**Objective:** Create a topic model for scientific research, with the hope of applying it to new, unseen papers to gain insight into what these new papers are about

**Progress made so far:**

I have completed downloading, cleaning, and preprocessing the data, as well as the Latent Dirichlet Allocation (LDA) model using pyLDAvis. The LDA model splits papers into 50 topics but does not label them. So I have also spent some time looking at topics and thinking how I could use this for making the topic model I actually want.

**Tasks Remaining:**

I have decided to switch focus away from LDA (reasons in the Challenges section) and will instead look at LLMs.

I need to train an LLM on my corpus and then use it for inference. This will likely require gathering training data (which I need to think about), as well as selecting and training a HuggingFace model. I have played with the 7B parameter Falcon model, which I think is small enough to train, but big enough to give a decent result. I need to train it, then I will need to write the code for inference.

**Challenges Faced:**

**LDA**: I have combed through the LDA result quite a bit, but don’t think it will be nearly enough to complete my task, as only some of the topics it selects are even comprehensible (i.e., I can tell what subject they’re about). With hundreds of hours more work (not really programmatic, mostly combing through topics and applying my domain knowledge), I think I could turn this into something, but given I don’t get anywhere near that time, I think I’m going to switch focus. I would like to instead train an LLM. I think this is a slight deviation from my original project but should give me the result I am looking for (to gain insight into new research papers).

**LLM:** I will be using HuggingFace, which I have never used before. As such, there will be a learning curve to downloading, training, uploading, redownloading and inference. I suspect the LLM itself is big enough to give me a good result, so this challenge is mostly around workflow and maybe some preprocessing of data.