Can LLMs outperform the classical OCR/HTR tools?

II-Meeting 08/9bre/2024

Seorin Kim

Illustration by The Atlantic. Source: Getty.

Objectives

To digitize the handwritten historical records

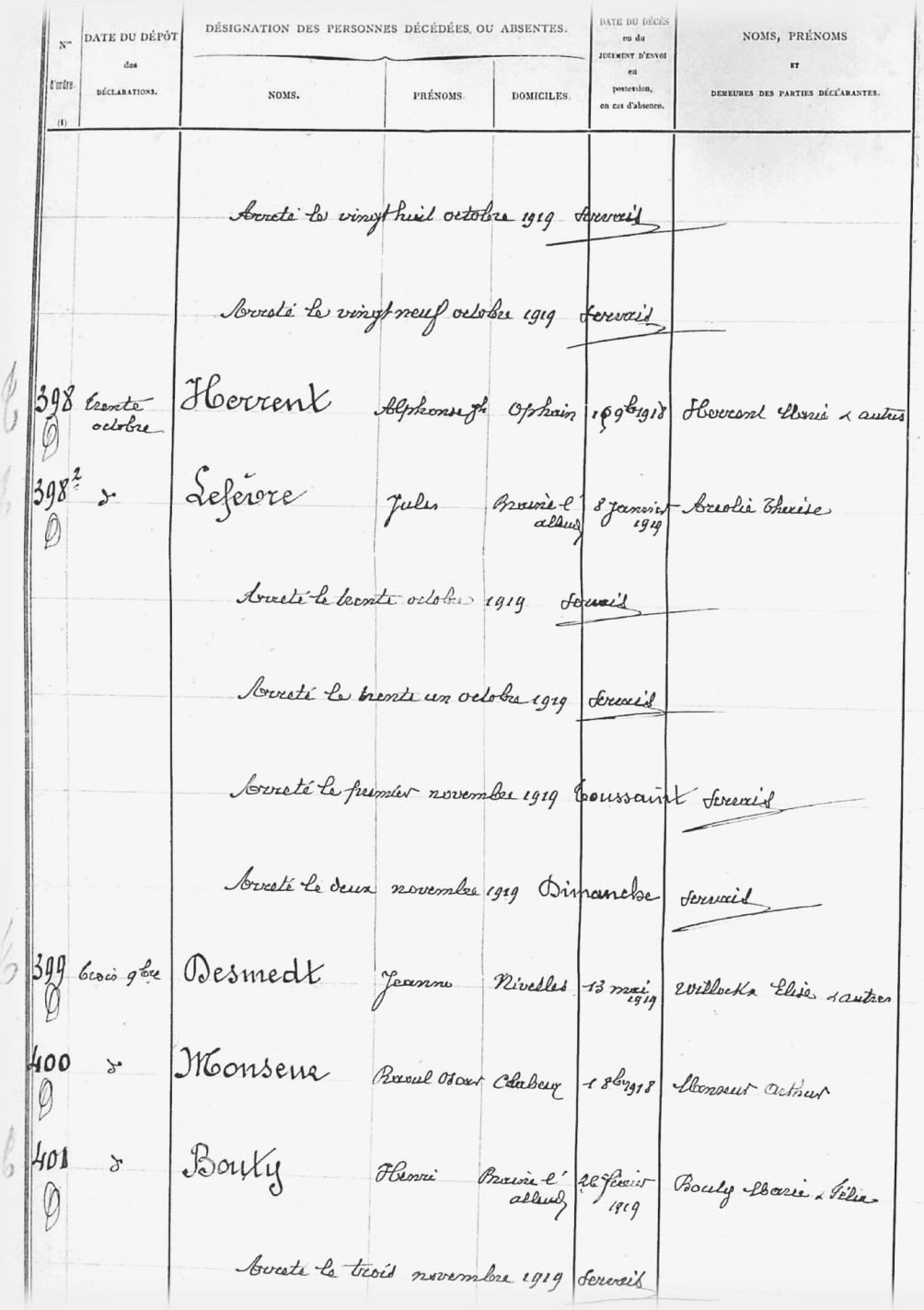
With LLMs, we want to achieve:

- Compatible results to the classical OCR/HTR pipelines
- Less ground truth (*GT = expensive!)

Historical Records

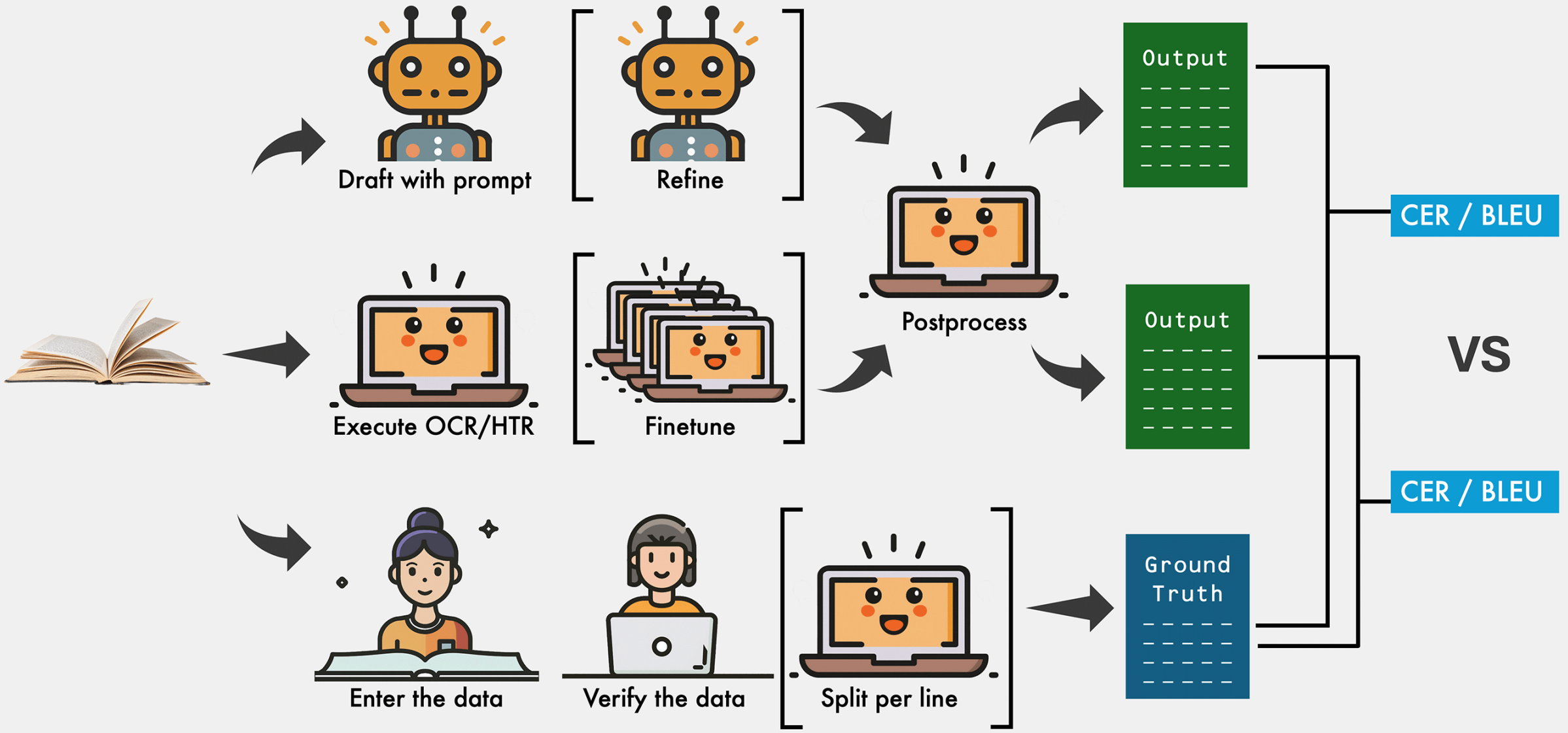
Challenges:

- Layout
- Mismatch: Content vs. Column names
  - E.g., memo
- Old symbols
- Numbers are hard to read



N° d'ordre (1)	DATE DU DÉPÔT des DÉCLARATIONS.	DÉSIGNATION DES PERSONNES DÉCÉDÉES OU ABSENTES. NOMS.	PRÉNOMS.	DOMICILES.	DATE DU DÉCÈS ou du JUGEMENT D'ENVOI en possession, en cas d'absence.	NOMS, PRÉNOMS ET DEMEURES DES PARTIES DÉCLARANTES.
		Arrêté le vingt huit octobre 1919 Service				
		Arrêté le vingt neuf octobre 1919 Service				
398 D	trente octobre	Herrent	Alphonse Jh	Ophain	19 Oct 1918	Herrent Henri + autres
398 bis D	id	Leferre	Jules	Braine l'Alleud	8 janvier 1919	Arnold Therese
		Arrêté le trente octobre 1919 Service				
		Arrêté le trente un octobre 1919 Service				
		Arrêté le premier novembre 1919 Toussaint Service				
		Arrêté le deux novembre 1919 Dimanche Service				
399 D	trois 9bre	Desmedt	Jeanne	Nivelles	13 mai 1919	Willocks Elisa + autres
400 D	id	Monsene	Raoul Oscar	Clabecq	18 Oct 1918	Monsene Arthur
401 D	id	Bouky	Henri	Braine l'Alleud	28 février 1919	Bouky Henri + Félix
		Arrêté le trois novembre 1919 Service				

Workflow


Draft with prompt
Refine
Output
Execute OCR/HTR
Finetune
Postprocess
Output
CER / BLEU
VS
CER / BLEU
Enter the data
Verify the data
Split per line
Ground Truth

Compared Models

LLMs:

- GPT 4o
- Claude Sonnet 3.5

OCRs:

- EasyOCR
- Keras OCR
- Pytesseract OCR
- TrOCR

Compared Types

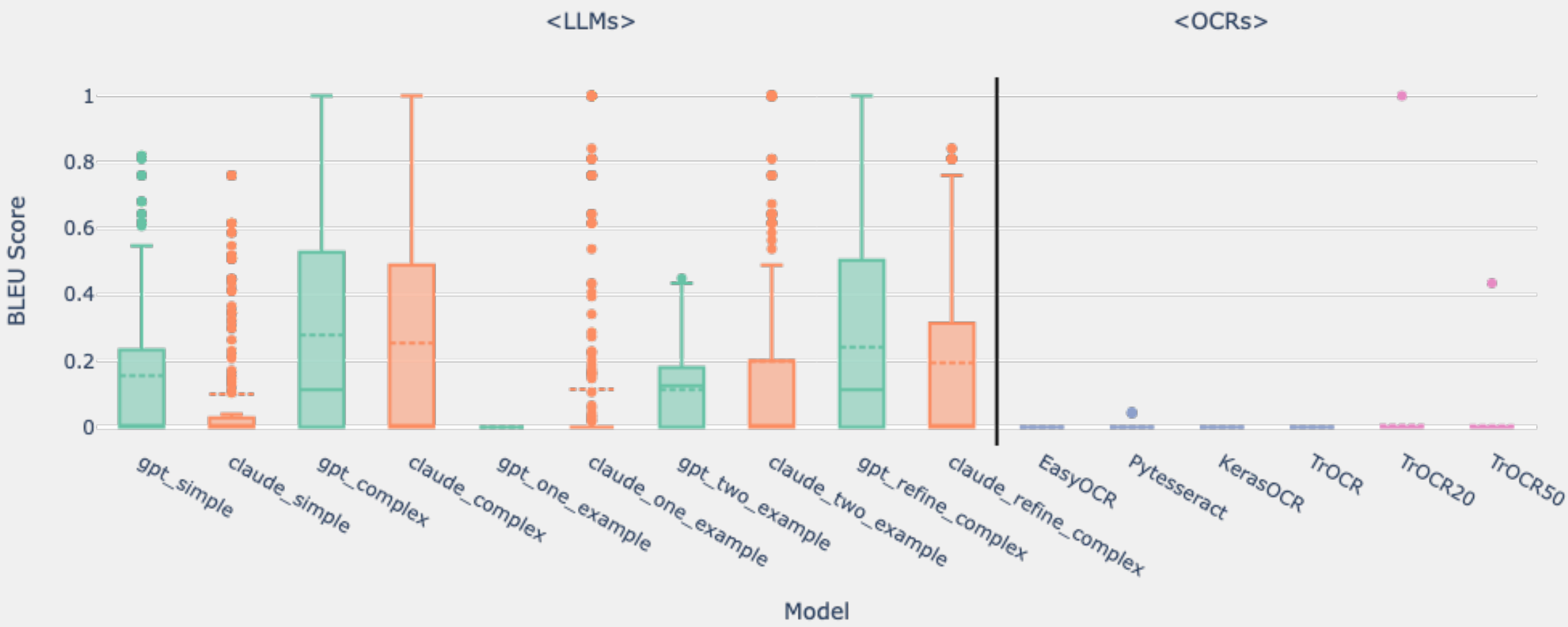
LLMs

(Zero-shot) Prompting?	• Simple Prompt • Complex Prompt
#Examples?	• One example • Two examples
Refine?	• Refine with the complex prompt output

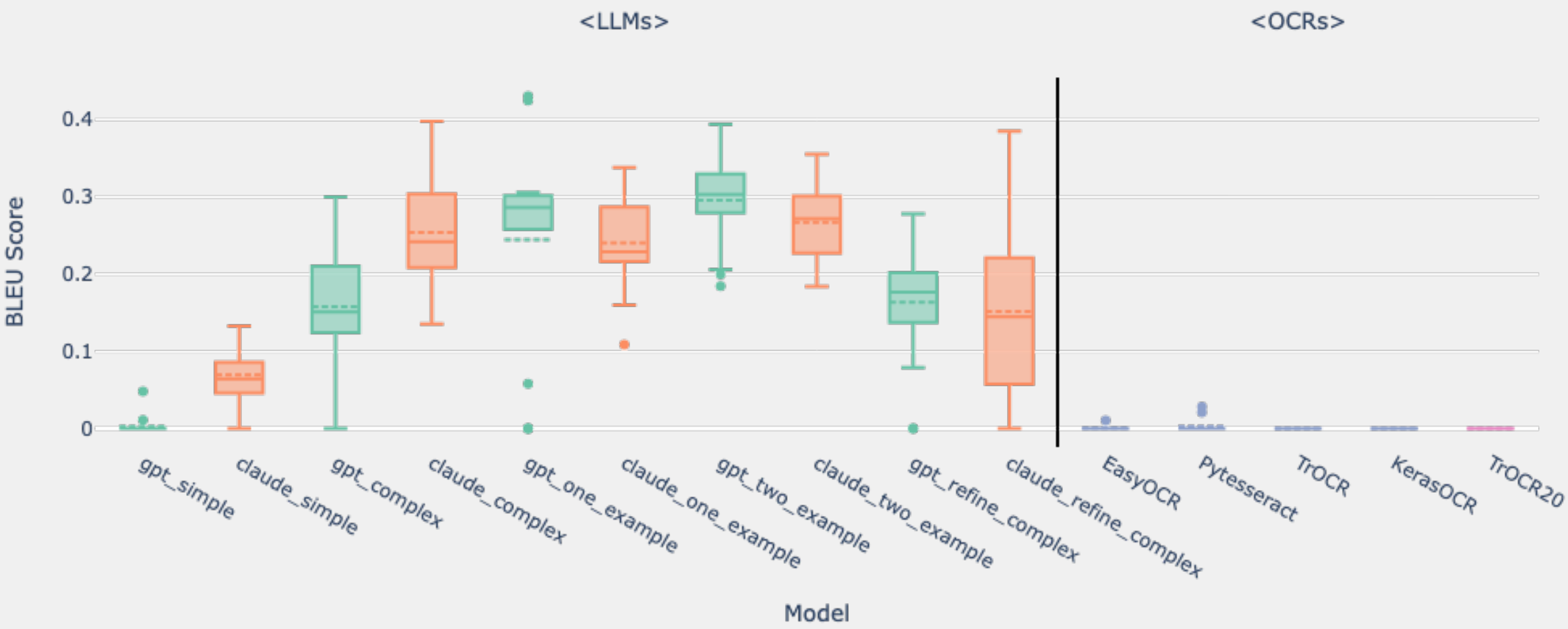
OCRs

~Zeroshot?	• Without finetuning
Conventional use	• Finetuning with 20% data (6^{th} epoch) • Finetuning with 50% data (6^{th} epoch)

BLEU Scores – line by line



BLEU Scores – whole scans

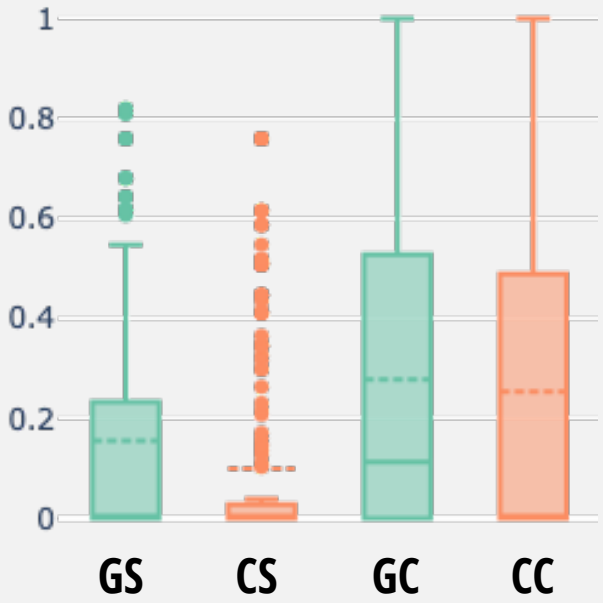


Are the results significantly different?

*BLEU Scores

Simple vs. Complex Prompts

		Line-by-line	
Model 1	**Model 2**	t-stat	p-value
GPT Complex	Claude Complex	1.307	0.192
GPT Simple	Claude Simple	4.088	0.000
GPT Complex	GPT Simple	7.217	0.000
Claude Complex	Claude Simple	8.596	0.000
GPT Complex	Claude Simple	9.057	0.000
Claude Complex	GPT Simple	4.709	0.000

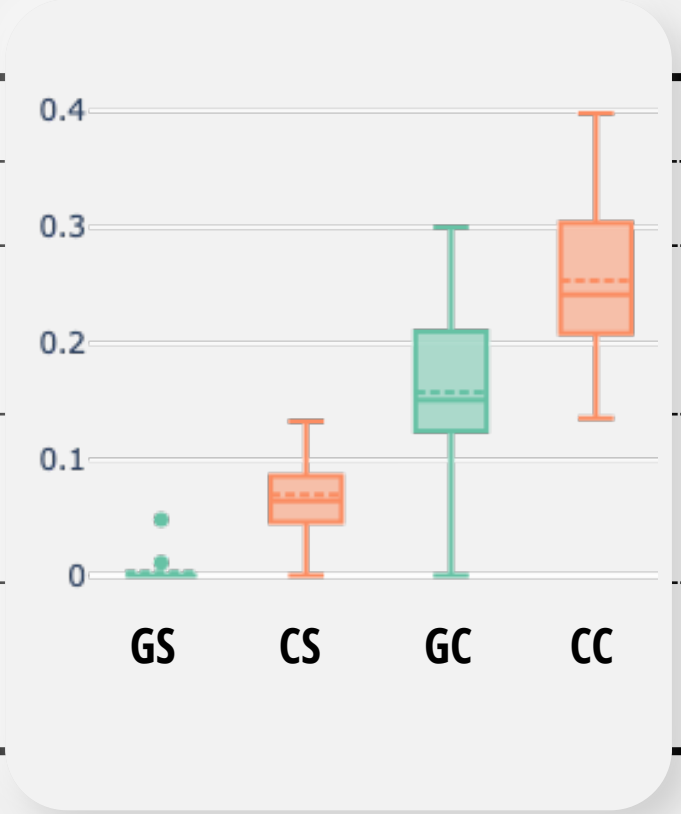


Are the results significantly different?

*BLEU Scores

Simple vs. Complex Prompts

		Whole Scans	
Model 1	**Model 2**	t-stat	p-value
GPT Complex	Claude Complex	4.532	0.000
GPT Simple	Claude Simple	7.256	0.000
GPT Complex	GPT Simple	8.870	0.000
Claude Complex	Claude Simple	12.255	0.000
GPT Complex	Claude Simple	5.708	0.000
Claude Complex	GPT Simple	18.086	0.000



Are the results significantly different?

*BLEU Scores

Simple vs. Complex Prompts

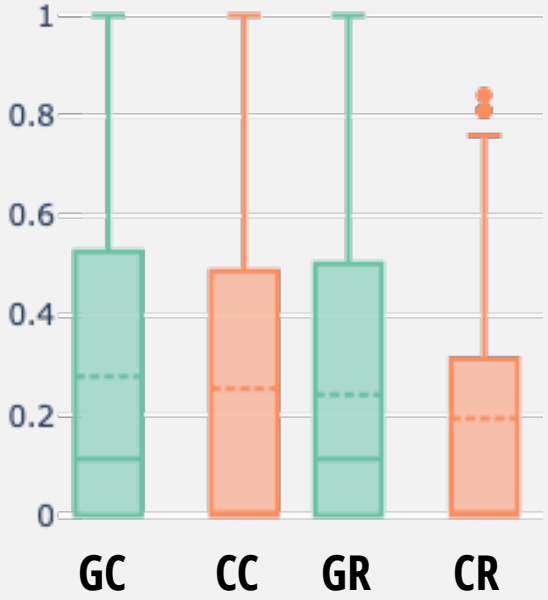
		Line-by-line		Whole Scans	
Model 1	Model 2	t-stat	p-value	t-stat	p-value
GPT Complex	Claude Complex	1.307	0.192	4.532	0.000
GPT Simple	Claude Simple	4.088	0.000	7.256	0.000
GPT Complex	GPT Simple	7.217	0.000	8.870	0.000
Claude Complex	Claude Simple	8.596	0.000	12.255	0.000
GPT Complex	Claude Simple	9.057	0.000	5.708	0.000
Claude Complex	GPT Simple	4.709	0.000	18.086	0.000

Are the results significantly different?

*BLEU Scores

Complex vs. Refine Complex

		Line-by-line	
Model 1	**Model 2**	t-stat	p-value
GPT Complex	Claude Complex	1.307	0.192
GPT Refine	Claude Refine	2.705	0.007
GPT Complex	GPT Refine	5.127	0.000
Claude Complex	Claude Refine	6.637	0.000
GPT Complex	Claude Refine	4.501	0.000
Claude Complex	GPT Refine	0.657	0.511

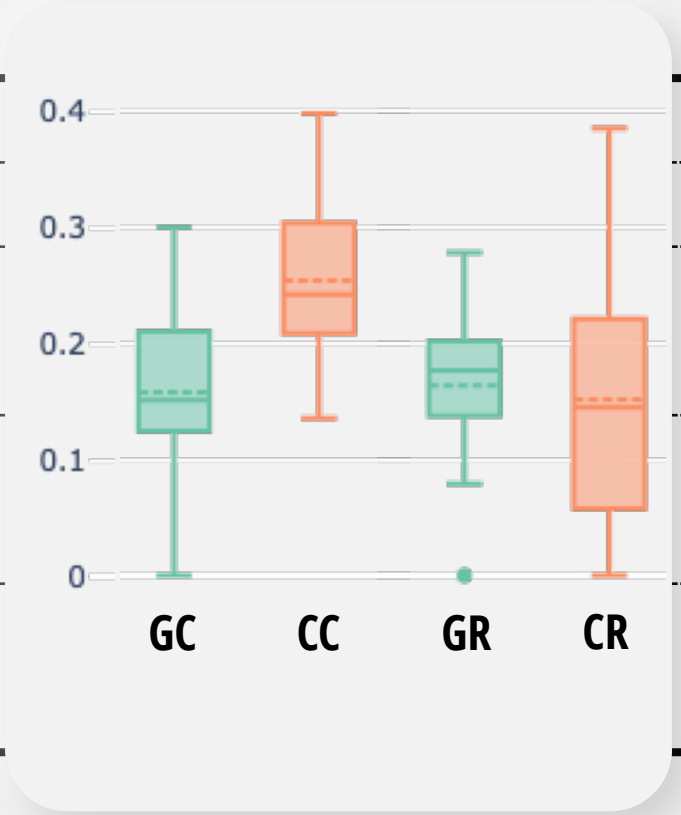


Are the results significantly different?

*BLEU Scores

Complex vs. Refine Complex

Model 1	Model 2	Whole Scans	
		t-stat	p-value
GPT Complex	Claude Complex	4.532	0.000
GPT Refine	Claude Refine	0.484	0.634
GPT Complex	GPT Refine	0.315	0.756
Claude Complex	Claude Refine	5.032	0.000
GPT Complex	Claude Refine	0.229	0.821
Claude Complex	GPT Refine	3.999	0.001



Are the results significantly different?

*BLEU Scores

Complex vs. Refine Complex

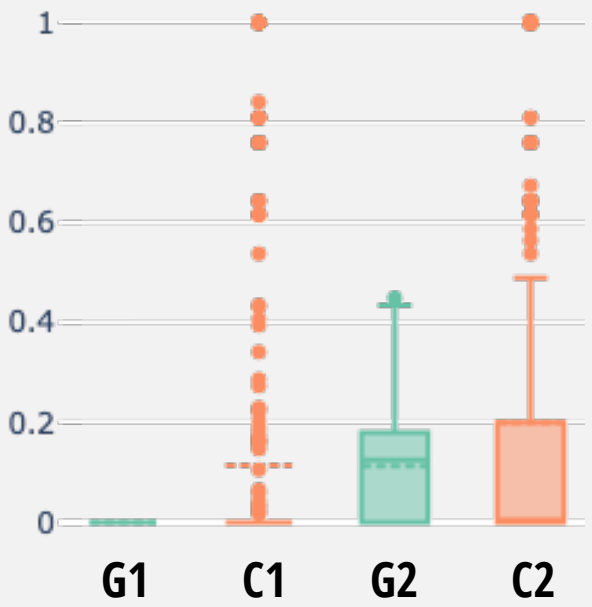
		Line-by-line		Whole Scans	
Model 1	**Model 2**	t-stat	p-value	t-stat	p-value
GPT Complex	Claude Complex	1.307	0.192	4.532	0.000
GPT Refine	Claude Refine	2.705	0.007	0.484	0.634
GPT Complex	GPT Refine	5.127	0.000	0.315	0.756
Claude Complex	Claude Refine	6.637	0.000	5.032	0.000
GPT Complex	Claude Refine	4.501	0.000	0.229	0.821
Claude Complex	GPT Refine	0.657	0.511	3.999	0.001

Are the results significantly different?

*BLEU Scores

One shot vs. Two shots

		Line-by-line	
Model 1	**Model 2**	t-stat	p-value
GPT One	Claude One	7.075	0.000
GPT Two	Claude Two	4.351	0.000
GPT One	GPT Two	18.948	0.000
Claude One	Claude Two	4.420	0.000
GPT One	Claude Two	9.489	0.000
Claude One	GPT Two	0.153	0.878

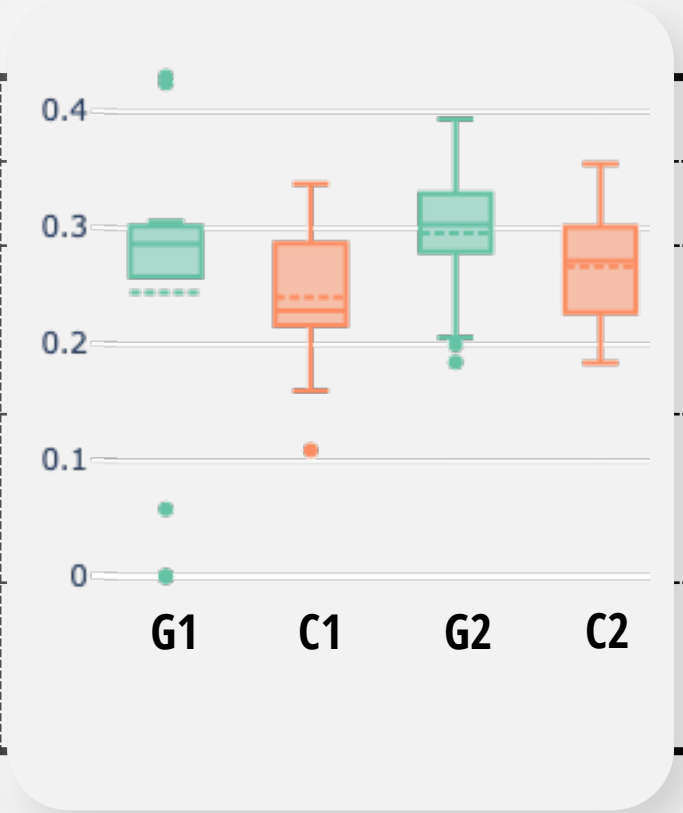


Are the results significantly different?

*BLEU Scores

One shot vs. Two shots

Model 1	Model 2	Whole Scans	
		t-stat	p-value
GPT One	Claude One	0.138	0.891
GPT Two	Claude Two	2.408	0.028
GPT One	GPT Two	1.544	0.141
Claude One	Claude Two	1.608	0.126
GPT One	Claude Two	0.745	0.467
Claude One	GPT Two	2.656	0.017



Are the results significantly different?

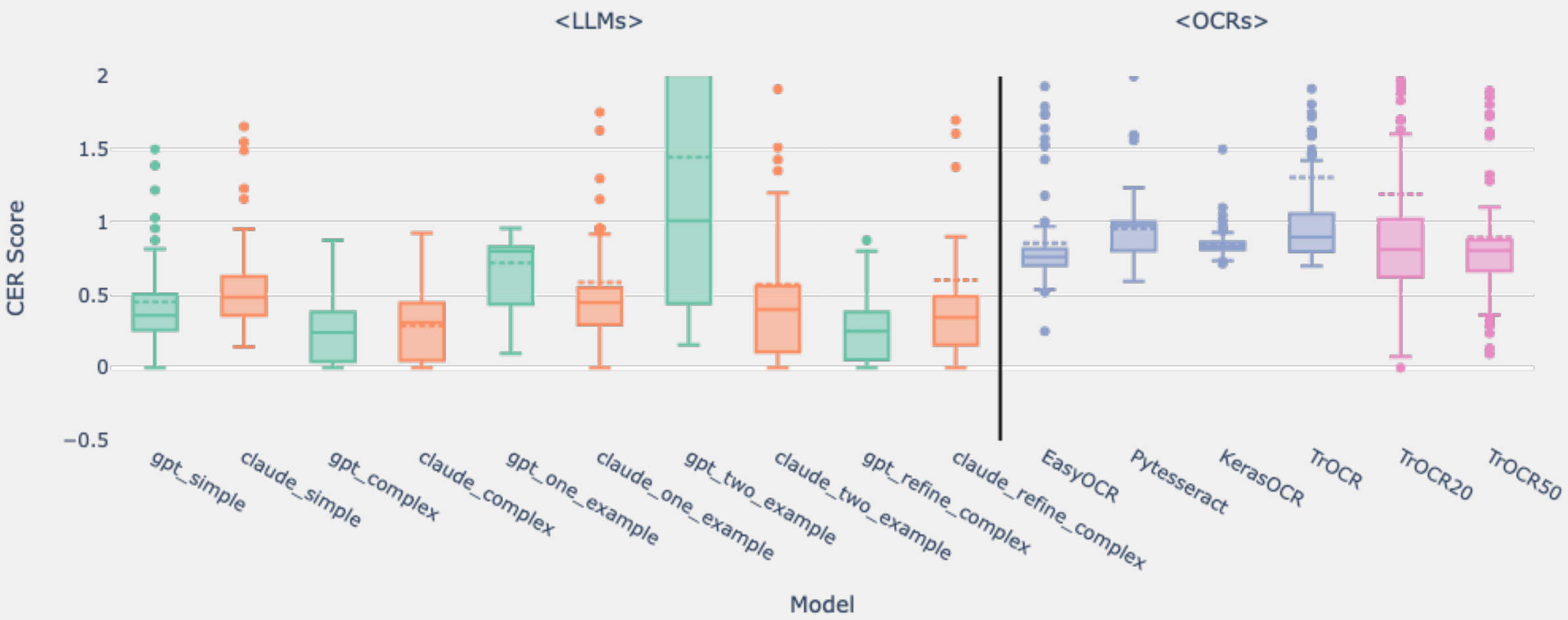
*BLEU Scores

One shot vs. Two shots

		Line-by-line		Whole Scans	
Model 1	**Model 2**	t-stat	p-value	t-stat	p-value
GPT One	Claude One	7.075	0.000	0.138	0.891
GPT Two	Claude Two	4.351	0.000	2.408	0.028
GPT One	GPT Two	18.948	0.000	1.544	0.141
Claude One	Claude Two	4.420	0.000	1.608	0.126
GPT One	Claude Two	9.489	0.000	0.745	0.467
Claude One	GPT Two	0.153	0.878	2.656	0.017

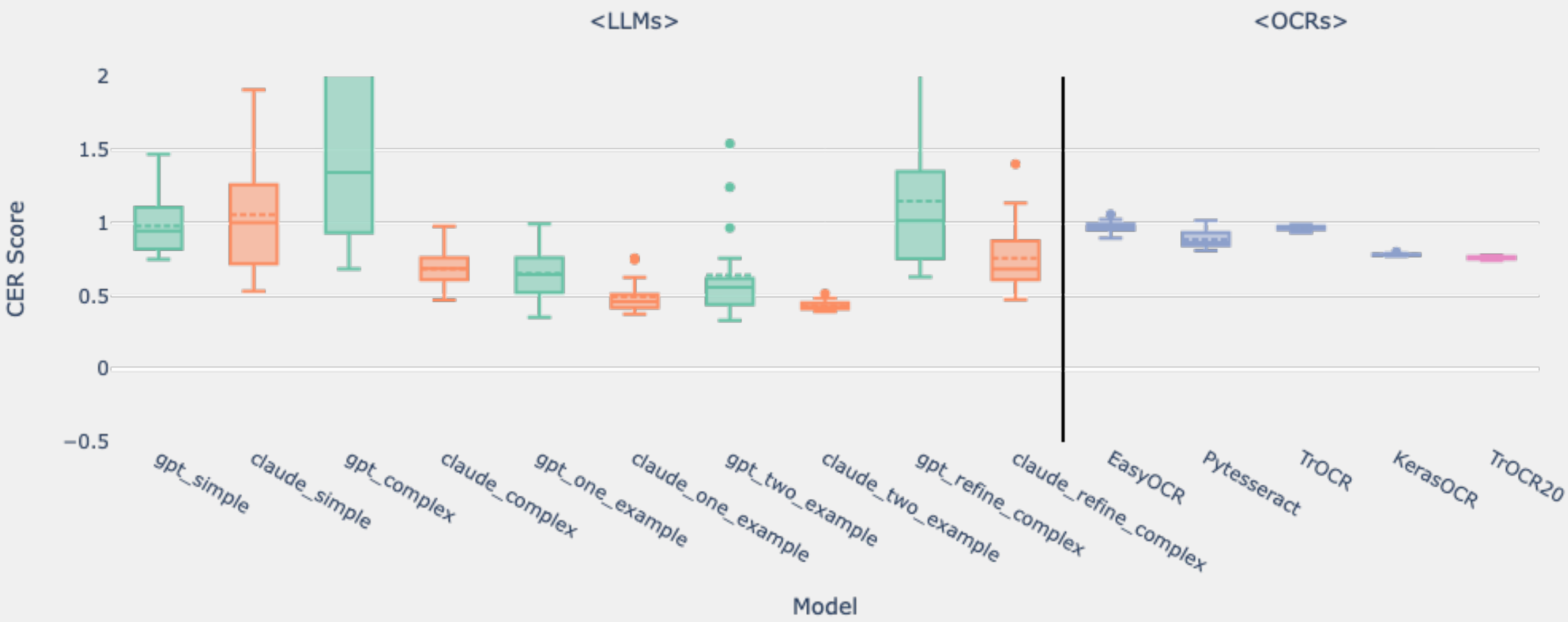
CER – line by line

Zoomed at [-0.5, 2]



CER – whole scans

Zoomed at [-0.5, 2]



BLEU vs. CER

[0,1] [0,+∞]

Line	BLEU	CER
9	0.341	0.300
10	0.000	0.547
9	0.000	0.815
10	0.000	0.842

BLEU is more conservative than CER.

*The use of one reference list leads to smaller BLEU.

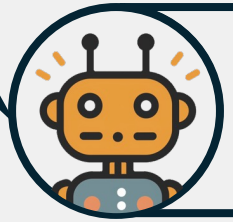
Document 1: line 9 and 10

GT

399 trois 9bre Desmedt Jeanne Nivelles 13 mai 1919 Willock Elise & autres 9480 530 8950 15 Db 1919 18 mars 1921 10 février 1920 39

400 d Monseur Raoul Oscar Clabecq 1 8b 1918 Monseur Arthur 69051 31417 659093 15 d 1 août 1919

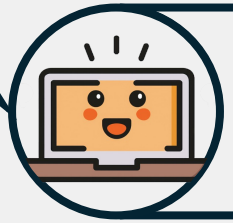
Claude 1 Ex.



399 trois 9bre Desmedt Jeanne Nivelles 13 mai 1919 Célibataire sans profession 9410 520 39_10 15 3/4 _919 13 mars 1920 10 février _21 _

400 _ Monseur Pascal Henri Célestin 1 8bre 1848 Receveur Débitant 69060 34478 34582 15 32 4 avril 1919

TrOCR50



Arrêté le vingt novembre 1919 Dimanche servais

Arrêté le vingt quatre novembre 1919 Dimanche servais

BLEU vs. CER

[0,1] [0,+∞]

Line	BLEU	CER
3	0.809	0.024
4	0.000	0.195
3	0.000	0.095
4	0.000	0.073

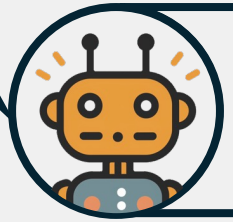
Document 9: line 3 and 4

GT

Arrêté le vingt cinq novembre 1919 servais

Arrêté le vingt six novembre 1919 servais

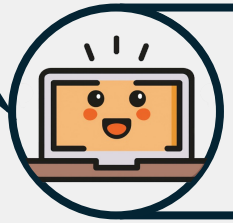
Claude 1 Ex.



Arrêté le vingt cinq novembre 1919 Servais

Arrêté le vingt huit octobre 1919 Servais

TrOCR50



Arrêté le vingt sept novembre 1919 servais

Arrêté le vingt sept novembre 1919 servais

OCRs

- Fine-tuning dependent
  - Without finetuning, no comprehensible outputs
  - With only 20% and 50% of data finetuned, the outputs are often repetitions of the trained data

LLMs

- Easy to use
  - With only a few examples, the quality increases significantly
  - No need layout analysis
  - No finetuning required
- line-by-line **>** whole scans
  - Keeps the same #rows
  - Layout analysis required?
- Whole **>** line-by-line
  - It understands the context better

CER Scores [0,+∞]

$$CER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

S = #Substitutions, D = #Deletions, I = #Insertions, C = #Correct Characters,

N = #Characters in the references (N=S+D+C)

- Not always [0,1] , especially in case of a high number of insertions
- Often associated to the % of characters that were incorrectly predicted.
- The lower the value, the better the performance (CER == 0 ⇔ Perfect)

CER Scores [0,+∞]

$$CER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

Reference: Cat , **Substitution**: Bat , **Insertion**: Cats , **Deletion**: _at

Candidate	Reference	CER
"hello world"	“hello”	I=6, N=5, CER=1.2
"this is the prediction"	"this is the reference"	S=6 N=21, CER=0.29
"there is another sample"	"there is another one"	S=3, I=7, D=2, N=20, CER=0.6

⚠️ *Careful with white spaces.*

BLEU* Scores [0,1]

*Bilingual Evaluation Understudy (Papineni et al. 2002)

① **Modified N-gram**: Candidate vs. Reference(s)
- Modified such that it punishes the random repetition of one or a few words of the reference in the candidate

② To combine the modified precisions for the various N-gram sizes: **geometric mean**
- Because the precision exponentially decays with the increase in N

③ Sentence brevity penalty (**BP**)
- E.g., Candidate: "**of the**" vs. Ref: "It is [...] the command **of the** Party" ➔ Modified n-gram precision == 1 (2/2 unigram, 1/1 bigram)
- (1) already punishes sentences longer than the refs
- BP == 1, if when the lengths are the same between the candidate (c) and ref (r)
- If c≤r, BP == $e^{1-\frac{r}{c}}$

BLEU Scores [0,1]

$$BLEU = \underbrace{BP}_{③} \cdot \underbrace{\exp\left(\sum_{n=1}^{N} w_n \log \underbrace{p_n}_{①}\right)}_{②}$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases}$$

$w_n = 1/N$ (*Default N == 4)

BLEU

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Document 9: line 4

GT

Arrêté le vingt six novembre 1919 servais

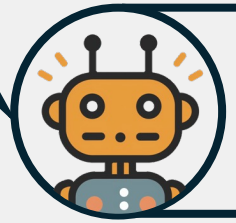
Precisions (BP=1)

[0.57, 0.33, 0.2, 0.0]

⇔

[4/7, 2/6, 1/5, 0/4]

Claude
1 Ex.



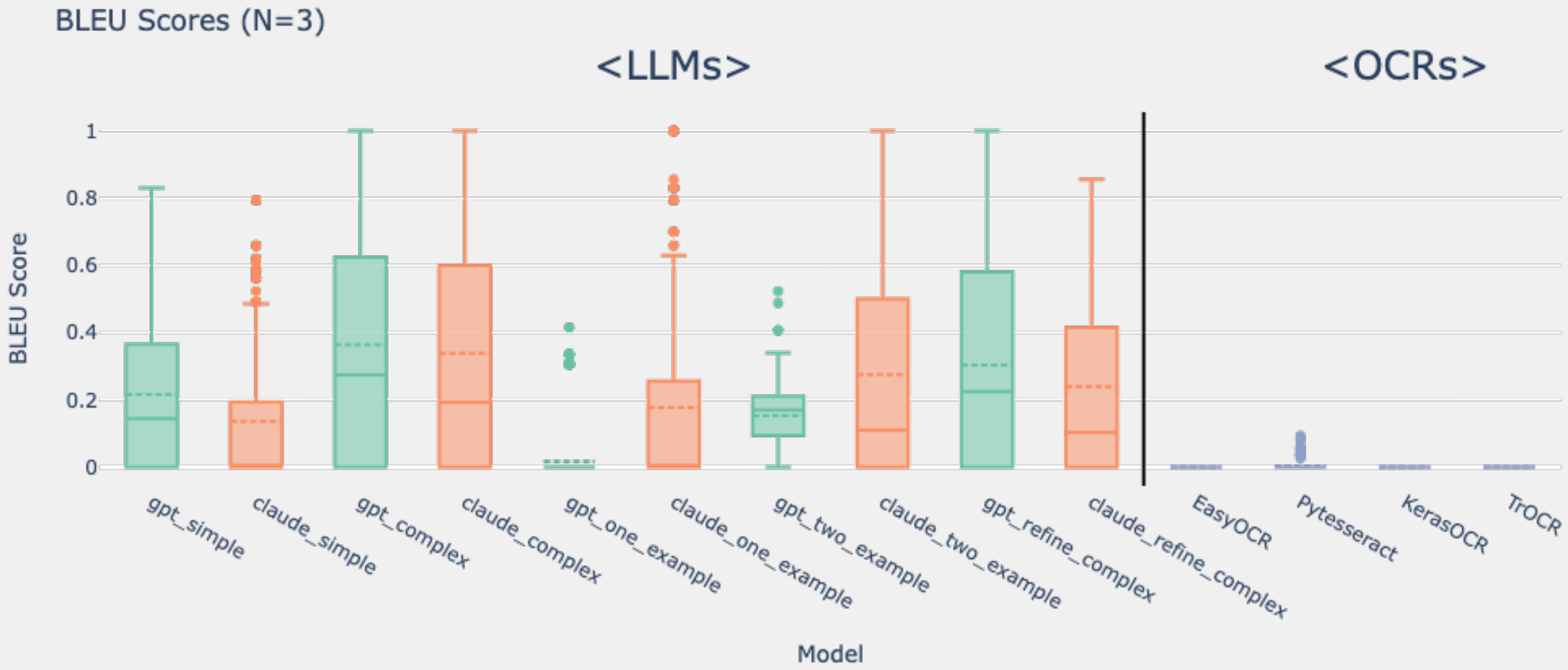
Arrêté le vingt huit octobre 1919 Servais

BLEU: 1*exp(1/3*(log(0.57)+log(0.33)+log(0.2)+log(0.0)) **== 0**

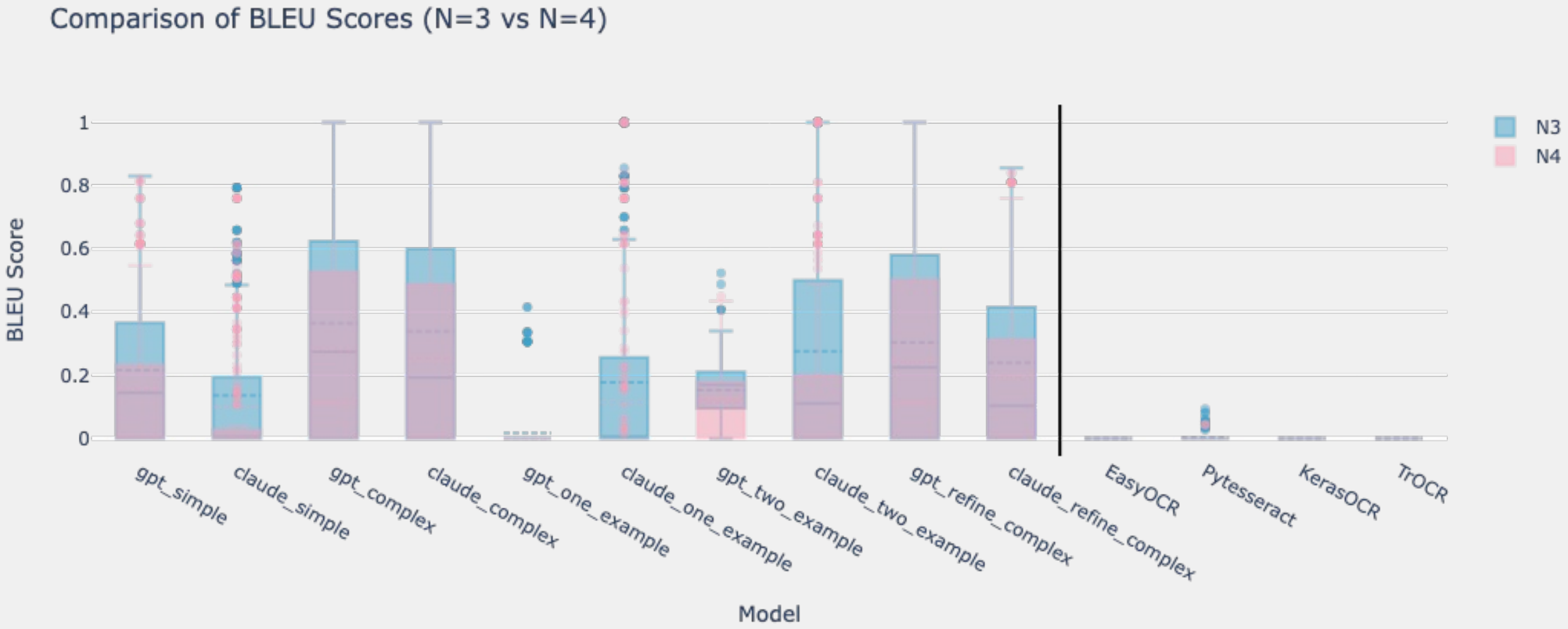
0.34

BLEU N=3	BLEU N=4	CER
0.336	0.000	0.195

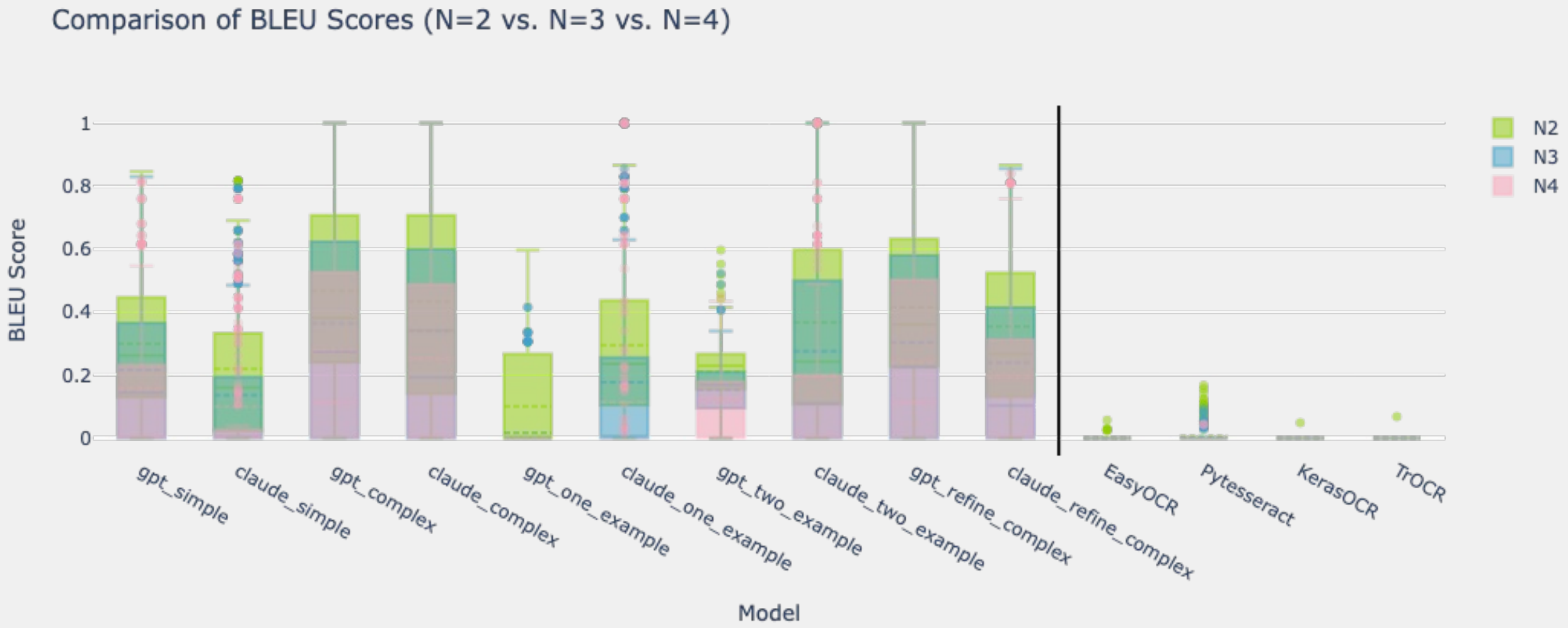
BLEU Scores – line by line



BLEU Scores – line by line



BLEU Scores – line by line



BLEU Scores vs. Human Evaluations