PDF????????
Automatic PDF Translation System

?????

?????????PDF???????????????????
??????????????????????????

?????
? ??50??????PDF??
? ?????????
? ?????????
? ?????????

1. ??????

??????PyMuPDF?PaddleOCR?Ollama?????PDF?????
??????????????????????????????
?????????????

1.1 ?????????
? extractor: PDF ??????
? layout_analyzer: ???????
? term_miner: ??????
? translator: LLM ?????
? post_processor: ???
? renderer: HTML/Markdown ??

1.2 ?????
1. PDF?????????
2. ???????????????
3. ???????????
4. ???????????
5. ???????
6. ???????

2. ???????????

??????????????????
? API (Application Programming Interface)
? OCR (Optical Character Recognition)
? LLM (Large Language Model)
? Transformer ???????
? PyTorch ??? spaCy ?????
? BERT, GPT-4, Gemma ????????

2.1 ????????
1. PDF????????
   - PyMuPDF?????????
   - ?????????

2. ?????????
   - LayoutLM?????????
   - ??????????

3. ????????????
   - spaCy????????
   - Wikipedia API??????

4. ???????????
   - Ollama/OpenAI API???
   - ?????????????

5. ???????
   - ??????????????

6. ????????
   - HTML/Markdown??????