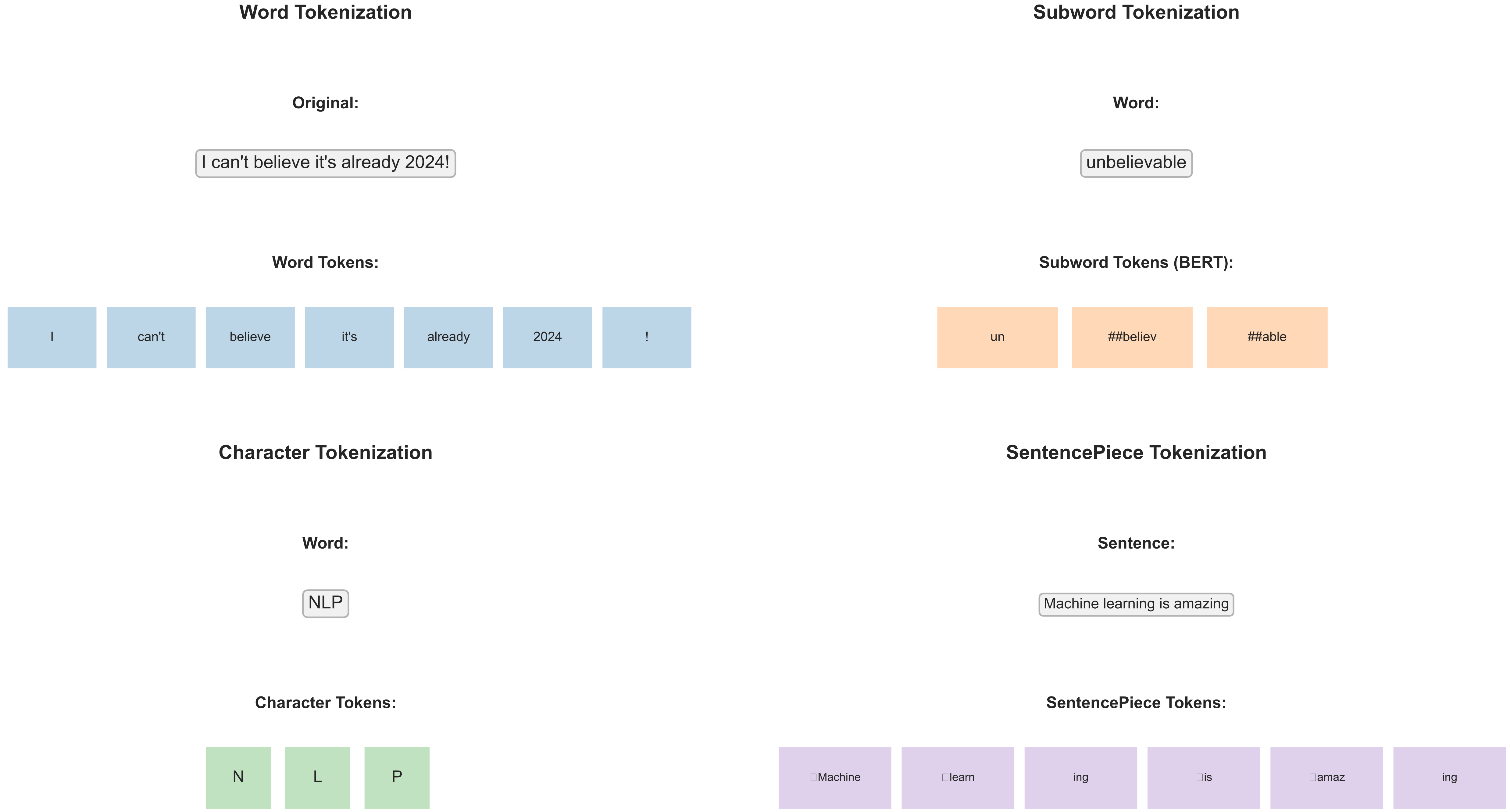
Tokenization Methods Comparison



Word Tokenization

Original:

I can't believe it's already 2024!

Word Tokens:

I | can't | believe | it's | already | 2024 | !

Subword Tokenization

Word:

unbelievable

Subword Tokens (BERT):

un | ##believ | ##able

Character Tokenization

Word:

NLP

Character Tokens:

N | L | P

SentencePiece Tokenization

Sentence:

Machine learning is amazing

SentencePiece Tokens:

□Machine | □learn | ing | □is | □amaz | ing