, in the future, to charge application developers for the use of GPT (the underlying Generative AI engine). They won't charge the final user directly; their revenue will come as a commission from companies using their platform.

The Generative AI ecosystem is just starting and will be unstoppable once it takes full steam. Some folks think Generative AI is overhyped, but I agree with Cerebras' CEO Andrew Feldman that it's underhyped instead. True, Stable Diffusion, DALL-E, and ChatGPT are toys, but they are just a taste of what is to come soon -the year we are entering, more precisely.