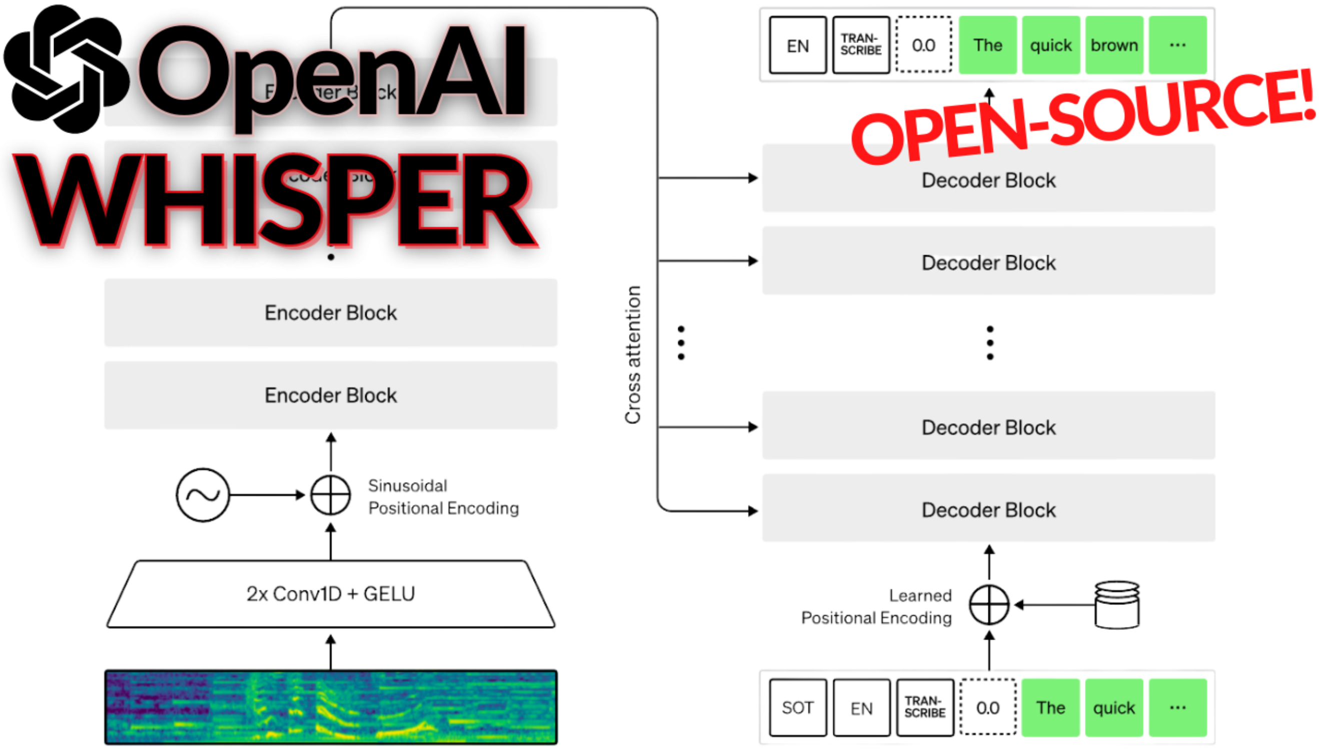
Louis-François Bouchard, aka What's AI

ARTIFICIAL INTELLIGENCE

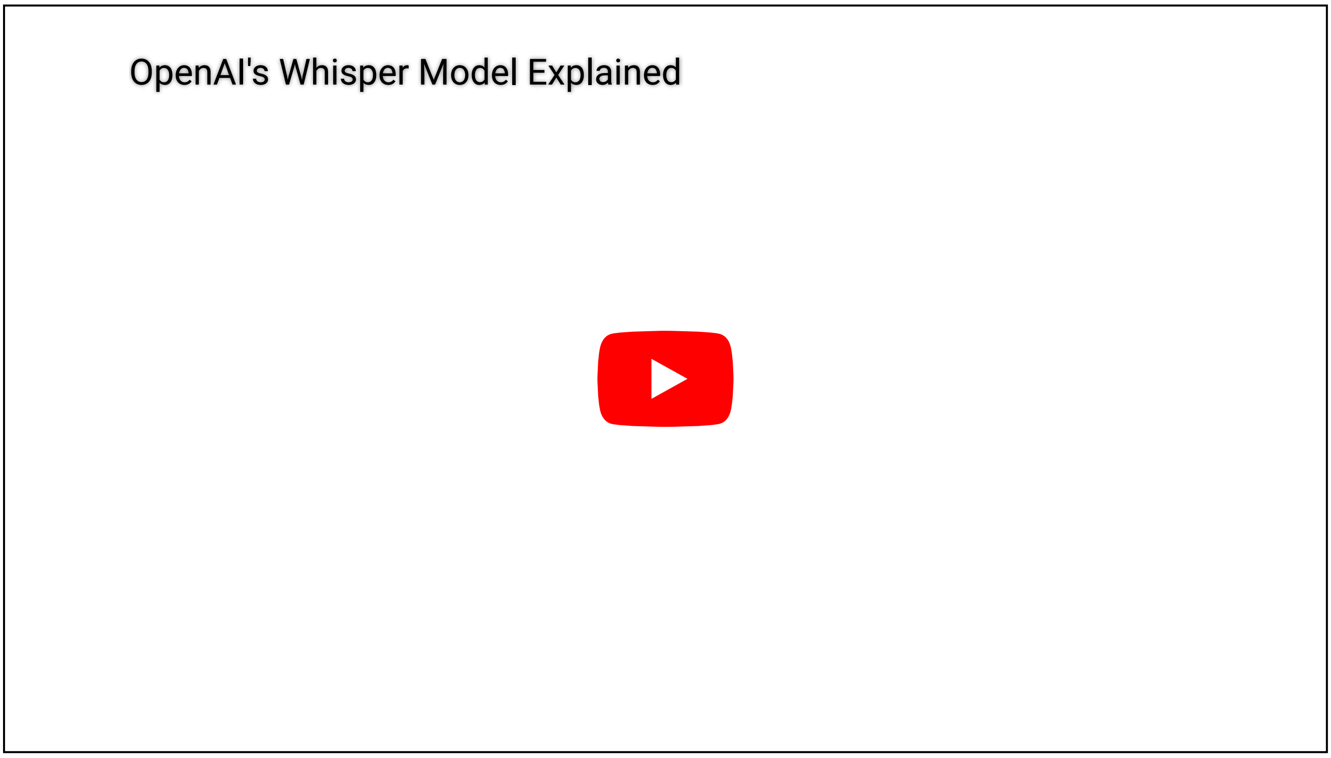
OpenAI's Most Recent Model: Whisper (explained)

A good transcription tool that would accurately understand what you say and write it down

Louis Bouchard
Oct 5, 2022 • 3 min read

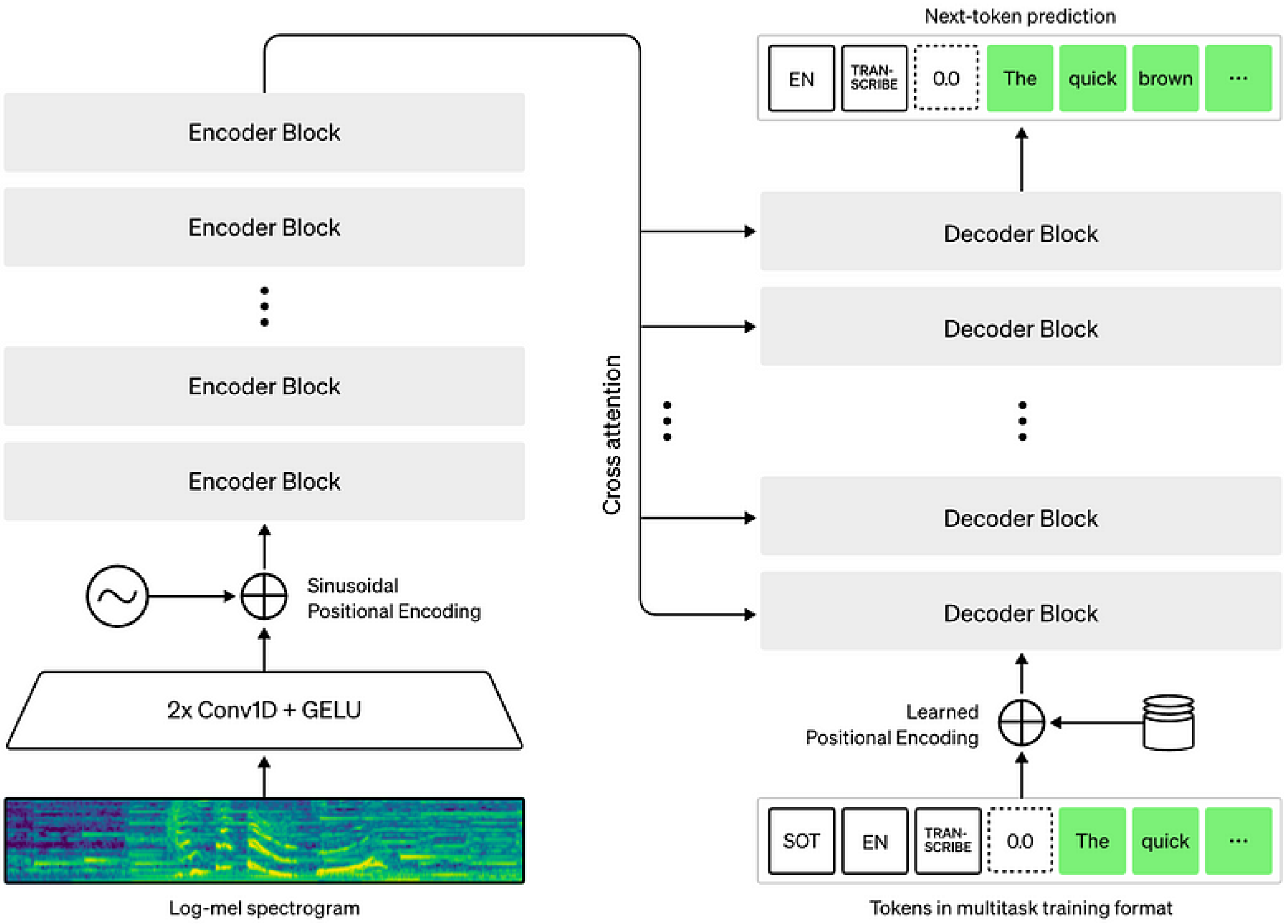


Watch the video



Have you ever dreamed of a good transcription tool that would accurately understand what you say and write it down? Not like the automatic YouTube translation tools… I mean, they are good but far from perfect. Just try it out and turn the feature on for my video above, and you'll see what I'm talking about. Well, OpenAI just released and open-sourced a pretty powerful AI model just for that: Whisper. It even understands stuff I can't even comprehend, not being a native English speaker (listen in the video)! And it works for language translation too!

The results and precision are incredible, but what's even cooler is how it works. Let's dive into it.



Overview of the Whisper encoder-decoder architecture. Image from the paper.

When it comes to the model itself, Whisper is pretty classic. It is built on the transformer architecture, stacking encoder blocks and decoder blocks with the attention mechanism propagating information between both.

It will take the audio recording, split it into 30-second chunks and process them one by one. For each 30-second recording, it will encode the audio using the encoder section and save the position of each word said, and leverage this encoded information to find what was said using the decoder.

The decoder will predict what we call tokens from all this information, which are basically each words being said. Then, it will repeat this process for the next word using all the same information as well as the predicted previous word, helping it guess the next one that would make more sense.

The overall architecture is a classic encoder-decoder that I covered in multiple articles, similar to GPT-3 and other language models, which I invite you to check for more architectural details.

This works as it was trained on more than 600'000 hours of multilingual and multitask supervised data collected from the web. Meaning that they trained their audio model in a similar way as GPT-3 with data available on the internet, making it a large and general audio model. It also makes the model way more robust than others. In fact, they mention that Whisper approaches human-level robustness due to being trained on such a diverse set of data ranging from clips, TED talks, podcasts, interviews, and more, which all represent real-world-like data with some of them transcribed using machine learning-based models and not humans.

Using such imperfect data certainly reduces the possible precision, but I would argue it helps for robustness when used sparsely compared to pure human-curated audio datasets with perfect transcriptions.

Having such a general model isn’t very powerful in itself, as it will be beaten at most tasks by smaller and more specific models adapted to the task at hand. But it has other benefits. You can use this kind of pre-trained models and fine-tune them on your task. Meaning that you will take this powerful model and retrain a part of it, or the entire thing, with your own data. This technique has been shown to produce much better models than starting training from scratch with your data.

And what's even cooler is that OpenAI open-sourced their code and everything instead of an API, so you can use Whisper as a pre-trained foundation architecture to build upon and create more powerful models for yourself.

Some people have already released tools like the youtube whisperer on huggingface by jeffistyping, taking a youtube link, and generating transcriptions.

They also released a google colab notebook to play with it right away.

While some think competition is key, I'm glad OpenAI is releasing some of its work to the public, as I am convinced such collaborations are the best way to advance in our field. Let me know what you think if you'd like to see more public releases of OpenAI or if you prefer the final products they build like DALLE.

As always, you can find more information about Whisper in the paper and code linked below.

I hope you enjoyed this article, and I will see you next week with another amazing paper!

References

▶ Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I., Robust Speech Recognition via Large-Scale Weak Supervision.
▶Project link: https://openai.com/blog/whisper/
▶Code: https://github.com/openai/whisper
▶Google Colab notebook:

https://colab.research.google.com/github/openai/whisper/blob/master/notebooks/LibriSpeech.ipynb

▶YouTube Whisperer app: https://huggingface.co/spaces/jeffistyping/Youtube-Whisperer

▶My Newsletter (A new AI application explained weekly to your emails!): https://www.louisbouchard.ai/newsletter/

2023: The best AI papers - A Review 🚀

A recap of the research progress and important news in AI in 2023!

Louis Bouchard
Dec 24, 2023 • 7 min read

What's AI Episode 25: Jerry Liu. From RAG Strategies to Gemini's Impact in Tech

The What's AI podcast episode 25 with Jerry Liu: LlamaIndex CEO and co-founder

Louis Bouchard
Dec 18, 2023 • 44 min read