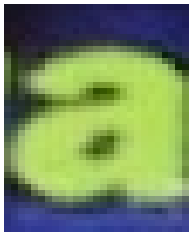
OCR Dataset

Connor Dempsey, Devin Seberino, Michael Cook

October 2025

About the Dataset

- The datasets consist of images of words, individual characters, and whole pages.
- One set has characters with a plain white background and the other sets have characters of different sizes, fonts, and backgrounds.
- The following are some examples:



(a) Character with non-white background

(b) Word with non-white background

(c) Word with non-white background

Limitations

- These data sets only contain individual characters or individual words
- This data does not compare well to our ideal inputs:pdf documents
- Data sets should have text that can be processed line by line