Here are some basic instructions for fine-tuning a BioBERT model to classify publications into two categories (CDK4 or bacterial cell division). We trained another model for extractive Q&A as well (see below).

The basic steps are:

1. Collect labelled data
2. Fine-tune a 🤗 Transformers models through a RunPod or Google Colab instance while tracking the performance using Tensorboard or wandb
3. Save fine-tuned model (as a tar.gz file) and transfer to an S3 bucket
4. Deploy fine-tuned model directly from an S3 bucket to an Amazon SageMaker endpoint for inferencing
5. Deploy a local Streamlit application that sends user input to this endpoint and displays the results.

**Fine-tuning your BERT model for classification**

* Use a Google Colab or RunPod GPU instance and work through the .ipynb notebook and fine-tune your BERT model
* Download the model.tar.gz file (which contains the model tensors/weights, and tokenizer settings) from your training instance or from Google Drive (if mounted)

**Deploying a 🤗 Transformers models in SageMaker for inference (classification)**

* See link [here](https://github.com/huggingface/notebooks/blob/main/sagemaker/10_deploy_model_from_s3/deploy_transformer_model_from_s3.ipynb) for more info
* Setup a S3 bucket and upload the tar.gz archive to your S3 bucket
* On your local computer, run the ‘setup-inference-endpoint.py’ to deploy the inferencing endpoint
  + The script tells AWS that the model files are in an S3 bucket and to setup an inferencing endpoint instance so we can use it (almost like an API)
* Run the app by going to the command line, cs to the folder of interest containing the streamlit\_app.py file and type ‘streamlit run streamlit\_app.py’
* This will launch the app in your default web browser at http://localhost:8501
* **Important:** Make sure you delete the endpoint and the bucket once you are done to avoid being charged!