Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.14.mlp	German	312	The highlighted tokens are predominantly German suffixes and word endings that form nouns, adjectives, and participles, often indicating grammatical features such as case, number, gender, or degree, as well as forming compound words and inflections.	0.92	0.915	0.977	0.86	0.83	0.828	0.837	0.82
model.layers.14.mlp	German	48455	The highlighted tokens are primarily German morphemes, function words, and noun or verb endings, often marking grammatical features such as case, number, tense, or forming compound words. There is a focus on suffixes, inflections, and connecting words that are essential for sentence structure and meaning in German text.	0.97	0.969	1	0.94	0.95	0.949	0.959	0.94
model.layers.15.mlp	German	17109	The highlighted tokens are primarily German morphemes, function words, and common suffixes or inflections, as well as high-frequency connectors and pronouns. These include verb and noun endings, conjunctions, prepositions, and pronouns, which are essential for sentence structure and meaning in German text. The activations also emphasize recurring grammatical patterns, such as subordinate clause markers, verb conjugations, and noun/adjective endings, reflecting the importance of syntactic and morphological elements in German language processing.	0.88	0.864	1	0.76	0.86	0.844	0.95	0.76
model.layers.15.mlp	German	75403	The highlighted tokens are predominantly German morphemes, function words, and common suffixes or inflections, often marking verb forms, noun endings, or grammatical particles. These tokens frequently appear at the boundaries of words or as part of compound constructions, reflecting the morphological richness and syntactic structure of German text.	0.91	0.901	1	0.82	0.86	0.854	0.891	0.82
model.layers.15.mlp	German	130833	The highlighted tokens are predominantly German function words, inflectional endings, and common noun or adjective suffixes, reflecting grammatical structure such as articles, pronouns, prepositions, conjunctions, and case or number markers essential for sentence construction and meaning.	0.87	0.854	0.974	0.76	0.83	0.821	0.867	0.78