| **Model Size** | **Parameters** | **Speed** | **Accuracy** |
| --- | --- | --- | --- |

|  |  |  |  |
| --- | --- | --- | --- |
| tiny | ~39M | Fastest ⚡ | Least accurate |

|  |  |  |  |
| --- | --- | --- | --- |
| base | ~74M | Fast | Moderate |

|  |  |  |  |
| --- | --- | --- | --- |
| small | ~244M | Slower | Good |

|  |  |  |  |
| --- | --- | --- | --- |
| medium | ~769M | Slower | Very good |

|  |  |  |  |
| --- | --- | --- | --- |
| large | ~1550M | Slowest 🐢 | Best accuracy ✅ |

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Size** | **Parameters** | **Speed** | **Accuracy** |
| tiny | ~39M | Fastest | Least accurate |
| base | ~74M | Fast | Moderate |
| Small | ~244M | Slower | Good |
| large | ~1550M | Slowest | Best Accuracy |

It’s been months that I started discovering the realm of Gen AI, and it is quite an adventure, from NLP specialization ( NLP with Vector space models (4 weeks covered in 2 weeks) -> NLP with probabilistic models (4 week covered in 1 week) -> NLP with sequential models (3 weeks covered in 1 week) -> NLP with attention models (3 weeks covered in 1 week), you get into transformers architecture, the paper “attention is all you need”, people mentioning LLMs, gen AI buzzwords. And then you trace how you will be continuing the learning path learning on one side coding on the other side (and of course with 8 hours of working as a data science engineer + consulting), how you will manage all of this, and then the path will be built much like how we build a gen ai project:

1. Data collection: based on the task you want to perform, input differ from audio to image to text. And what we need for acturate modeling is that the data can be unlabeled but it contains the label as a response (self supervised modeling)
2. Data Preprocessing: It’s an amazing step that usually you won’t be performing it as it will be taken apart by the gen ai and LLM models already existing (It involves all techniques that clean the data for text we can perform removing stop words, punctiuation, html tags, lemmatization and stemming, remove email addresses, handle abbreviations as well as emojis .. you can find all what you need through this book that I found interesting once scrolling in linkedin)
3. Data Tokenization: Split the preprocessed data into words, sentences, characters (depends from the tokenization method and you’ll find an initial understanding of the tokenization methods within this github repo). This step is crucial as a computer doesn’t understand human language so the latter will be transformed in token which are ids that map to the words from the vocabulary. (a demo will be shared to showcase the different type of tokenization with reference videos for better understanding)
4. Modeling: use a pretrained model that you can fine tune over a labeled dataset for a specific task
5. Evaluation of the models through bleu, rouge, perplexity .. scores. I’ll be sharing soon a file about the different scoring methods (at least the ones that I get to know, their definition and with examples)
6. Model Deployment
7. Monitoring based on feedback

Currently I’m in the modeling phase and alternating with evaluation and model fine tuning. In this phase I inspire myself from existing project, plan on how I can enhance them and test my learning through these challenges. Here’s one of the project that I’m currently working on it:

Speech 2 Text and speech 2 Image: the project involves a flask application where the user will be uploading a recording, and enter the task that he wanted from the model to perform such as translate, transcript, detect the language or generate an image that explains the speech with a text. In this phase the user can’t enter any creative prompt other than one singular word which summarize the needed task from the model.

For that, multiple models were combined and for each model, I added the intuition behind it, how it works the architecture and without going further with mathematical details ( with that I’ve already added some reference for better explanation on mathematical model behind especially the diffusion model)). It all begun with transcription using Open AI whisper model and due to its limitations for translation task as it can only translate to English language, I combined with it the famous T5 model for the translation task and finally Dreamlinke for image generation.

Now when getting to use the project, you’ll see the need of integrating human prompting as the models such T5 can perform different other tasks than translating such as summarizing which can be needed for an audio especially that nowadays we don’t want really to hear our friends vocal to interrupt our music. So the next phase will be integrating prompt engineering, to allow the user to input the prompt as he want and answer him along the way.

🌟 **Bridging Speech, Text, and Images: A New Chapter in Generative AI** 🚀

Months of diving deep into **Generative AI** have led me to a new, exciting challenge: combining **Speech-to-Text**, **Translation**, and **Image Generation** into a single project! After exploring areas like **NLP**, **Transformers**, and **Diffusion Models**, I’m now working on a project that brings together **text**, **audio**, and **images** in a cohesive way — and it’s an adventure I’m excited to share! 🎙️➡️📚➡️🎨

**What’s this project about?** The app allows users to upload an audio recording and choose a task, like transcribing speech, translating it, or generating an image based on the content. But the cool part? This isn’t my first project; it’s just the first time I’m merging **text**, **image**, and **audio** into one. The journey has been incredible, seeing how these different modalities can work together seamlessly.

🛠️ **Models Used**:

* **Whisper** for speech-to-text transcription.
* **MBart** for translation to a target language selected by the user (and soon, summarization).
* **Dreamlike Diffusion** for generating images from spoken content.

What makes this project different from others I’ve worked on is the interplay between these models. Each one has its own strengths and limitations, and in this project, I’m blending them to deliver a smooth, user-friendly experience. The result is a simple interface where users can input tasks (like “translate” or “image”), and the system handles everything in the background — **audio** is transcribed, **text** is processed, and **images** are generated.

🔍 **What you’ll find inside**:

* A **ReadMe** with a breakdown of each model’s **architecture**, **pros and cons**, and how they interoperate in this project.
* A mini-book explaining the **architectural details** of models like **Seq2Seq**, **Transformers**, and **Diffusion Models**, and how I used them to achieve multimodal results.

💬 **Next Steps**: While the app currently allows users to choose basic tasks, the next phase is integrating **prompt engineering** to allow users to issue more complex, natural language commands. Imagine being able to say, “Create an image from this story” or “Summarize this speech” — that’s where I’m headed!

This project is a milestone in my journey with **Generative AI**, and I’m thrilled to see how these diverse models can work together in a real-world setting. It’s all about experimenting, learning, and pushing the boundaries of what AI can do!

#GenerativeAI #AIProjects #NLP #SpeechToText #TextToImage #LLMs #DiffusionModels #PromptEngineering #AIInnovation #TechJourney #WomenInTech #DataScience #DeepLearning #OpenAI