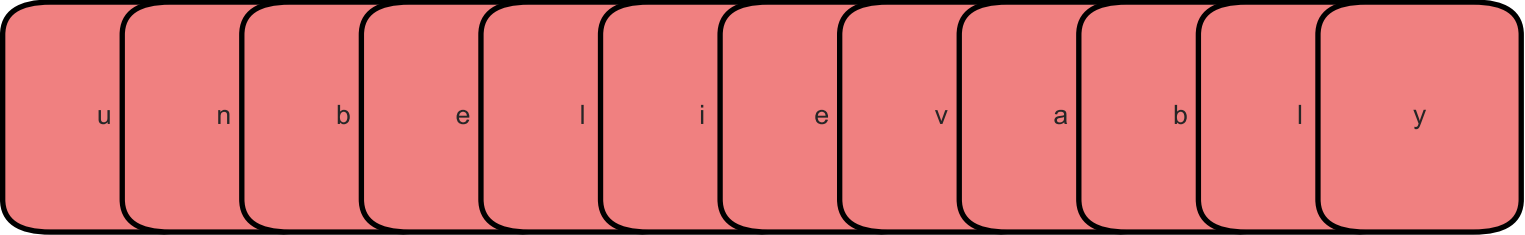


The Tokenization Spectrum

Example: "unbelievably"

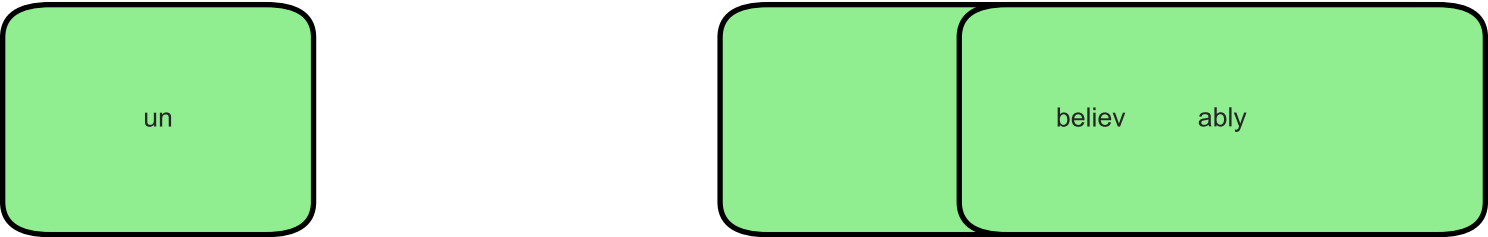
Character



12 tokens

Too long sequences
Model learns spelling

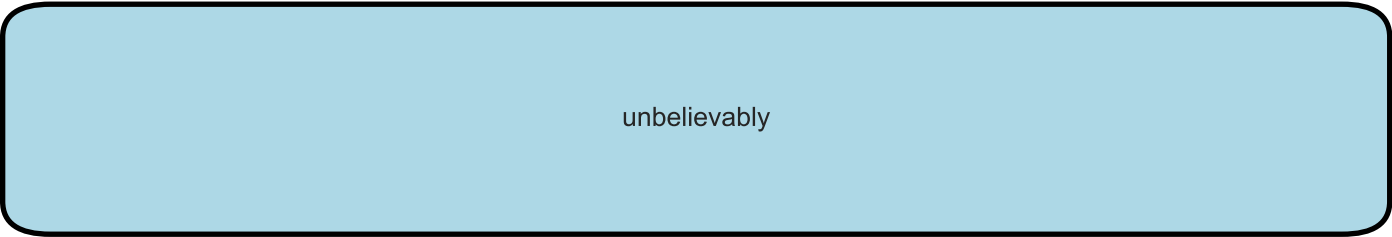
Subword



3 tokens

Balanced
Optimal for most tasks

Word



1 tokens

Vocabulary explosion
Can't handle new words

Granularity