Performance Comparison Across NLP Tasks

Word Analogy Task (Google Analogy Dataset) — Accuracy (%)

Model	Accuracy (%)
N-grams	45
Word2Vec	75
ELMo	82
BERT	89
GPT-3	93
GPT-4	96
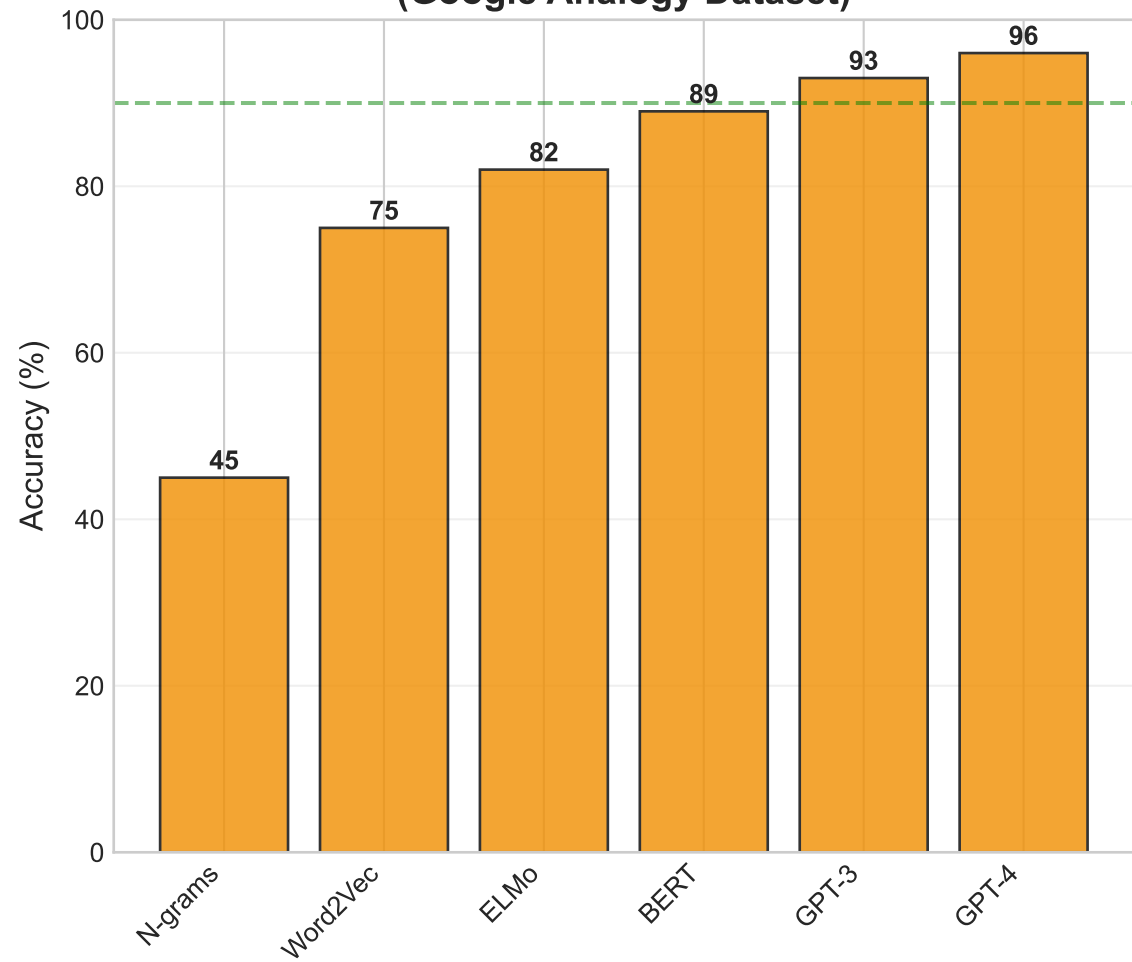
Sentiment Analysis (Stanford Sentiment Treebank) — F1 Score (%)

Model	F1 Score (%)
N-grams	72
Word2Vec	81
ELMo	88
BERT	94
GPT-3	96
GPT-4	97
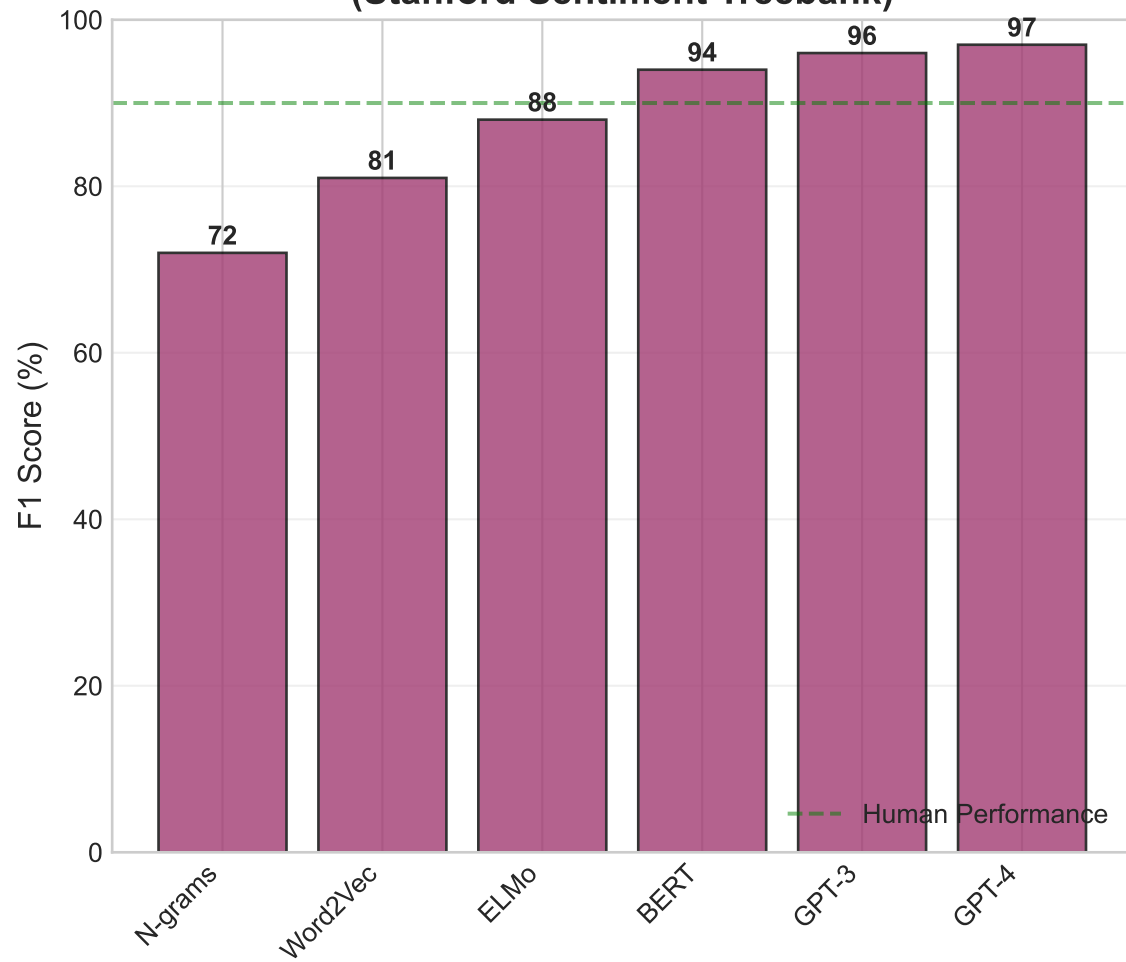
Named Entity Recognition (CoNLL-2003) — F1 Score (%)

Model	F1 Score (%)
N-grams	68
Word2Vec	76
ELMo	86
BERT	92
GPT-3	94
GPT-4	95

Human Performance (dashed line)