Python-based information extraction

Python, a programming language mainly used for task automation, data analysis, and data visualization is required to be used in the project for its appropriate application into the project goal and its versatility with available suitable libraries such as “docx” or “tkinter” to be used during the project for text extraction. It also provides an integrated solution with Large Language Models (LLMs), a type of artificial intelligence (AI) algorithm that uses deep learning techniques and large data sets in order to understand, summarize, generate and predict new content. This type of technology was used in this project to check for inconsistencies in the Animal Ethics documents for its capability of accurately extracting useful records of complex scientific knowledge (Dunn et al., 2022), helping to reduce the manual workload and enhance document review accuracy.

LLM

First and foremost, the team had to install specific Python libraries and tools in Google Colab. This installation included '*transformers*,' a Python library developed by Hugging Face for handling and utilizing various Large Language Models (LLMs), including Llama 2. In this project, it was used for loading and employing the Llama 2 model to carry out text processing tasks. 'B*itsandbytes*' was employed for binary data processing specific to the project, '*accelerate*' to boost the training process of Llama 2 or other deep learning models, thereby enhancing training speed, and '*sentencepiece*,' a Python library used for tokenization and text processing.

Subsequently, it was essential to define the identifier or name of the model required for the project. This identifier should point to the Llama-2-13b-chat-hf model stored in The Hugging Face Hub model repository. Following this, the team configured the model quantization settings. In this case, the model was configured for 4-bit quantization to reduce memory usage and enhance efficiency during model execution. Once these preparatory steps were completed, the specified pre-trained model was able to be downloaded from The Hugging Face model repository.

Next, the team devised a form of interaction with the Llama 2 model. This involved creating text containing task descriptions and input information for the model to process. The model accepted this text and generated response information, which was subsequently obtained. This process served the ultimate purpose of performing a consistency check on textual information. Specifically, the team conducted consistency checks on Musculus Strain, Mortality, and Animal pain/distress level with extracted AEC application documents in JSON files.