Based on my analysis of the entire codebase, here's how the code splits training and test data:

Training and Test Data Splitting in the Refact Finetune System

Overview

The data splitting process happens in multiple stages with sophisticated logic to handle different scenarios:

1. Initial File Assignment (process\_uploaded\_files.py)

Configuration-Based Assignment:

Files are initially assigned to train/test sets based on a configuration file (PP\_CONFIG\_HOW\_TO\_UNZIP)

Each uploaded file/archive can be explicitly assigned to "train" or "test" set via the which\_set parameter

Default assignment is "train" if not specified

Assignment Process:

2. Automatic Test Set Selection (file\_sets\_context.py)

Auto-selection Logic:

If no files are manually assigned to test set (len(self.test\_files) == 0), the system automatically selects test files

Uses autoselect\_test\_files\_num parameter (default: 3 files from train\_defaults.py)

**Test files are selected as: min(autoselect\_test\_files\_num, len(train\_files) // 2)**

3. Final Set Construction (process\_uploaded\_files.py)

Filtering and Final Assignment:

4. Deduplication Logic

Handling Duplicates:

During deduplication, if files are found in both train and test sets, test takes priority

5. Key Configuration Parameters

From train\_defaults.py:

Minimum File Requirements:

6. Output Files

The system generates these output files:

train\_set.jsonl - Training files

test\_set.jsonl - Test files

PP\_TRAIN\_FILTERED\_FILEPATH and PP\_TEST\_FILTERED\_FILEPATH - Filtered versions used in training

Summary

The splitting strategy follows this hierarchy:

Manual Configuration - Explicit assignment via configuration files

Automatic Selection - Random selection of 3 files (or half of total files, whichever is smaller) for test set

Deduplication - Test set takes priority over train set for duplicate files

Filtering - Apply file type and quality filters before final assignment

This approach provides flexibility for both manual control and automatic handling of train/test splits while ensuring data integrity and preventing overfitting through proper test set isolation.