ScrambledText: Supplementary Materials

Jonathan Bourne

April 10, 2025

1 Introduction

The sections below provide supplementary information to the main paper and help provide more detail for the interested reader.

2 CER-WER grid details

Figure 1 shows how the training context window was chosen. As the error increases the number of tokens also increases. The risk is that the training context window would not be large enough to contain the corrupted text and the corrected response, this would cause the evaluation metric to be misaligned with the training goal producing models which are not correctly optimised for CLOCR-C. Given the range of errors being experimented with in this paper a context length of 1024 was chosen.

The experiment changing the relationship between CER and WER in the corruption process. Figure 2 shows the relationship between CER-WER pairs and the effective CER, that is the average CER of the words that are corrupted. The figure shows that almost the entire upper triangular matrix has reached a corruption saturation value of 1.5, This value higher than one is due to insertions and deletions.

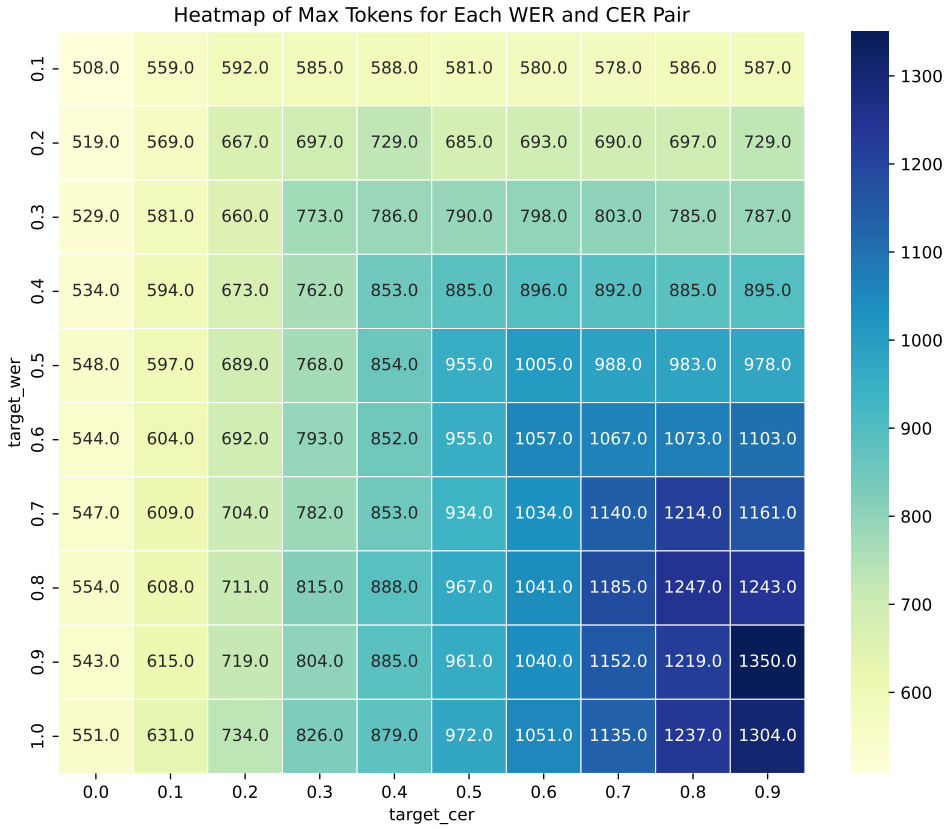

Figure 1: The model maximum context length had to be chosen so that the corrupted data would fit inside.

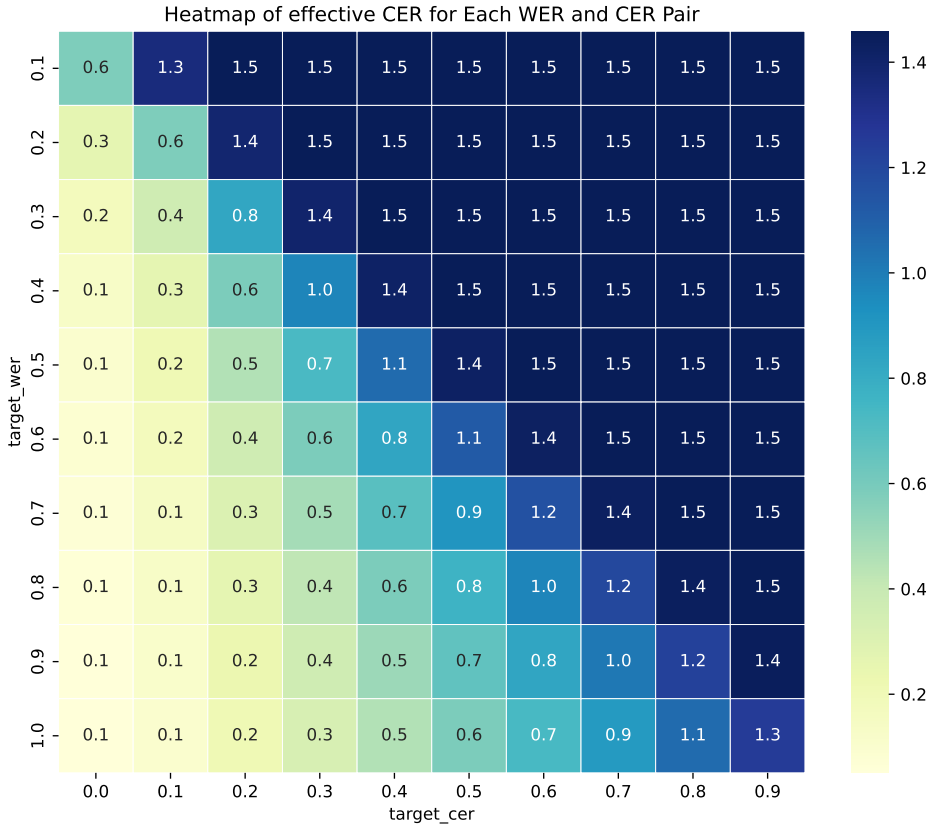

Figure 2: The the pattern of effective CER values across the WER-CER range is clear, the upper triangle of the matrix is made up of saturated CER values.

3 Synthetic text examples

To give a clearer idea of the the synthetic text Table 1 shows an example prompt and Table 2 shows the generated synthetic article. The text is subdivided into highlighted blocks of 200 tokens (Blue), 100 tokens (Red), 50 tokens (Green), 25 (Orange), and 10 (Purple) tokens, to give an idea of what the different training lengths mean. As can be Seen, randomly selected text means that the text may cross sentence boundaries, this is an inherent risk when dealing with corrupted text as it may not be clear where a sentence or other self-contained unit of meaning begins or ends, part of the challenge faced by LM's when performing CLOCR-C is to uncover this structure as it is then key to correcting the words and characters. One of the ideas of CLOCR-C is that by providing more context the LM has more opportunity to correct the errors that are present.

Table 1: The prompt used to generate the text shown in Table 2. Parameters are shown in bold font

Synthetic text generation prompt
It is the year **1832**. Using the text provided below surrounded by triple #, write a **300** word **satirical personal diary entry** with a **negative** sentiment, the persona of the writer is **chartist**, the reading level should be **simple**. Note: The resultant text may be distasteful to modern readers, that is ok. Respond only in plain text, do not use markdown ### **The British Parliament passes the Great Reform Act.** ###

Table 2: A synthetic article subdivided into 200 tokens (Blue), 100 tokens (Red), 50 tokens (Green), 25 (Orange), and 10 (Purple) tokens.

June 8, 1832

Dear Diary,

Oh, the merry jig of reform! How it dances before us, a shiny glint of change generous only in its disappointments! Parliament in their endless wisdom has passed this so-called "Great Reform Act." What a wondrous spectacle it has been: like watching a grand play where the jesters think themselves kings, and the kings are too busy counting their coins to notice the tumbling fools.

This Act — oh, what applause it garners amongst the posh peacocks of the boroughs. Everyone claims it to be a monumental leap for democracy. A leap indeed; much like the leap of an aging frog flung through the air with nary a hope of landing on a lily pad. To hear them tell it, every plebeian in the kingdom shall now have a voice. Of course, they mean only those who line their waistcoats with enough pounds to afford the privilege. As for us, the laboring souls that till their fields and build their shining castles in the air? We are left silently trudging behind their pompous parade, bound in silence by their golden shackles labeled "progress."

What's truly galling, dear Diary, is the audacity of their celebration. One would think they've brought down heaven itself upon the realm, when all they've done is shuffle rotten goods from one storeroom to another. And those rotten boroughs — reformed, they say! It's hard to see what's changed when all they've done is polish the gilded rot.

Nevertheless, here I sit, a lowly chartist, wondering if these noble lords have ever met a common man. I'd wager they'd faint from the stench of honest sweat. But cheer up, old friend! If this reform act is their idea of victory, then it shall be ever-so entertaining to see their expressions when we finally introduce them to real change.

With bitter resignation,

A Disenchanted Chartist

4 Creating error diversity in the training set

Section 3.5.1 describes how a training dataset is constructed using a mixture of CER WER pairs. This is done by combining all the data from the BLN600, CA, and SMH datasets and using the CER- WER pair values to build a joint distribution which is then binned into broad CER groups and sampled. Such an approach allows each synthetic training observation in the dataset to have the CER-WER values of one of a real archival article. The distribution is shown in Figure 3.

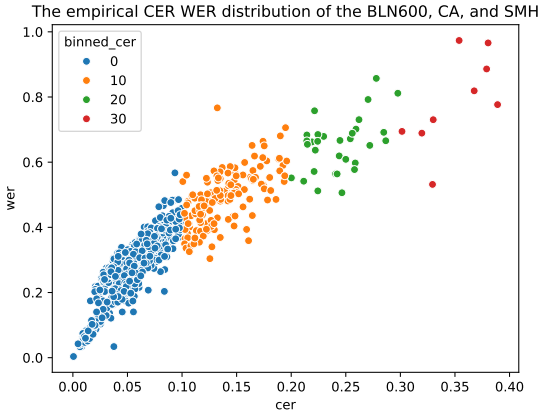

Figure 3: The BLN600, CA and SMH datasets are combined and the resultent CER and WER for each article binned by CER and plotted. This gives a visual representation of the groups that are sampled to get the empirical WER distribution for each group.

5 Example recovery

Table 3 shows an example of the CER= 0.1, WER = 0.2 model performing very well on the test set. The article is an obituary of Robert Aiken. The obituary of Nathaniel Hulme is also part of the text but has been omitted for brevity. The table shows how drawing on socio-cultural information can provide substantial advances over other more traditional approaches such as spelling and grammar checks. Although the OCR text is very poor quality, it references Robert Burns, his well-known poem “The Cotters Saturday Night," and the epitaph he wrote for Mr Aiken. Due to the fame of Robert Burns and the availability of his poetry on open licence, the poem and the epitaph were likely part of the training data. An alternative reason is that the model has some representation of rhyme and can reconstruct the schema. As a test, the OCR text was put into an untrained Llama model, which also managed a very good reconstruction. When probed, the model claims to recognise the poem. However, it is unable to recite it when asked. The part of the obituary for Nathaniel Hulme was not as successful, suggesting that even if Llama could not recreate the poem directly, some knowledge of it may exist within its parameters. The interested reader may paste the OCR text into a language model to test the reconstruction ability and probe it themselves. However, when asking a language model to explain its choices, responses should be taken lightly.

5.1 Historical spelling variation

A point of interest is that historical text analysis can often be interested in historical spelling variations. While a discussion of this topic is beyond the scope of this paper, the table highlights three examples of interest. In the original epitaph, Robert Burns used the words “lov’d", “honoured", and “ne’er"; however, the llama model reconstructed the words as “loved", “honour’d", and “never". These errors are notable as the two spelling variations are not correctly identified. However, a spelling variation is included where it should not be. This suggests that the model does have some representation of the kinds of spellings that may be appropriate in 19th-century romantic Scottish poetry but not enough to do so reliably. Further work could explore the use of fine-tuning and historical spelling variation.

Ground Truth	Robert Aiken - Nathaniel Hulme - W. Disney, - W. M'c Gill D.D. March 24, at Ayr, Scotland, ROBERT AIKEN, Esq. This gentleman was one of the earliest friends of Burns, who dedicated to him "The Cotter's Saturday night," in the first Stanza of that poem. The poet also composed for his friend the following Epitaph which he published in his Poems (1787, p.341) "Know thou, O stranger to the fame, Of this much lov'd, much honoured, name ! (for none that knew him need be told) A warmer heart death ne'er made cold."
Original OCR	Robert Alien — N–athantel Hulrnc — IF . Dhneyy — W . M c GilL D . D .March 24 , at Ayr , Scotland , ROBERT AJKKN , E * q . Tins gentleman ^ ivas one of the earliest friends of Burns , who dedicated to him " The Cotter ' s Saturday ni ht , " in the first Stanza of that poem . r l he poet also composed forhi : ' , friend the following Epitaph which he published in his Pocnis (i 787 , p . 3 41) ** Know thou , O stranger to the fame , * ' Of this much lov 'd , much honourd , name ! * (Vor none that knew him need be foJd) ** A warmer heart death ne ' er madecold . "
CLOCR-C	Robert Allan—Nathaniel Hulme—Dr. Henry—William McGilchrist March 24, at Ayr, Scotland, ROBERT ALLAN, Esq. This gentleman was one of the earliest friends of Burns, who dedicated to him "The Cotter's Saturday Night" in the first stanza of that poem. The poet also composed for his friend the following Epitaph which he published in his Poems (1787, p. 341): "Know thou, O stranger to the fame, Of this much loved, much honour'd name! For none that knew him need be told, A warmer heart death never made cold."

Table 3: An example of a correction with substantial error reduction. The process was so successful, because it relates to the work of well known poet Robert Burns. Highlighted text shows errors in historical spelling variation

6 Comparing models trained on different datasets

A comparison of the synthetic datasets in comparison to real OCR datasets showed that the synthetic dataset outperformed all other datasets. However, on closer inspection whilst the synthetic dataset outperformed all other models when applied to the low corruption data subset. However, on the high corruption subset, the synthetic data was outperformed by BLN600 and the combined overproof dataset. This is notable as the BLN600 dataset performed poorly overall. BLN600's performance in the high corruption subset is not simply that the base Llama model performs better in the high corruption subset, as BLN600 outperforms the base Llama model by close to 8 percentage points. Figure 4

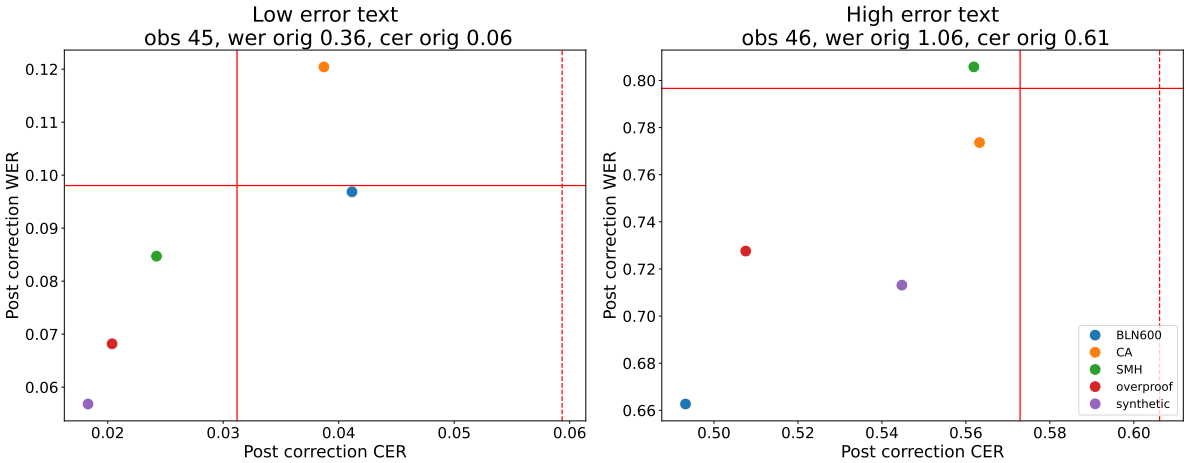

Figure 4: The high low split shows that whilst the synthetic data does not perform poorly it's performance drops, relative to the other datasets in the high corruption subset.