Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.2.mlp	German	56064	The highlighted tokens are primarily German morphemes, word stems, or inflections, often marking noun or verb endings, compound word parts, or grammatical function words. Many are common in German-language text, including noun endings (-en, -heit, -keit, -ung), verb forms (-iert, -te, -en), and connecting words (und, von, in, der, die, das, eine). There is also frequent activation on recurring roots in compound nouns and on inflectional suffixes, reflecting the morphological structure of German.	0.67	0.548	0.87	0.4	0.72	0.65	0.867	0.52
model.layers.5.mlp	German	31366	The highlighted tokens are primarily German (and some Russian) morphemes, word stems, and inflectional endings, often marking grammatical features such as case, number, tense, or forming compound nouns and verbs. Many are parts of longer words or compounds, reflecting the agglutinative and inflectional nature of German, and are often found at the beginning, middle, or end of words.	0.8	0.783	0.857	0.72	0.8	0.8	0.8	0.8
model.layers.8.mlp	German	79729	The highlighted tokens are primarily German function words, verb forms, noun and adjective endings, and common prefixes or suffixes, often marking grammatical relationships, tense, plurality, or case, as well as frequent collocations and set phrases. These tokens are essential for sentence structure and meaning in German text.	0.69	0.608	0.828	0.48	0.66	0.585	0.75	0.48
model.layers.9.mlp	German	9083	The highlighted tokens are primarily function words, verb forms, noun and adjective endings, and common collocations in German sentences, often marking grammatical structure, tense, or case, as well as frequent phrase boundaries and connectors. These tokens are crucial for sentence construction and meaning in German text.	0.85	0.831	0.949	0.74	0.83	0.817	0.884	0.76
model.layers.10.mlp	German	44371	The highlighted spans are predominantly German multi-word expressions, noun or verb phrases, and compound words, often marking grammatical constructs, inflections, or idiomatic units. Many are functionally important for sentence structure, such as verb-final clauses, participial forms, or noun compounds, and frequently include suffixes or prefixes that signal tense, plurality, or case. These patterns reflect the morphological richness and syntactic dependencies characteristic of German text.	0.87	0.851	1	0.74	0.86	0.837	1	0.72
model.layers.11.mlp	German	27401	The highlighted tokens are predominantly German morphemes, word stems, and grammatical endings, often marking inflections, compounds, or function words. These tokens frequently appear at the boundaries of words or phrases, indicating their importance for understanding German syntax, morphology, and sentence structure.	0.86	0.837	1	0.72	0.82	0.791	0.944	0.68
model.layers.11.mlp	German	52681	The highlighted tokens are primarily German function words, verb forms, pronouns, and common noun or adjective suffixes, especially those marking person, tense, or plurality. There is a strong focus on grammatical structures, such as verb conjugations, modal verbs, and case endings, as well as on frequently used collocations and phrasal patterns in German sentences.	0.97	0.969	1	0.94	0.94	0.94	0.94	0.94
model.layers.12.mlp	German	107855	The highlighted tokens are predominantly German morphemes, suffixes, and function words that form or modify nouns, verbs, and adjectives, as well as common conjunctions and prepositions. These tokens often appear at the ends of words (e.g., -ung, -en, -heit, -lich, -keit, -en, -t, -te, -en, -er, -es, -e, -en, -nen, -heit, -keit, -ung, -schaft, -lich, -bar, -ig, -end, -ieren,	0.88	0.867	0.975	0.78	0.83	0.809	0.923	0.72