Alpaca Methodology for training text2text models

**Models Interactions**

**Direct – Pipeline**

This method is used by directly downloading the weights and the full model on the system, this allows you interact with the model however you prefer and modify any of the parameters. The pipeline function, allows you to generate responses with less code and more user friendly.

**Advantage:** Fastest interaction, no breakdowns, allows model modifications,

**Disadvantage**: Requires downloading full model, some models require GPU for fast interactions, some models work slightly different important to read each one documentation.

**Uses**: All cases

**Notes**:   
  
https://huggingface.co/docs/transformers/main\_classes/pipelines

**InferenceClient**

The InferenceClient is a function that allows you to interact with models that are hosted and running on Hugging Face Cloud services (Inference API or Inference Endpoint). This allows you to quickly interact with the models that are less than 10GB.

**Advantage:** No need to download, not much code. You can interact with bigger models if they are self hosted, by a price.

**Disadvantage**: Most of NLP tasks are allowed but limited, maximum result up to 10Gb, models need time to “wake up” on first request, the function will normally send an error message on first request. Not all parameters can be used. I struggled at viewing full responses.

**Uses**: All NLP tasks used during this research and more

**Notes**:   
  
https://huggingface.co/docs/huggingface\_hub/v0.20.1/en/package\_reference/inference\_client#inference

**Requests (Hosted Inference API)**

Test and evaluate, for free, over 150,000 publicly accessible machine learning models, or your own private models, via simple HTTP requests, with fast inference hosted on Hugging Face shared infrastructure. Similar to Inference Client but by doing HTTP Requests

**Advantage:** No need to download, each interaction is different allows to use context and different prompts.

**Disadvantage**: Most of NLP tasks are allowed but limited, maximum result up to 10Gb, models need time to “wake up” on first request, the function will normally send an error message on first request. Not all parameters can be used. Each interaction is different according to the task.

**Uses**: All NLP tasks used during this research and more

**Notes**:   
  
https://huggingface.co/docs/api-inference/index

**LangChain**

A framework that allows you to interact with multiple LLM, it has multiple uses in the interaction, evaluation, analysis and deployment of models. It allows to directly interact with some models own API, for this research we only used the HuggingFace capabilities on their API models.

**Advantage:** No need to download model, it gives the best capability for the creation of prompts and model interactions with entering parameters. You can interact with heavier models but the time of response is long. It has some error detection mechanisms and will stop if the response takes more than 7 minutes to generate.

**Disadvantage**: For the Hugging Face Hub library, only a small amount of tasks are permitted: Text Generation, Text2Text Generation and Summarisation.

**Uses**: Limited tasks allowed on the Hugging Face Hub. Other libraries allow more interactions and task, each one requires you to get an API token, some are free others require payment like OpenAI

**Notes**:

https://python.langchain.com/docs/get\_started/introduction  
https://python.langchain.com/docs/integrations/llms/

**Spaces (Inference Endpoint)**

For any model, Hugging Face offer the solution to host your model on the cloud spaces (this could is hosted by a third party: Gradio, Streamlit, Docker) for no GPU and 16GB this spaces are for free. The more memory and compute required a price needs to be paid. This spaces only charge by the time of usage of the model.  
  
For or case of interacting with the models, this can be done on the HuggingFace space site (which is similar like interaction to ChatGPT) if the space is hosted in Gradio, it allows to send requests in Python.

**Advantage:** It allows to interact with really big models, best models are found here, it can be really fast interaction. Some models (if specified) allow to change parameters while interacting with.

**Disadvantage**: Not all spaces allow python HTTP requests, some spaces are slow and have a waiting list for request.

**Uses**: All cases depending on the space.

**Notes**:

https://huggingface.co/docs/hub/spaces-overview