Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.15.mlp	Hindi	101032	The important tokens are primarily Hindi morphemes, suffixes, and inflections that mark grammatical relationships, case, number, tense, and honorifics, as well as common function words and noun/adjective endings. These tokens often appear at the end of words or as standalone grammatical markers, reflecting the agglutinative and inflectional nature of Hindi syntax.	0.85	0.824	1	0.7	0.85	0.828	0.973	0.72
model.layers.15.mlp	Hindi	112041	The highlighted tokens are primarily common Hindi syllables, morphemes, or short word fragments, often appearing at the start or within proper nouns, place names, and compound words, reflecting the agglutinative and inflectional nature of Hindi text. These tokens frequently serve as building blocks for larger words, especially in names, titles, and technical terms.	0.92	0.915	0.977	0.86	0.92	0.918	0.938	0.9
model.layers.15.mlp	Hindi	114764	The highlighted tokens are primarily Hindi grammatical suffixes, verb endings, and particles that indicate tense, number, gender, case, or emphasis, as well as sentence-ending punctuation. These elements are essential for sentence structure and meaning in Hindi text.	0.8	0.75	1	0.6	0.8	0.756	0.969	0.62
model.layers.15.mlp	Hindi	119973	The most prominent pattern is the frequent occurrence of the Hindi postposition \"क\" (ke/ka/ki/ko), which functions as a grammatical marker for possession, relation, or object, and is highly activated in various inflected forms and contexts throughout the text.	0.8	0.75	1	0.6	0.76	0.684	1	0.52