**Detecting Hallucinations in Large Language Models Using Semantic Entropy**

**Files required:**

* model\_loader.ipynb
* json\_data\_loader.ipynb
* Semantic\_entropyqa.ipynb

**1. Request Access from Meta**

Since LLaMA-2 is licensed by Meta, you need permission to use the model through Hugging Face.

* Go to the LLaMA-2 model page on Hugging Face.
* Click **Request Access** and fill out the form with the required details about your intended usage.
* After submitting the form, Meta will review your application. This may take some time, depending on their process.
* Once your request is approved, Meta will notify you by email, confirming access to the model.

**2. Set Up a Hugging Face Account and Access Token**

You’ll also need a Hugging Face account to access the model and a token to authenticate.

* Create a Hugging Face account if you don’t have one.
* Generate an access token:
* Go to your Hugging Face profile, click Settings > Access Tokens > New Token.
* Select Read permissions, then click Generate New Token.
* Replace "hf\_eVdzQnpyxAlfErkpQIaOqRcxOVMJDJAyru" in the model\_loader\_ipynb notebook code with your own token. This will allow you to load the model.

1. **Set Up a Google Colab Account and Workspace**

* If you don’t already have a Google account, create one to access Google Colab.
* Colab provides free GPU access, though for more intensive tasks, you may want to consider upgrading to Colab Pro as this method requires extensive computational resources.

1. **Prepare Your Notebooks and Files**

* Upload model\_loader.ipynb, json\_data\_loader.ipynb, and Semantic\_entropyqa.ipynb to your Google Drive.
* Organize these files within a specific project directory in Google Drive (e.g., Colab Notebooks/LLM\_Hallucination\_Detection).

1. **Run Semantic Entropy Calculations (Semantic\_entropyqa.ipynb):**

* Adjust your file paths according to your google drive directory for saving generated answers and entropy calculations.
* Execute the cells in this notebook to calculate semantic entropy for LLaMA-2’s outputs.
* Analyse the entropy scores for patterns indicative of hallucinations.