OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches

Jenna Kanerva, Cassandra Ledins, Siiri Käpyaho, Filip Ginter

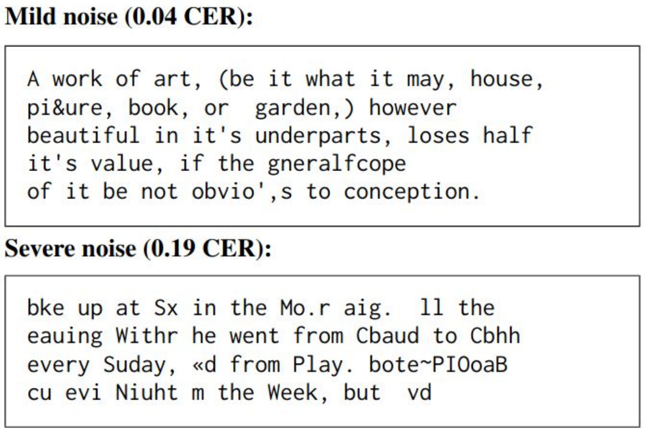
TurkuNLP
University of Turku, Finland

TURKUNLP
.ORG

UNIVERSITY
OF TURKU

Background

- Many historical text collections are digitized through scanning and OCR
    - Noise level varies
- **OCR error post-correction** for improving quality and usability of historical collections
    - No access to original images or full OCR output, only text
    - Re-OCR too expensive



Mild noise (0.04 CER):

```
A work of art, (be it what it may, house,
pi&ure, book, or  garden,) however
beautiful in it's underparts, loses half
it's value, if the gneralfcope
of it be not obvio',s to conception.
```

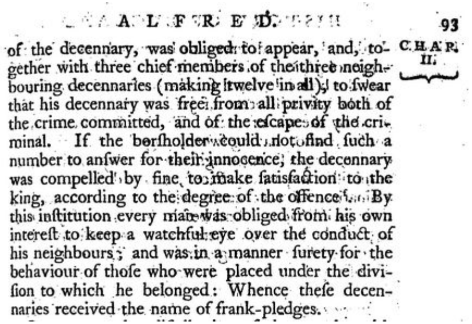
Severe noise (0.19 CER):

```
bke up at Sx in the Mo.r aig.  ll the
eauing Withr he went from Cbaud to Cbhh
every Suday, «d from Play. bote~PIOoaB
cu evi Niuht m the Week, but  vd
```

Objective / RQs

- Can LLMs be prompted to correct OCR errors from historical datasets?
  - **Input:** OCRed text → **Output:** Clean version
  - In related studies, no clear consensus whether LLMs can be applied zero-shot
- Experiments on two large historical datasets (English and Finnish)
- Special focus on open-weight models
  - Commercial models infeasible cost-wise for large datasets

English Data



ALFRED. 93

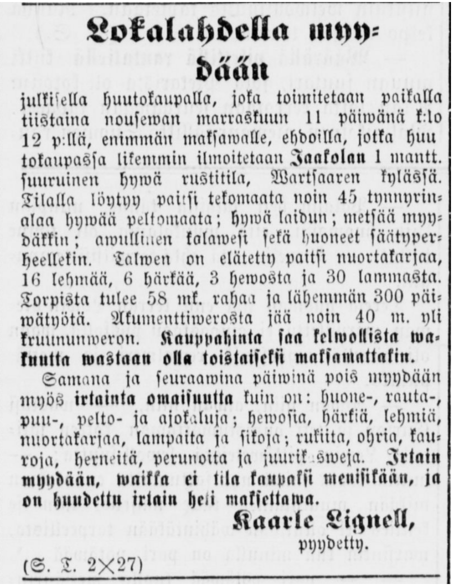
CHAP. II.

of the decennary, was obliged to appear, and, together with three chief members of the three neighbouring decennaries (making twelve in all), to ſwear that his decennary was free from all privity both of the crime committed, and of the eſcape of the criminal. If the borſholder could not find ſuch a number to anſwer for their innocence, the decennary was compelled by fine, to make ſatisfaction to the king, according to the degree of the offence. By this inſtitution every man was obliged from his own intereſt to keep a watchful eye over the conduct of his neighbours; and was in a manner ſurety for the behaviour of thoſe who were placed under the diviſion to which he belonged: Whence theſe decennaries received the name of frank-pledges.

- **ECCO:** Eighteenth Century Collections Online
  - over 180,000 publications originally printed in the 18th century Britain
  - Scan and OCR by Gale company (owns the data!)
- **ECCO-TCP:** Text Creation Partnership
  - 2,000+ manually created full-text transcriptions of ECCO books (CC-0 1.0)
- OCR post-correction dataset by page-level pairing of ECCO OCR and ECCO-TCP texts (Helsinki Computational History Group)

Finnish Data

- OCR ground truth[1] by National Library of Finland
  - Original image (not used in this project)
  - OCR engine output
  - Human made ground truth
- Digitized newspapers published 1836–1918, Fraktur font
  - Individual pages



Lokalahdella myydään

julkiſella huutokaupalla, joka toimitetaan paikalla tiistaina nouſewan marraskuun 11 päiwänä k:lo 12 p:llä, enimmän makſawalle, ehdoilla, jotka huutokaupaſſa likemmin ilmoitetaan **Jaakolan** 1 mantt. ſuuruinen hywä rustitila, Wartſaaren kyläſſä. Tilalla löytyy paitſi tekomaata noin 45 tynnyrinalaa hywää peltomaata; hywä laidun; metſää myydäkſin; awullinen kalaweſi ſekä huoneet ſäätyperheellekin. Talwen on elätetty paitſi nuortakarjaa, 16 lehmää, 6 härkää, 3 hewosta ja 30 lammasta. Torpista tulee 58 mk. rahaa ja lähemmän 300 päiwätyötä. Akumenttiwerosta jää noin 40 m. yli kruunuweron. **Kauppahinta ſaa kelwollista wakuutta wastaan olla toistaiſekſi makſamattakin.**

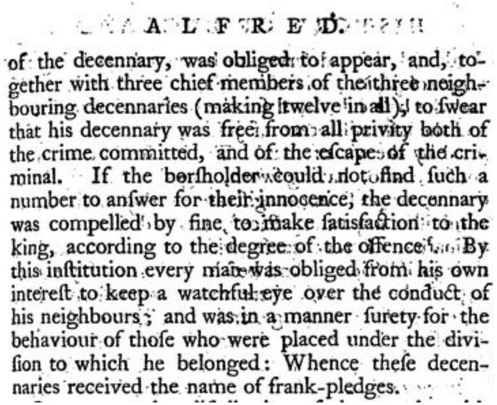
Samana ja ſeuraawina päiwinä pois myydään myös **irtainta omaiſuutta** kuin on: huone-, rauta-, puu-, pelto- ja ajokaluja; hewoſia, härkiä, lehmiä, nuortakarjaa, lampaita ja ſikoja; rukiita, ohria, kauroja, herneitä, perunoita ja juurikasweja. **Irtain myydään, waikka ei tila kaupakſi meniſikään, ja on huudettu irtain heti makſettawa.**

Kaarle Lignell,
pyydetty.

(S. T. 2×27)

[1] http://digi.nationallibrary.fi/

Overview of the data



ALFRED.

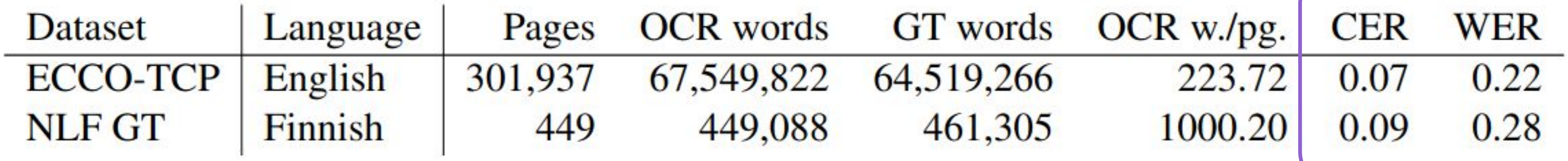
of the decennary, was obliged to appear, and, together with three chief members of the three neighbouring decennaries (making twelve in all), to ſwear that his decennary was free from all privity both of the crime committed, and of the eſcape of the criminal. If the borſholder could not find ſuch a number to anſwer for their innocence, the decennary was compelled by fine to make ſatisfaction to the king, according to the degree of the offence. By this inſtitution every man was obliged from his own intereſt to keep a watchful eye over the conduct of his neighbours; and was in a manner ſurety for the behaviour of thoſe who were placed under the diviſion to which he belonged: Whence theſe decennaries received the name of frank-pledges.

OCR output

of-'the' deCenary, was~ obliged toappear,' and, to:- c:a t.:. gether with three chiefimeribers e of thethie tremoigh- ;. bouring decennaries (makins itwelvei aii) to .swear i that his decennary was frei,fromi allt privity h6th of the crime, committed, anid 6f: the' efcapet ds ,ihd ,cri+. rinal. -If the berfholder'.;Eould .jriotofind fudh. a number to answer foritheislr.nnocei.ce, the deconnary was compelled'by. fine,-toj.bake fatisfaa iori'd to lthe king, .according to the degree: of- the fncei.s n aBy this institution every ma'^tr*bbliged fom his own intereft.to keep a watchf.eye .qover the cdndut. of his neighbours.' and was:in aamanner. furety-foir the behaviour of those who-were placed under the divi- sion to which he belonged: Whence these decen- naries received-the. -name of frank-pledges.

Ground truth

of the decennary, was obliged to appear, and, together with three chief members of the three neighbouring decennaries (making twelve in all) to swear that his decennary was free from all privity of the crime, committed, and of the escape of the criminal. If the berholder could not find such a number to answer for their innocence, the decennary was compelled by fine to make satisfaction to the king, according to the degree of the offence. By this institution every man was obliged from his own interest to keep a watchful eye over the conduct of his neighbours, and was in a manner surety for the behaviour of those who were placed under the division to which he belonged: Whence these decennaries received the name of frank-pledges.



Dataset	Language	Pages	OCR words	GT words	OCR w./pg.	CER	WER
ECCO-TCP	English	301,937	67,549,822	64,519,266	223.72	0.07	0.22
NLF GT	Finnish	449	449,088	461,305	1000.20	0.09	0.28

Metrics

$$\text{CER\%} = \frac{(\text{CER}_{\text{orig}} - \text{CER}_{\text{post}})}{\text{CER}_{\text{orig}}} \times 100$$

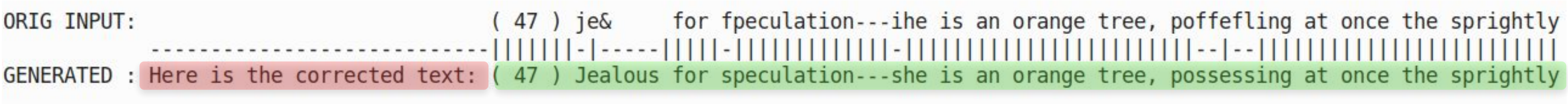
- Relative CER reduction
    - By how much (%) are the remaining OCR errors reduced
    - The overall CER% is an weighted average of example-wise CER%
- Normalization before evaluation: Systematic differences between historical and modern spellings
    - English: **Long-s** to **s**
    - Finnish: **w** to **v**

Experiments

- Split page-level data into segments of 300 subwords
- Random sample of 200 test segments for each language
- **Experimental setting:** Given a prompt and a segment of historical English/Finnish, how much (%) of the OCR errors does the LLM reduce?

Experiments

- Post-processing: LLM Overgeneration Removal
  - LLMs are talkative, usually generate additional explanations
  - Generated output aligned against the original LLM input on character level, and extra leading and trailing texts filtered out



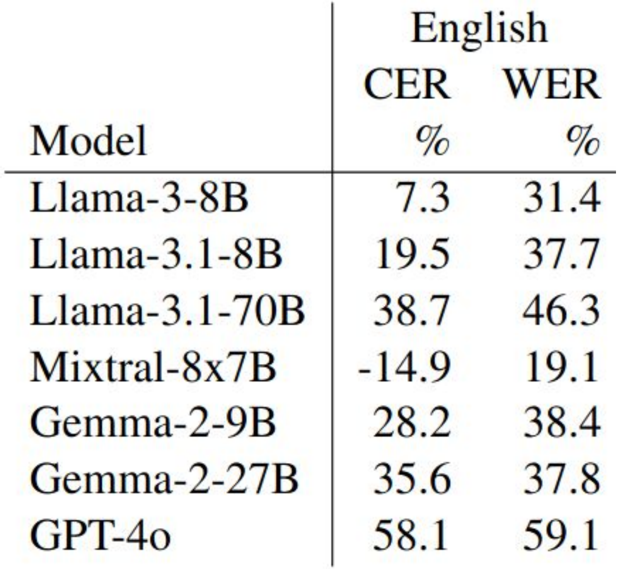
```
ORIG INPUT:                             ( 47 ) je&     for fpeculation---ihe is an orange tree, poffefling at once the sprightly
            ----------------------------|||||||-|-----|||||-|||||||||||||-||||||||||||||||||||||||--|--|||||||||||||||||||||||||
GENERATED : Here is the corrected text: ( 47 ) Jealous for speculation---she is an orange tree, possessing at once the sprightly
```

Results

English

➡ Most models positive on both metrics

- Llama 3.1 70B best open model
- GPT-4o still notably better than open models



	English	
	CER	WER
Model	%	%
Llama-3-8B	7.3	31.4
Llama-3.1-8B	19.5	37.7
Llama-3.1-70B	38.7	46.3
Mixtral-8x7B	-14.9	19.1
Gemma-2-9B	28.2	38.4
Gemma-2-27B	35.6	37.8
GPT-4o	58.1	59.1

→ In general, clear improvement can be obtained

Results

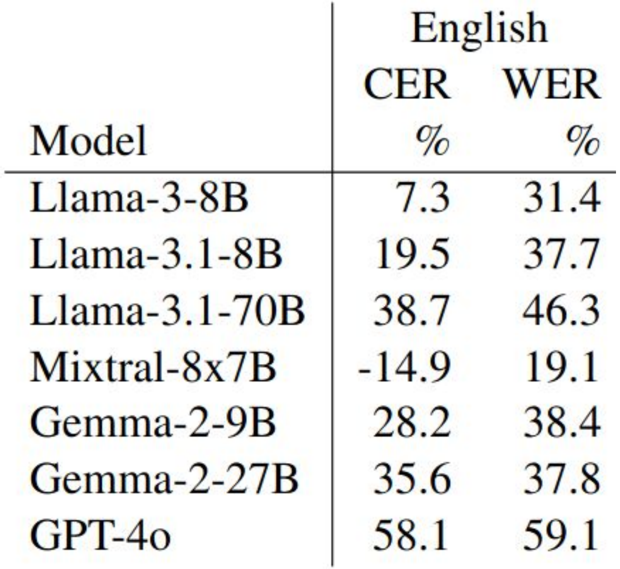
English

- Most models positive on both metrics

➡ Llama 3.1 70B best open model

- GPT-4o still notably better than open models

→ In general, clear improvement can be obtained



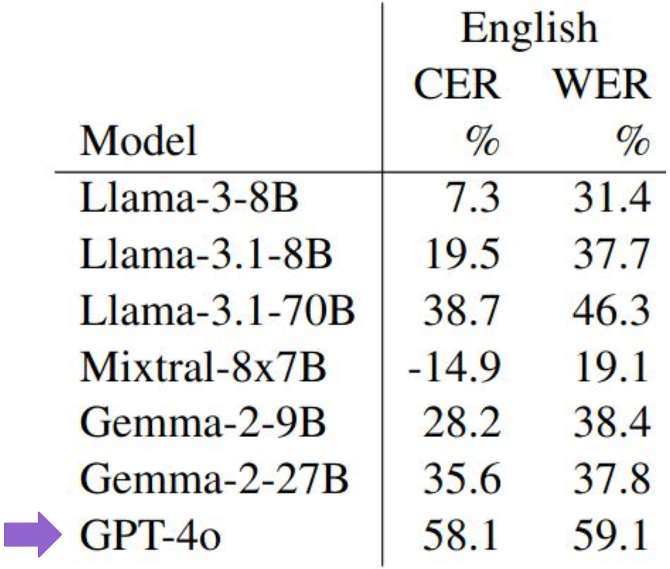
	English	
	CER	WER
Model	%	%
Llama-3-8B	7.3	31.4
Llama-3.1-8B	19.5	37.7
➡ Llama-3.1-70B	38.7	46.3
Mixtral-8x7B	-14.9	19.1
Gemma-2-9B	28.2	38.4
Gemma-2-27B	35.6	37.8
GPT-4o	58.1	59.1

Results

English

- Most models positive on both metrics
- Llama 3.1 70B best open model

➡ GPT-4o still notably better than open models



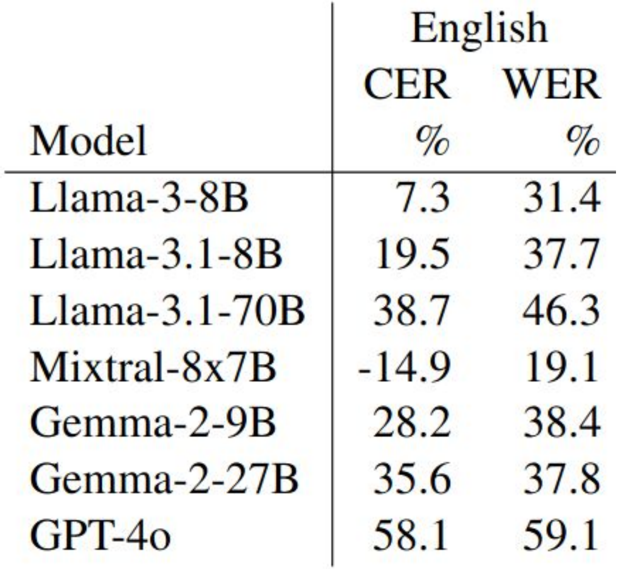
	English	
	CER	WER
Model	%	%
Llama-3-8B	7.3	31.4
Llama-3.1-8B	19.5	37.7
Llama-3.1-70B	38.7	46.3
Mixtral-8x7B	-14.9	19.1
Gemma-2-9B	28.2	38.4
Gemma-2-27B	35.6	37.8
➡ GPT-4o	58.1	59.1

→ In general, clear improvement can be obtained

Results

English

- Most models positive on both metrics
- Llama 3.1 70B best open model
- GPT-4o still notably better than open models



	English	
	CER	WER
Model	%	%
Llama-3-8B	7.3	31.4
Llama-3.1-8B	19.5	37.7
Llama-3.1-70B	38.7	46.3
Mixtral-8x7B	-14.9	19.1
Gemma-2-9B	28.2	38.4
Gemma-2-27B	35.6	37.8
GPT-4o	58.1	59.1

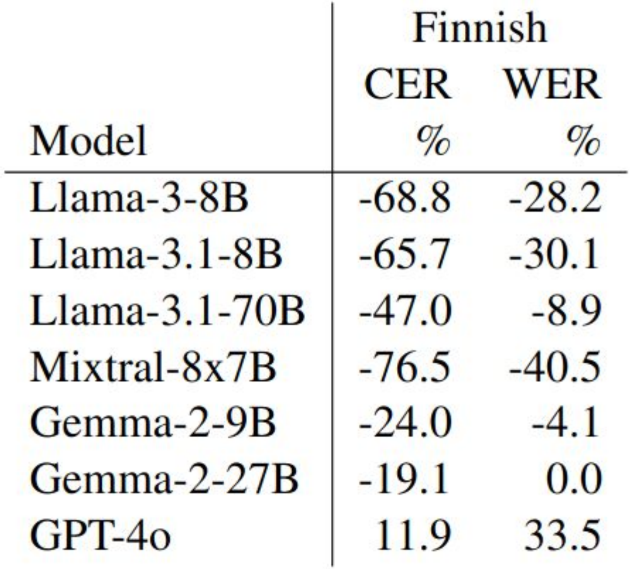
➡ In general, clear improvement can be obtained

Results

Finnish

➡ Most models negative on both metrics

- Gemma 2 27B least worse, but still negative
- GPT-4o positive on both metrics, but less so compared to English



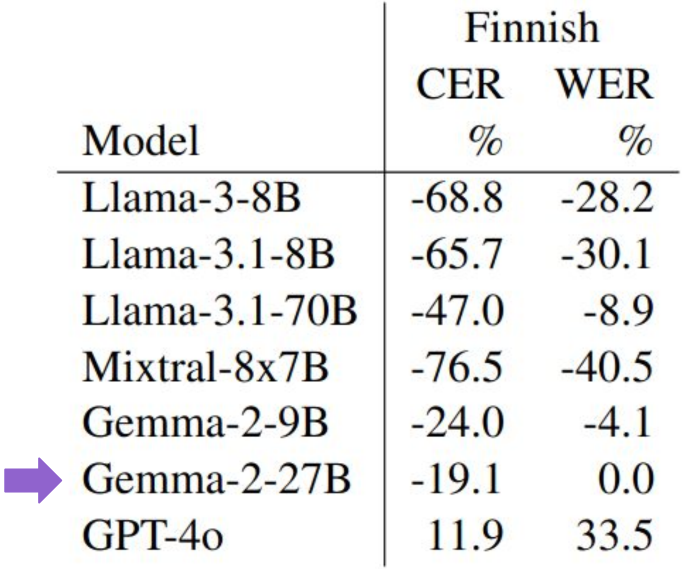
	Finnish	
	CER	WER
Model	%	%
Llama-3-8B	-68.8	-28.2
Llama-3.1-8B	-65.7	-30.1
Llama-3.1-70B	-47.0	-8.9
Mixtral-8x7B	-76.5	-40.5
Gemma-2-9B	-24.0	-4.1
Gemma-2-27B	-19.1	0.0
GPT-4o	11.9	33.5

→ zero-shot post-correction currently out of reach for historical Finnish

Results

Finnish

- Most models negative on both metrics
- ➡ Gemma 2 27B least worse, but still negative
- GPT-4o positive on both metrics, but less so compared to English



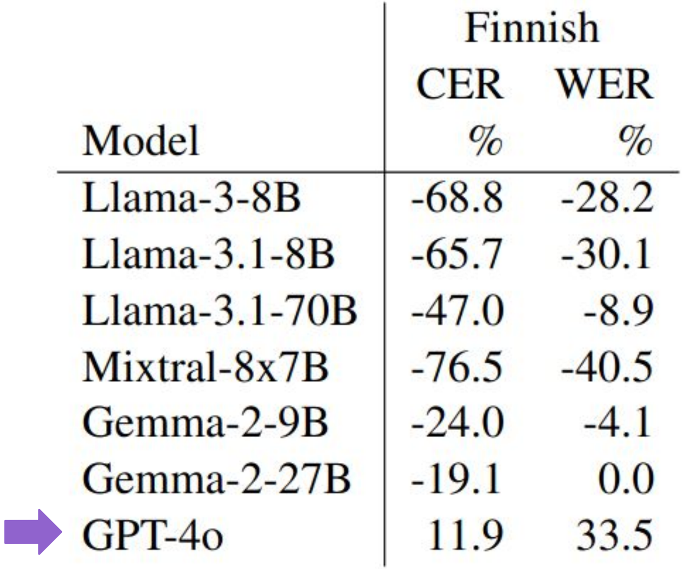
	Finnish	
	CER	WER
Model	%	%
Llama-3-8B	-68.8	-28.2
Llama-3.1-8B	-65.7	-30.1
Llama-3.1-70B	-47.0	-8.9
Mixtral-8x7B	-76.5	-40.5
Gemma-2-9B	-24.0	-4.1
➡ Gemma-2-27B	-19.1	0.0
GPT-4o	11.9	33.5

→ zero-shot post-correction currently out of reach for historical Finnish

Results

Finnish

- Most models negative on both metrics
- Gemma 2 27B least worse, but still negative
- ➡ GPT-4o positive on both metrics, but less so compared to English



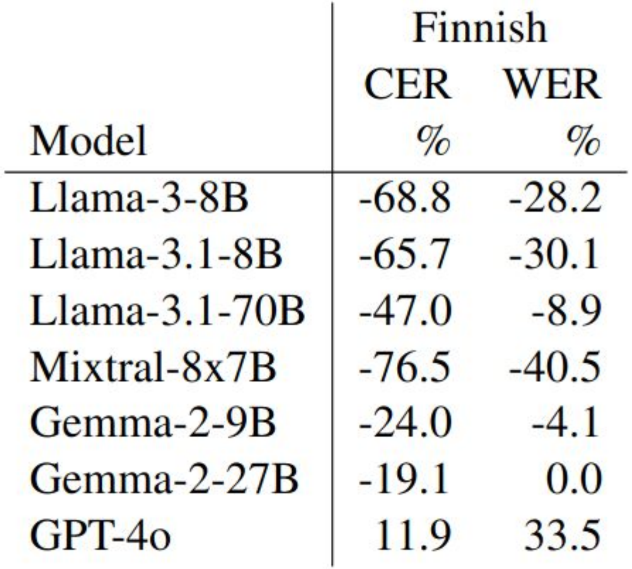
	Finnish	
	CER	WER
Model	%	%
Llama-3-8B	-68.8	-28.2
Llama-3.1-8B	-65.7	-30.1
Llama-3.1-70B	-47.0	-8.9
Mixtral-8x7B	-76.5	-40.5
Gemma-2-9B	-24.0	-4.1
Gemma-2-27B	-19.1	0.0
➡ GPT-4o	11.9	33.5

→ zero-shot post-correction currently out of reach for historical Finnish

Results

Finnish

- Most models negative on both metrics
- Gemma 2 27B least worse, but still negative
- GPT-4o positive on both metrics, but less so compared to English



	Finnish	
	CER	WER
Model	%	%
Llama-3-8B	-68.8	-28.2
Llama-3.1-8B	-65.7	-30.1
Llama-3.1-70B	-47.0	-8.9
Mixtral-8x7B	-76.5	-40.5
Gemma-2-9B	-24.0	-4.1
Gemma-2-27B	-19.1	0.0
GPT-4o	11.9	33.5

➡ zero-shot post-correction currently out of reach for historical Finnish

Summary of the feature ablation studies

Effect of post-processing:

- Depends on the model, varies between *no effect* and *must do*!

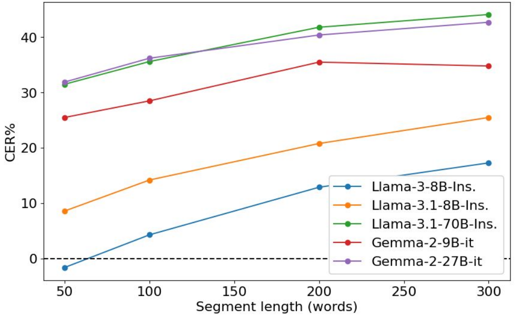
Effect of quantisation:

- Unquantized models (fp16) slightly better (0–4.5%)
  - GPU memory requirements increase from 43GB to 132GB (Llama 3.1 70B)

Summary of the feature ablation studies

Effect of segment length:

- Degrades notably if the segment is too short!
  - Not enough long documents to conclude the maximum length (page-level data)



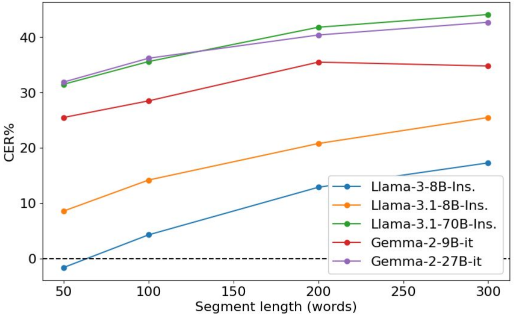
Summary of the feature ablation studies

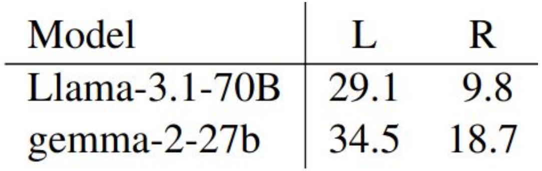
Effect of segment length:

- Degrades notably if the segment is too short!
  - Not enough long documents to conclude the maximum length (page-level data)

Effect on segment boundary:

- Performance worse on the right-side of the boundary (previous context missing!)
  - Attempts to account for this yielded mixed results





Model	L	R
Llama-3.1-70B	29.1	9.8
gemma-2-27b	34.5	18.7

CER% around segment boundary

Conclusions

- LLMs can be utilized to OCR post-correct historical English
  - Best open model: 38.7% relative CER improvement (Llama 3.1 70B)
  - GPT-4o: 58.1% relative CER improvement
- For Finnish, poor performance with open models
- Details matter (post-processing, segment length etc.)
- Evaluation is not straightforward

Future work

- Apply the best open model to correct the full ECCO OCR
  - 180,000 books
  - LLM correction run done with Llama 3.3 70B and LUMI supercomputer
  - Evaluation on-going
- For Finnish, fine-tune an LLM for the task?
  - Or wait for better LLMs?

Thank you!

TURKUNLP
.ORG

UNIVERSITY
OF TURKU