**Sentence Completion**

The first risk assessment variable analyzed within the WikiText dataset was sentence completion, a variable that characterizes the extent to which sentences are fully formed and grammatically correct. Sentence completion is crucial in evaluating the overall coherence and quality of language within the dataset. The analysis involved examining the distribution of sentence completion scores, assessing the frequency of incomplete or poorly structured sentences, and identifying any patterns or outliers. A BERT sequence classification model and tokenizer were used to determine completeness scores for the training examples, and the distribution was visualized in the violin plot shown in Figure 1. As shown, there was a broad range of completeness scores, with a minimum of around 0.375 and a maximum of around 0.550. The scores on the lesser end of the spectrum could indicate training examples that lack syntactical proficiency and are, therefore, of lower quality.

**Figure 1:** ***A violin plot of completeness scores of the first 100 training examples in the C4 dataset as evaluated by the BERT model and tokenizer for sequence classification.***

**Readability**

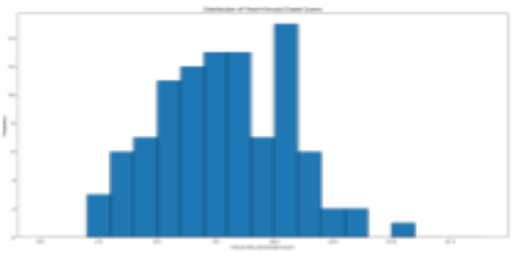

**Figure 2:** ***A histogram of Flesch-Kincaid Grade Scores across the training examples present within the subset***

The second risk assessment variable analyzed within the subset of training examples was readability, a metric of how difficult language is to understand. The Flesch-Kincaid Grade Score was used via the textstat Python library implementation to quantify readability. Flesch-Kincaid Grade Score returns a numerical value representing the grade level of education needed for a reader to understand a given text excerpt. Extremes in readability score could potentially indicate something wrong with a training example, such

as incorrect labels, ambiguous content, or outliers. Therefore, the distribution and variability were plotted in the histogram shown in Figure 2.

**Lexical Diversity**

The third risk-assessment variable analyzed within the data was lexical diversity, a metric representing the diversity of the vocabulary and semantics within the text. Lexical diversity was scored for the training examples across the subset using the Moving Average Type-Token-Ratio (McCarthy and Jarvis, 2010). It was then plotted and analyzed within the kernel density plot in Figure 3. Another potential variable for filtering out poor training examples.

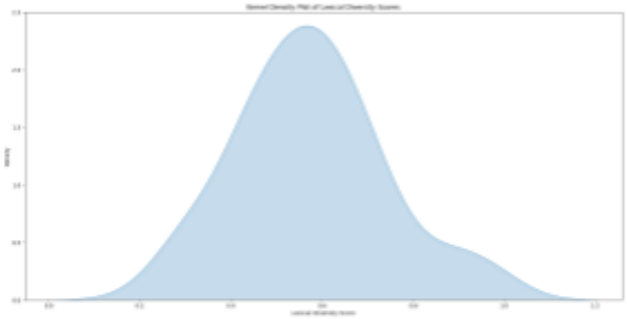

**Figure 3:** ***A kernel density plot of the lexical diversity scores as determined using MATTR for the training examples in the subset***