- - - - *Only for training labels (**not at inference time**)*

TEX

Labeling script

Annotates the proofs and theorems in LaTeX source file

Compilation

PDF

PDFAlto

Extracting font information across lines

Grobid

Extracting text and layout information per paragraph, through coordinates

Image generation script

Generates images of the paragraphs, using layout information

CSV

Merging script

Merging information about layout, font, and text per paragraph

...

Multimodal dataset

Text, font sequence and image for each paragraph