RenAIssance Project Tests for Prospective GSoC 2026 Applicants

Below are the tests we will use to evaluate prospective GSoC students for the **RenAIssance** project. Please thoroughly complete the specific test for your project of interest and (optionally) other tests if you would like to also be considered for additional projects at the same time. This may increase your chances of success, but make sure you don't do it at the expense of the specific project you are interested in.

Note: please work in your own github branch (i.e. NO PRs should be made). Send us a link to your code when you are finished by following the instructions below, and we will evaluate it. We encourage you to submit your solutions at least 1 week before the GSoC Proposal Submission deadline, or earlier, so that you have enough time to write the proposal.

General Dataset: choose any public dataset appropriate for this text recognition task.

Specific Dataset:

Dataset: text scans PDF sources scans
Dataset reference Transcribed sources text

Description of the Specific Dataset
The dataset consists of 6 scanned early modern printed sources and another set of 5 handwritten sources – choose the dataset that best applies to the proposal you're interested in applying into. The images have a simple recognition applied that reflects the limitations of the OCR already used (missed letters, incorrectly recognized words...), each source is saved as separate PDF file. The PDF is editable and can be saved as JPEGs if that helps with processing. Marginalia can be ignored, applicants should only focus on the main text, which may vary in layout and page organization depending on the source. The dataset also includes a transcription of the initial parts of each PDF source – they should be used as ground truth reference while training the AI models for the project. The transcriptions also include a few notes that should help manage expectations of slight variability in spelling errors or limitations throughout the text. Each source has more content than the first transcribed part, so that renAIssance can evaluate the degree of accuracy and viability of the test method employed.

Tasks:

- Those interested in working on Optical Character Recognition of **printed** text:
  - Complete Test I
- Those interested in working on Large Language Model pipeline creation for **handwritten** text:
  - Complete Test II

Specific Test I. Optical Character Recognition of printed sources

Task: Build a model based on convolutional-recurrent, transformer, or self-supervised architectures for optically recognizing the text of each data source. Your model should be able to detect the main text in each page, while disregarding other embellishments. Integrate an LLM or VLM model as a late-stage step to the OCR process (such as cleaning up the OCR output). Pick the most appropriate approach and discuss your strategy.

Evaluation Metrics: discuss which evaluation metrics you are using to evaluate your model performance

-- **Specific Test II. Text recognition of handwritten sources**

Task: Build an OCR pipeline based on a LLM or VLM, and optionally integrate with an OCR algorithm. The OCR algorithm may be of your own creation, or a preexisting one that you have selected. Since they all make errors, the crucial part of this task is to finetune the LLM/VLM for OCR, or embed an OCR method into a pipeline that also uses an LLM (such as Gemini) or VLM throughout the process: reading the image, interpreting it, producing the OCR output, and correcting or at least predicting the most likely correct spelling of the recognized text. The LLM/VLM should not be a late-stage step for this test, it should be used at all stages of the pipeline, where it makes most efficient gains.

Evaluation Metrics: discuss which evaluation metrics you are using to evaluate your pipeline performance

----

Submission Guidelines:

Please send us your CV and a link to all your completed work (github repo, Jupyter notebook + pdf of Jupyter notebook with output) to human-ai@cern.ch with Evaluation Test: RenAIssance in the title.