1 Idea

1.1 Seorin comments:

- Transcription: I don't think we should expect full manual transcription from my project.
- I've already done one (transcription_ex1.xlsx) by myself. I verified the surnames with the Belgian genealogy database (https://nl.geneanet.org/genealogie/).
- The "actif" and "passif" and the "restant" amounts are checked by calculation.
- For locations, consult this: https://fr.wikipedia.org/wiki/Arrondissement$_a dministrati f_d e_N ivelles$
- With one example, it doesn't seem to improve the quality of the output. Although I used the next page of the same book as a new image.
- Just to see the effect, I will transcribe one more document.
- Or zero-shot + looking for a better prompt may be the way but I haven't got satisfactory output.
- We need to modify the function for the image adjustment so that (1) cropping depends on the type of documents (whether the entire scan of two pages is one table or each page contains one) and (2) colour adjustments are best for all.

1.2 Julien comments

- Maybe other model than gpt4o while work better, for some task gpt3 is known to perform better
- I added comparison of OCR methos and our LLM idea in the code so we can try to quantify our improvements (Folder: results/comparison)
- Test CropColor Image VS Not CropColor, there is an improvement ?

2 Tests

2.1 Example1.jpeg

	Human	LLM	LLM cc	EasyOCR	EasyOCR cc	Pytesseract	Pytesseract cc
Human	0.000000	0.870813	0.933702	0.919689	0.918575	0.879562	0.863281
LLM	0.870813	0.000000	0.803279	0.961538	0.962585	0.937500	0.924528
LLM cc	0.933702	0.803279	0.000000	0.971660	0.972549	0.948905	0.933333
EasyOCR	0.919689	0.961538	0.971660	0.000000	0.853774	0.901493	0.903427
EasyOCR cc	0.918575	0.962585	0.972549	0.853774	0.000000	0.903790	0.909091
Pytesseract	0.879562	0.937500	0.948905	0.901493	0.903790	0.000000	0.752577
Pytesseract cc	0.863281	0.924528	0.933333	0.903427	0.909091	0.752577	0.000000

Table 1: Jaccard distance

	Human	LLM	LLM cc	EasyOCR	EasyOCR cc	Pytesseract	Pytesseract cc
Human	0.000000	0.957368	0.978122	0.973497	0.973130	0.960255	0.954883
LLM	0.957368	0.000000	0.935082	0.987308	0.987653	0.979375	0.975094
LLM cc	0.978122	0.935082	0.000000	0.990648	0.990941	0.983139	0.978000
EasyOCR	0.973497	0.987308	0.990648	0.000000	0.951745	0.967493	0.968131
EasyOCR cc	0.973130	0.987653	0.990941	0.951745	0.000000	0.968251	0.970000
Pytesseract	0.960255	0.979375	0.983139	0.967493	0.968251	0.000000	0.918351
Pytesseract cc	0.954883	0.975094	0.978000	0.968131	0.970000	0.918351	0.000000

Table 2: Masi distance

	Human	LLM	LLM cc	EasyOCR	EasyOCR cc	Pytesseract	Pytesseract cc
Human	0	275	299	294	293	283	284
LLM	275	0	302	257	260	151	130
LLM cc	299	302	0	305	305	308	306
EasyOCR	294	257	305	0	214	246	246
EasyOCR cc	293	260	305	214	0	248	248
Pytesseract	283	151	308	246	248	0	141
Pytesseract cc	284	130	306	246	248	141	0

Table 3: Levenshtein distance

2.2 Example2.jpeg

	Human	LLM	LLM cc	EasyOCR	EasyOCR cc	Pytesseract	Pytesseract cc
Human	0.000000	0.777778	0.797235	0.935927	0.934091	0.887879	0.898148
LLM	0.777778	0.000000	0.596154	0.955224	0.955752	0.932203	0.924444
LLM cc	0.797235	0.596154	0.000000	0.965318	0.965714	0.955823	0.945148
EasyOCR	0.935927	0.955224	0.965318	0.000000	0.392045	0.908235	0.924171
EasyOCR cc	0.934091	0.955752	0.965714	0.392045	0.000000	0.906542	0.922353
Pytesseract	0.887879	0.932203	0.955823	0.908235	0.906542	0.000000	0.714801
Pytesseract cc	0.898148	0.924444	0.945148	0.924171	0.922353	0.714801	0.000000

Table 4: Jaccard distance

	Human	LLM	LLM cc	EasyOCR	EasyOCR cc	Pytesseract	Pytesseract cc
Human	0.000000	0.926667	0.933088	0.978856	0.978250	0.963000	0.966389
LLM	0.926667	0.000000	0.866731	0.985224	0.985398	0.977627	0.975067
LLM cc	0.933088	0.866731	0.000000	0.988555	0.988686	0.985422	0.981899
EasyOCR	0.978856	0.985224	0.988555	0.000000	0.799375	0.969718	0.974976
EasyOCR cc	0.978250	0.985398	0.988686	0.799375	0.000000	0.969159	0.974376
Pytesseract	0.963000	0.977627	0.985422	0.969718	0.969159	0.000000	0.905884
Pytesseract cc	0.966389	0.975067	0.981899	0.974976	0.974376	0.905884	0.000000

Table 5: Masi distance

	Human	LLM	LLM cc	EasyOCR	EasyOCR cc	Pytesseract	Pytesseract cc
Human	0	287	287	322	323	324	317
LLM	287	0	145	307	310	213	203
LLM cc	287	145	0	310	313	207	197
EasyOCR	322	307	310	0	77	309	308
EasyOCR cc	323	310	313	77	0	312	312
Pytesseract	324	213	207	309	312	0	182
Pytesseract cc	317	203	197	308	312	182	0

Table 6: Levenshtein distance