DOGE Paper

Issues :- ① Suboptimal annotation quality
② Lack of diversity in task format

features :- ① 1.4M document parsing data
text box annotation → word, phrase, line, paragraph
↓
Chart, Poster, Pdf

② 700k instruction tuning dataset
- text in location out
- location in text out

DOGE - Engine :- ① Poster → Crello dataset
Re-rendering + ① render with meta annotation
② modify color and opacity of 1 textbox + re render
③ pixel wise substraction

Repeat for every text box

→ then Normalized.

② Chart → Chart QA - bar, chart line, pie, ... etc
+ JSON / CSV

matplotlib → creat chart image +
+ Random padding + Remove text for 1/3 data
+ Randomly mask half of text in 2/3
+ Re-rendering strategy

③ PDF document :- CCMAIN 2021 31 PDF UNTRUCATED

moderately sized file → ? [size?]
Pdf parsing tool → PyMuPDF
+ MinerU

MinerU → problems → missing table, footnotes, other elements
PyMuPdf → issues → doesnt provide reading order.