

██████████	SentencePiece BPE (██████████████████.)	mT5 (██████████████████.)	MiniLM (██████████████████.)	SentencePiece BPE (██████████████.)	mT5 (██████████████.)	MiniLM (██████████████.)
Vocabulary size (real)	4230.0	5921.0	5387.0	3859.0	4549.0	4462.0
Total tokens	376610.0	281587.0	318945.0	447549.0	519084.0	444623.0
Total characters	875040.0	875040.0	875040.0	1170005.0	1170005.0	1170005.0
Compression ratio (chars/tokens)	2.3235	3.1075	2.7435	2.6143	2.254	2.6315
Bits per token (theoretical)	12.0464	12.5316	12.3953	11.914	12.1513	12.1235
Bits per token (real)	7.9555	8.4705	8.2136	7.2074	6.2199	6.9613
Total info (theoretical, Mbit)	4.54	3.53	3.95	5.33	6.31	5.39
Total info (real, Mbit)	3.0	2.39	2.62	3.23	3.23	3.1