Multilingual NLP Homework 2

Benucci Lorenzo
Università "La Sapienza"/ Roma
benucci.2219690@studenti.uniroma1.it

D'Addario Giuseppe
Università "La Sapienza"/ Roma
daddario.2177530@studenti.uniroma1.it

Abstract

This study addresses the task of Optical Character Recognition (OCR) post-correction, converting noisy OCRed text into clean output using both large and small LLMs. Performances are evaluated with the paradigm known as LLM-as-a-Judge, and correlation metrics between human and LLM-generated scores are computed to assess the reliability of automatic evaluation.

1 Introduction

Automatically correcting OCR-generated text is a non-trivial challenge, especially when dealing with sources where noise patterns are irregular or unpredictable. Hence, OCR post-correction solutions, such as using generative models to rewrite the sentences without errors, are required.

2 Dataset

For this project the text of "The Vampyre", by Polidori, was used, split in indexes from 0 to 47. The file named "the_vampyre" contains the text with synthetic OCR errors, which needs to be corrected, while the file "the_vampyre_clean" contains the clean version, representing the gold correction. An example is shown in [Tab: 2]. Two other datasets were used for finetuning purposes, namely "LIMA" (Less is More for Alignment)[1], and "Post-OCR Correction"[2]

3 Translation [OCRed→ Clean]

In order to have a wide view, five different LLMs were tested to analyze how correction quality depends on the number of parameters. Specifically, the question is: are Llama's $17B$ parameters truly required for this task?

3.1 t5 + SpellChecker

As a small correction model, the pretrained and finetuned t5 (seq2seq) model[3], enhanced with a custom pre-processing pipeline was chosen. After a manual correction of common OCR errors, a two-step segmentation, imposed by the model's tokenizer limitations, was carried out: first, texts into sentences, then sentences into smaller chunks, with a maximum length (100 tokens). Since OCR-post-correction involves lexicon, typography, syntax and orthography, model inputs were further refined with a spell checker to help address orthographical errors.

3.2 Minerva

Three different versions of Minerva were used: the base model, the model finetuned on the LIMA dataset and the same model finetuned also on the Post-OCR Correction dataset. Both finetunings were carried out on the Leonardo Supercomputer. The reason behind the first finetuning was to improve Minerva's abilities to stick to the prompt, while the second one aimed to improve its capability in OCR error detection and correction.

3.3 Llama4

As an additional model, LLaMA-4-Scout-17B-16E-Instruct was tested out, using the public API for the generation of the corrected sentences. This model, having $17B$ parameters, is the largest one used in the project.

4 Evaluation

4.1 Human and Machine annotation

To evaluate the quality of the corrections, a parallel procedure was adopted: first, a human evaluation of the corrections was carried out; then, two different LLMs -Gemini and Prometheus- were asked to

[1] https://huggingface.co/datasets/GAIR/lima
[2] https://huggingface.co/datasets/PleIAs/Post-OCR-Correction
[3] https://huggingface.co/yelpfeast/byt5-base-english-ocr-correction

perform the same evaluation, relying on the same set of criteria, shown in the appendix [Tab: 1].
`gemini-1.5-flash-latest`: the evaluation was executed through the model's public API.
`M-Prometheus-7B`: the evaluation was executed on the boost partition of the Leonardo Supercomputer, and took more than one hour per model to be completed.

The final prompt, provided in appendix [A: 7.2], was obtained through experimentation on Prometheus, opting for a minimal but effective description of the scores. The reason was to avoid ambiguities due to over-explanations and allow the model to generalize better, reducing the variability of the evaluations and improving their consistency.

4.2 Human-Metrics correlation

After the quality evaluation, an analysis of human-metrics correlation was carried out, measuring the agreement between human annotations and the automated evaluation scores produced by the different language models. Two statistical measures were employed: Cohen's Kappa coefficient [A: 7.0.1], which assesses the level of agreement beyond chance between categorical ratings, and the standard accuracy. These metrics provide insight into how reliably the automated systems reflect human evaluation.

5 Results and Conclusions

5.1 Correction quality

Models showed markedly different performance, as seen by the *ROUGE* scores in [Tab: 4].
t5 + Spell Checker attains a relatively high Rouge, but low correlation with human judgments ($\rho \approx [0.23, 0.29]$), suggesting fluent outputs with lexical overlap. This means the model understands the syntactic but not the semantic, due to its small dimension and its limited tokenizer size. Very few outliers were observed. **Minerva** shows a high performance and a very strong correlation with human scores ($\rho = 0.705, p < 10^{-7}$), while **LLaMA4**, despite very high Rouge scores, has a very low correlation ($\rho \approx 0.04, p \approx 0.78$). This underlines how the model is very close to generating perfect outputs, but its range of improvement is still inside the second-last level.
The **LIMA-finetuned** model improves slightly in Rouge overlap scores but not in correlation. On the other hand, the **double fine-tuned** version (on Post-OCR) exhibits a noticeable decline in both Rouge and correlation, revealing a deterioration for this specific task. Furthermore, it's evident how bigger models imply more outliers. This behavior is shown in the appendix: [A:fig. 6-7]

5.2 Evaluation quality

For Evaluation quality, results are shown in the table [Tab: 3]. Focusing on the correction scores: for the **t5 + Spell Checker** model, results suggest a low agreement of both evaluators with human annotations.
For the **LLaMA4** model, there is a striking discrepancy in the evaluations[A:fig. 1]. This highlights that both Prometheus and Gemini appear unable to capture the subtle nuance between a perfect and an almost perfect output, and that the agreement is statistically random. This is in line with the correction performance of the model discussed before. **Minerva** model's C-K and Accuracy values are generally higher [A:fig. 3], indicating more reliable corrections and a good agreement with human scores.
Gemini may rely on surface-level features, such as typographical details or formatting errors (e.g. spaces before or after a comma), which cause, for LLaMA4 corrections, the distinction between scores 4 and 5 to be slightly skewed. This behavior may explain why Gemini aligns better with the surface-level corrections of t5, but fails to capture deeper semantic quality in more complex models.
Prometheus tends to overrate outputs overall but shows better agreement with human judgments when evaluating Minerva's corrections, indicating more reliable performance on that model.

6 Future works

The main problems observed in this project are linked to the discrepancy between human and LLM evaluation scores, as well as an insufficient correction performance to consider the task solved. For future work, in-context learning should be explored, to better align model corrections with human judgments, and also to generate better corrections. Furthermore, the LLaMA 4 model also underlined that the prompt is not effective at capturing subtle improvements near the optimal output. To address this, a more discretized scoring prompt could be adopted, incorporating a larger number of intermediate scores, either equidistant or not.

7 Appendix

7.0.1 A - Cohen's Kappa Coefficient

Cohen's Kappa coefficient (κ) quantifies the level of agreement between two raters who each classify items into mutually exclusive categories.

The formula is:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the **observed agreement**, and p_e is the **expected agreement** by chance. The resulting κ value ranges from -1 to 1 with: $\kappa = 1 \rightarrow$ perfect agreement, $\kappa = 0 \rightarrow$ agreement equivalent to chance, $\kappa < 0 \rightarrow$ agreement worse than chance. Unlike simple accuracy, κ accounts for the possibility of agreement occurring by chance, hence, in our case, κ offers a more robust assessment of the alignment.

7.1 A - Evaluation Score

Evaluation Score

Score	Description
1	Completely unacceptable. The output is severely incomplete or entirely incorrect.
2	Major issues. The output is complete but significantly inaccurate or misleading.
3	Some errors. Mostly correct, but with noticeable mistakes or altered meaning.
4	Minor issues. Largely accurate, with only small grammatical or factual deviations.
5	Perfect. No errors; completely faithful and correct.

Table 1: Criteria for the evaluation score. The term "factual deviations" in the score 4 comprises all those errors which change a fact but keep the meaning equal. (E.g. 'Johri iz my fother –> Jeorge is my father'. The fact is different, because John and Jeorge are different people, but the meaning of the sentence is the same.)

7.2 A - Prompt

Gemini/Prometheus prompt

Evaluate the quality of the [GENERATED] text in comparison to the [EXPECTED] text.

Use the following scale:
1 = Completely unacceptable. The output is severely incomplete or entirely incorrect.
2 = Major issues. The output is complete but significantly inaccurate or misleading.
3 = Some errors. Mostly correct, but with noticeable mistakes or altered meaning.
4 = Minor issues. Largely accurate, with only small grammatical or factual deviations.
5 = Perfect. No errors; completely faithful and correct.
Output only the score after the [NUMERIC SCORE] tag.
Format: [NUMERIC SCORE] <number>

Do not explain. Do not write anything else.

7.3 A - Dataset

Dataset Excerpt	
OCR	“Tbe world was alI beforehim, where to chaose his place of rest, and Providence his guide.”
Corrected	“The world was all before him, where to choose his place of rest, and Providence his guide.”

Table 2: A single-entry excerpt from the dataset

7.4 A - Metrics

Model	Metric	Gemini	Prometheus
T5 + Spell Checker	Acc.	0.583	0.396
	C-K	0.377	0.204
Minerva	Acc.	0.500	0.604
	C-K	0.303	0.486
LLaMA4	Acc.	0.438	0.792
	C-K	0.112	0.000
Minerva on llima	Acc.	0.500	0.458
	C-K	0.230	0.230
Minerva Post-OCR	Acc.	0.479	0.312
	C-K	0.244	0.056

Table 3: Evaluation metrics: Accuracy (Acc.) and Cohen’s Kappa (C-K) for each translation model and evaluator

7.5 A - Confusion matrices

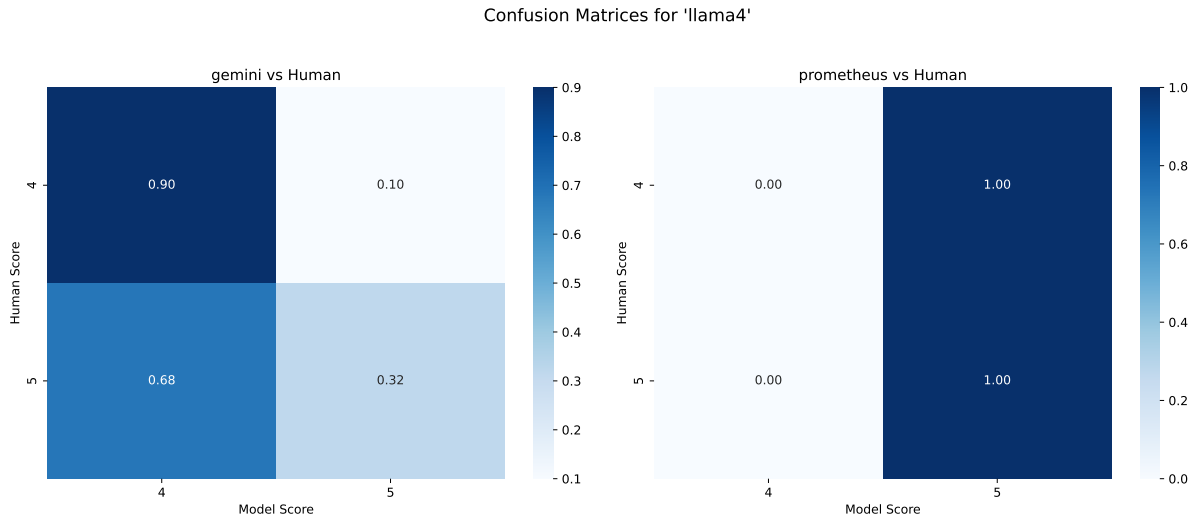


Figure 1: Confusion matrices of llama4-model corrections. On the left, Gemini Vs Human. On the right, Prometheus Vs Human.

Confusion matrices for t5 and Minerva models show an overall high discrepancy between human scores and both Gemini and Prometheus scores. For the t5 model, Gemini is unable to distinguish a sentence with minor errors from a sentence with major issues; while Prometheus has assigned low scores even with perfect, or almost, outputs (e.g. "THE VAMPYRE; -> THE VAMPYRE;" scored 2 instead of 5). Both models behave better on Minerva corrections, but still with imprecisions. The worst correlation is given by Prometheus on the corrections of the Minerva model double-finetuned on LIMA and Post-OCR datasets.

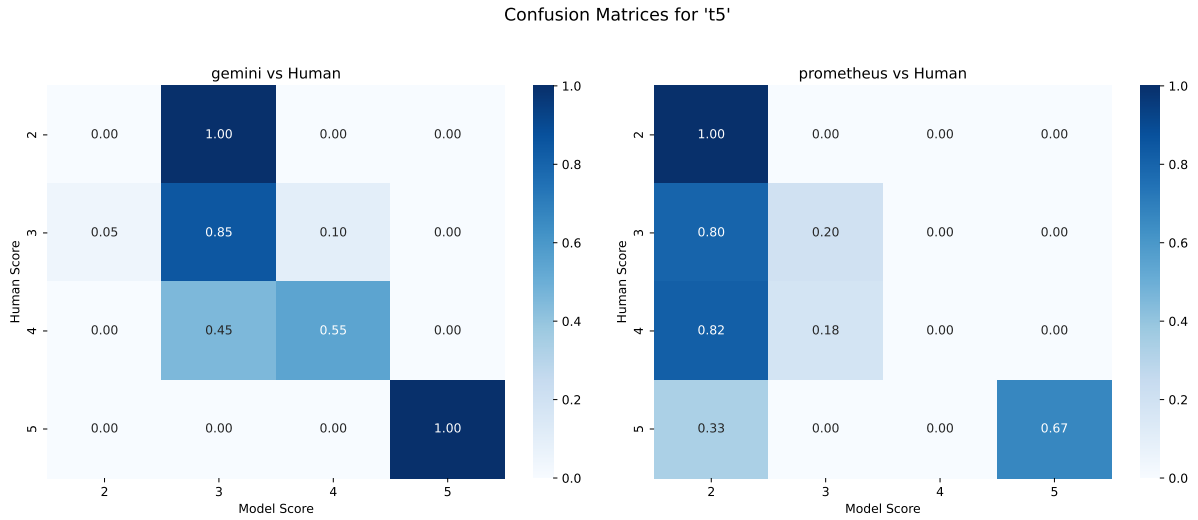


Figure 2: Confusion matrices of t5-model corrections. On the left: Gemini vs. Human. On the right: Prometheus vs. Human.

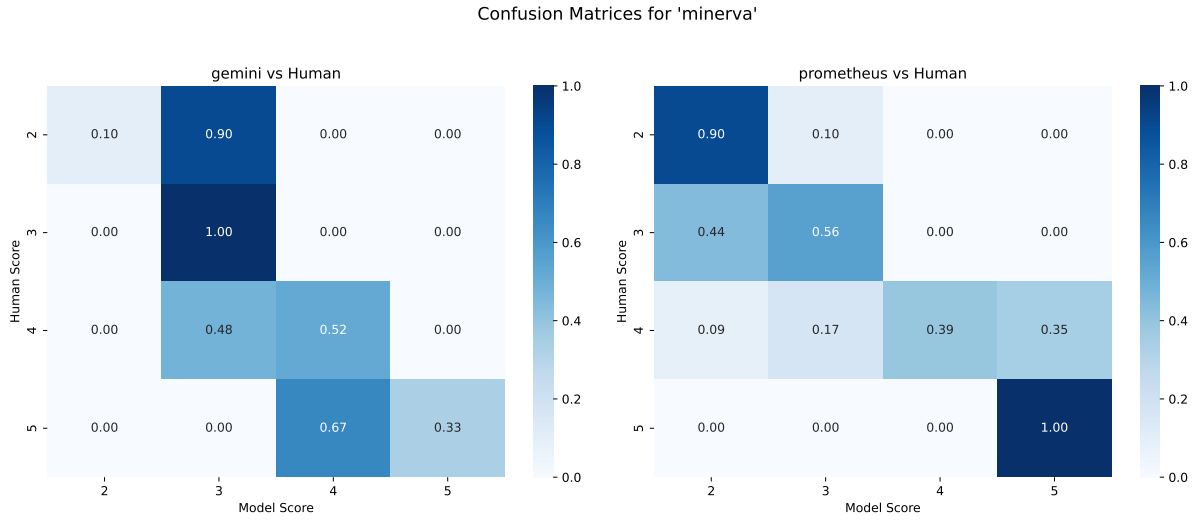


Figure 3: Confusion matrices of minerva-model corrections. On the left: Gemini vs. Human. On the right: Prometheus vs. Human.

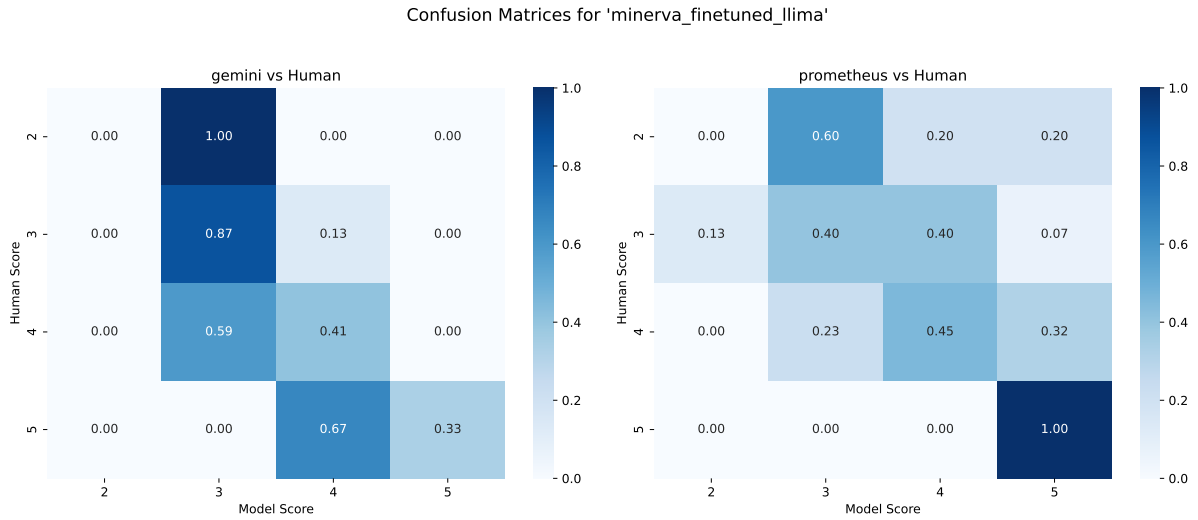


Figure 4: Confusion matrices for the corrections produced by the once-fine-tuned on LIMA Minerva model. On the left: Gemini vs. Human. On the right: Prometheus vs. Human.

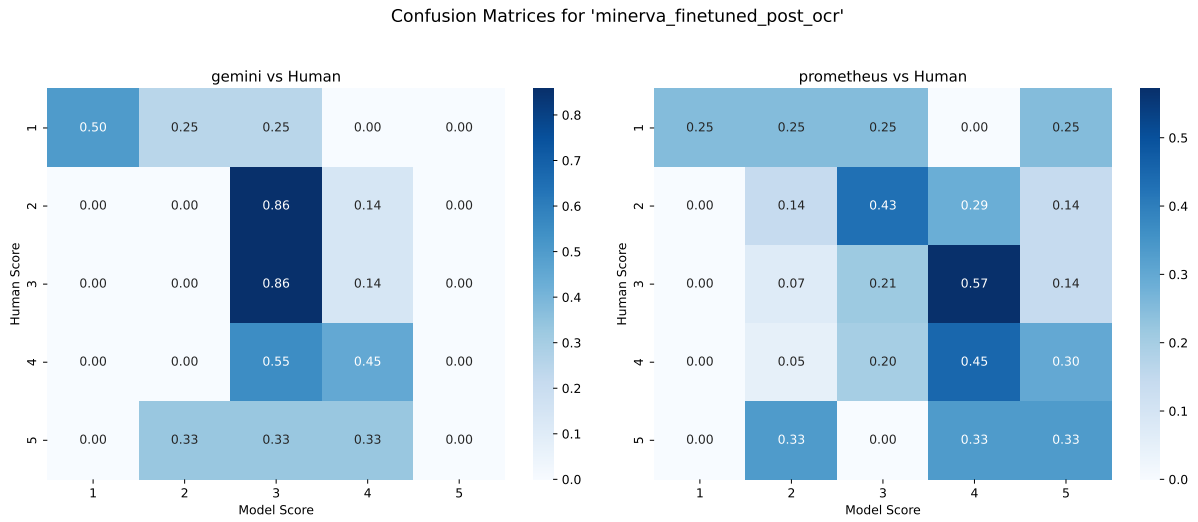
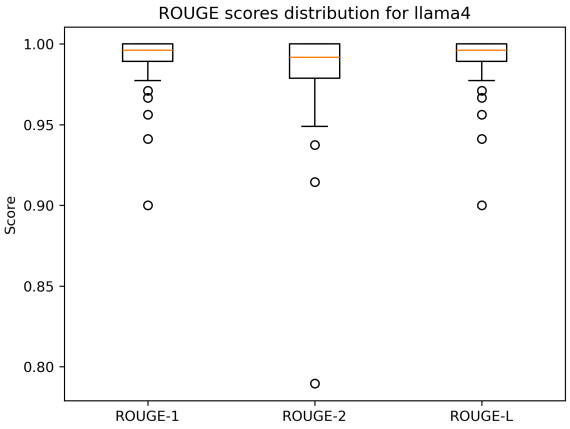


Figure 5: Confusion matrices for the corrections produced by the double-fine-tuned Minerva model. On the left: Gemini vs. Human. On the right: Prometheus vs. Human.

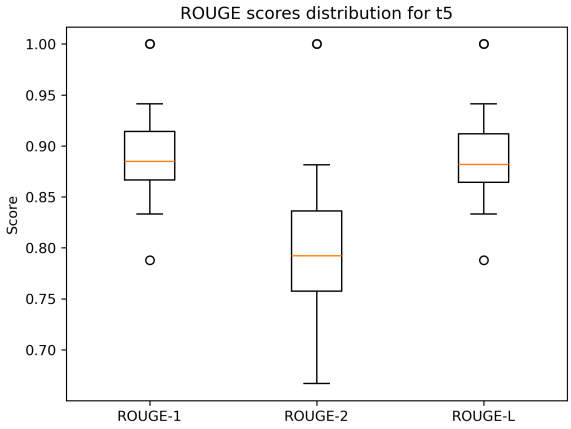
7.6 Correction quality

Model	ROUGE-1			ROUGE-2			ROUGE-L		
	Mean	ρ	p-val	Mean	ρ	p-val	Mean	ρ	p-val
T5 + Spell Checker	0.897	0.234	0.110	0.811	0.289	0.047	0.895	0.231	0.115
Minerva	0.858	0.701	2.83e-08	0.773	0.651	5.58e-07	0.851	0.705	2.16e-08
LLaMA4	0.990	0.042	0.779	0.982	0.049	0.740	0.990	0.042	0.779
Minerva on LIMA	0.872	0.645	7.60e-07	0.793	0.686	7.41e-08	0.864	0.669	2.03e-07
Minerva Post-OCR	0.784	0.443	0.0016	0.675	0.405	0.0043	0.771	0.438	0.0019

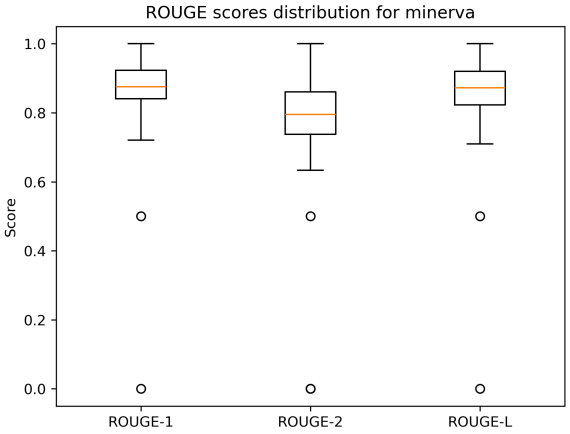
Table 4: Mean ROUGE scores, Spearman ρ and corresponding p-values with human annotations.



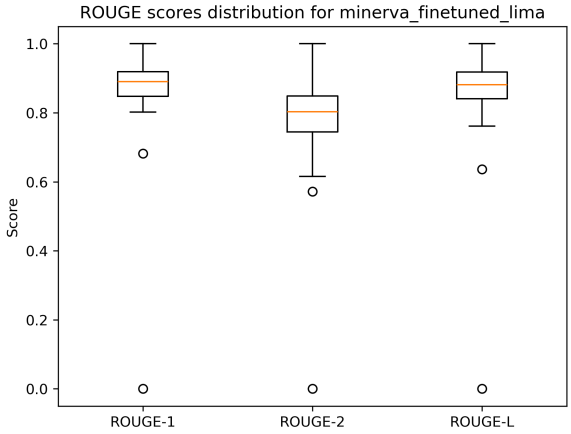
(a) Rouge scores for llama4



(b) Rouge scores for t5 + SpellChecker



(c) Rouge scores for minerva



(d) Rouge scores for minerva finetuned on LIMA

Figure 6: Rouge Scores: orange line is the median, the height of the box is the variance, small circler are outlier

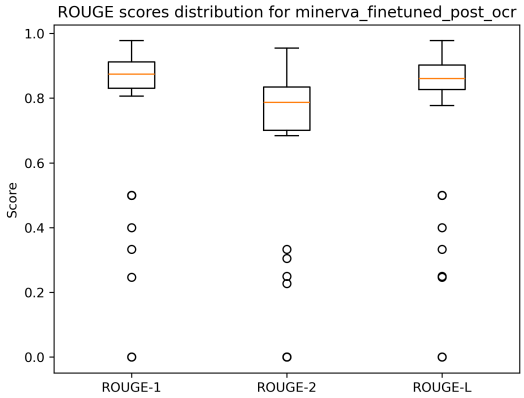


Figure 7: Rouge Scores for minerva finetuned on post OCR: orange line is the median, the height of the box is the variance, small circler are outlier