Explorez la diversité de l’intelligence artificielle générative au-delà de ChatGPT

OpenAI’s GPT model is not the only option for generating content and information using AI. In response, here is a non-exhaustive list of large language models that are on par with ChatGPT, as well as frameworks and projects related to generative AI.

Today, many companies see artificial intelligence as the future, and many executives consider ChatGPT to be synonymous with AI. But OpenAI’s flagship project is far from being the only large language model, and for some software projects or domains, it may not even be the best choice. Competitors are appearing almost daily. It seems like everyone wants to build the next generation of AI tools that will either save or destroy the world, depending on who you ask.

Are some models better than others? Perhaps. They all have flaws, quirks, bugs, and weaknesses that become more apparent as they are used. Generative AI may seem amazing at first glance, but over time, its strange and unpredictable side starts to emerge.

Scientifically measuring the quality of generative AI responses is difficult due to the scope of the models and how they are used. A data scientist could input thousands, if not millions, of test questions and evaluate the responses, but the results will be limited if the test sets focus on only one type of question.

Consulting a resource like Hugging Face’s Open LLM Leaderboard is interesting but not necessarily accurate. While finding an accurate way to compare LLMs is challenging, it is at least easier to switch from one to another. Projects like OpenLLM or

FastChat, for example, simplify the wiring of various models despite their different APIs and interfaces.

Perhaps the biggest underlying question is the cost. While everyone is excited about the surge of interest and investments, building a large language model can take months, if not years. Teams first gather training data, then run it through expensive hardware that consumes electricity. Finally, they produce the model. The best way to monetize and support this work is an evolving question. Some companies experiment with free access to their findings, while others gladly rely on services with their own billing models. Open-source LLMs can be a true gift, but only if you are capable of handling the model deployment work and making it run.

In this regard, we have compiled an overview of 14 large language models that are not GPT. They may or may not fit the needs of a particular business project. The only way to know is by sending messages and carefully evaluating the results.

AgentGPT

One existing tool for gathering all the necessary code for an application is AgentGPT. It is designed to create agents that can be sent to perform tasks such as planning vacations or writing code for a specific type of game, hence the name “Agent.” AgentGPT is built using Next.js with the T3DotGG stack, LangChainAI, and GPT-3. It draws inspiration from Yohei Nakajima’s BabyAGI and SigGravitas’ AutoGPT. The source code for much of the technology stack is available under a GPL 3.0 license. A running version is also available as a service. AgentGPT is currently in beta.

Alpaca

Several researchers from Stanford took Meta’s Llama 7B and trained it on a set of prompts that mimic instructionfollowing models like ChatGPT. This fine-tuning resulted in Alpaca 7B, an LLM that unlocks the knowledge encoded in the Llama LLM for an average person to access by asking questions and giving instructions. According to some estimates, this lightweight LLM could run on less than $600 worth of hardware. The creators of Alpaca 7B are releasing the training set and the code that built it. Anyone can reproduce the model or create something new from a different set.

Cerebras

When specialized hardware and a general model evolve together, a highly efficient and fast solution can be achieved. Cerebras offers its LLM on Hugging Face in a variety of sizes, ranging from the smallest (111 million parameters) to the largest (13 billion parameters) for those who want to run it locally. However, many will want to use the cloud services, which run on Cerebras’ own embedded processors optimized for large-scale training analysis.

Claude

Anthropic has created Claude to be a helpful assistant that can handle many business text tasks, from research to customer service. The user sends a message and gets a response. Anthropic intentionally designed the prompts to be long to encourage more complex instructions, enabling users to have better control over the results. Anthropic currently offers two versions: the full model called Claude-v1, and a simplified version called Claude Instant, which is significantly cheaper. The former is intended for tasks requiring more complex and structured reasoning, while the latter is faster and better suited for simple tasks like classification and moderation.

Falcon

The large-scale Falcon-40b and the smaller Falcon-7b were built by the Technology Innovation Institute (TII) in the United Arab Emirates. The teams trained the Falcon model on a large set of general examples from RefinedWeb, with a focus on inference improvement. They then flipped it around and released it under Apache 2.0, making it one of the most open and free models available for experimentation.

FrugalGPT

FrugalGPT is not a different model, but rather a careful strategy to find the cheapest model possible to answer a particular question. The researchers behind FrugalGPT recognized that many questions do not require the largest and most expensive model. Their algorithm starts with the simplest one and works its way up a cascading list of LLMs until it finds a good answer. The researcher’s experiments suggest that this cautious approach can save up to 98% of costs.

Source: www.lemondeinformatique.fr