Replacement Language Modeling

Bill Ray

November 15, 2022

1 Setup

1.1 Empty Token

A token sequence can be decoded into a string by concatenating the characters of the tokens. The empty token is a special token that represents the empty string. Inserting the empty token into a token sequence at any point does not change the string that is decoded from the sequence. Token sequences that decode to the same string are functionally equivalent. Consider the empty token to be part of the vocabulary going forward.