This is an explanation of every file under the “pj” folder in the repository

**OCR\_docs:** A folder of text-files I was looking to test our fine-tuning model on. These files were not included in the original fine-tune, but very well can be if needed to.

**text\_docs:** A folder of text-files used to train the fine-tune initially. There are 50 text-files in this folder and as stated, these were edited to fit GPT-3 token-limit and were stripped of all new-line or unrecognizable characters. Additionally, the “\n\n###\n\n” character was added to each of these text-files, and it is required that this character is used in training and testing prompts in the future.

**trainingDocs:** folder of OCRd PDFs which were intended to be tested on. These are the OCR PDF counterparts for all of the text-files in OCR\_docs. I recommend to look over these when comparing the completions from fine-tuning testing to the correct information, so you can see exactly why/where the fine-tune went wrong.

**finetuneTrain0608:** csv of prompt-completion pairs which was prepared to initially train our fine-tuning job. Used the OpenAI data-preparation tool found on the OpenAI website in order to convert this into JSONL format. Was created using the jupyter notebook found in this file.

**finetuneTrain0608\_prepared2.jsonl:** The prepared jsonl format used to train our fine-tuning model. This was created using the data-preparation tool, and this is the file used to create a new fine-tuning job. The instructions to use/create this data are found in the fine-tuning tab on the OpenAI website.

**finetuning\_script.ipynb:** This is the jupyter notebook where all of the scripts that prepared this data/tested this data can be found. The notebook has comments throughout which explain the workflow of everything, and all of the directories found on that notebook were made to fit my computer. If used on another computer, make sure to edit directories accordingly.

**mlWorkflow:** csv of the ML Workflow spreadsheet found on the drive. This sheet is not up-to-date, but was used to train the data for the fine-tune.