Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.4.mlp	Hindi	37982	The highlighted tokens are primarily common Hindi morphemes, suffixes, and function words such as case markers, postpositions, pronouns, and verb endings. These elements are essential for grammatical structure and meaning in Hindi sentences, often appearing at word boundaries or as inflections, and are crucial for parsing and understanding the syntactic and semantic relationships within the text.	0.96	0.959	0.979	0.94	0.94	0.939	0.958	0.92
model.layers.4.mlp	Hindi	76042	The highlighted tokens are primarily single Hindi characters or short morphemes that function as grammatical markers, case endings, or parts of compound words, often appearing at word boundaries or within inflected forms, indicating their importance in the structure and meaning of Hindi sentences.	0.94	0.936	1	0.88	0.93	0.929	0.939	0.92
model.layers.4.mlp	Hindi	98694	The tokens correspond to Hindi postpositions and case markers, especially \"का\", \"की\", and \"के\", which indicate possession or relation, and are frequently attached to nouns to form genitive constructions.	0.69	0.551	1	0.38	0.68	0.529	1	0.36
model.layers.5.mlp	Hindi	128732	The highlighted tokens are primarily function words, verb forms, and common morphemes in Hindi, such as case markers, auxiliary verbs, pronouns, and postpositions, which are essential for sentence structure and meaning.	0.98	0.98	0.962	1	0.95	0.952	0.909	1
model.layers.7.mlp	Hindi	129521	The highlighted tokens are primarily Hindi morphemes, syllables, or short word fragments that frequently appear as grammatical markers, inflections, or connectors within and between words. These include common verb endings, case markers, postpositions, and other functional elements that are essential for sentence structure and meaning in Hindi text.	0.69	0.551	1	0.38	0.7	0.583	0.955	0.42
model.layers.8.mlp	Hindi	13166	The highlighted tokens are suffixes, inflections, or postpositions in Turkish, Hindi, and related languages, often marking grammatical relationships such as possession, case, plurality, or tense, and are frequently attached to nouns, verbs, or proper names.	0.81	0.765	1	0.62	0.74	0.74	0.74	0.74
model.layers.8.mlp	Hindi	17924	The highlighted segments are primarily Hindi phrases, words, or morphemes that serve as key grammatical, semantic, or syntactic units within sentences. These include verb forms, noun phrases, postpositions, and connectors that are essential for sentence structure, meaning, or emphasis. The selections often mark important actions, attributions, or relationships, and frequently appear at clause or sentence boundaries, or as part of idiomatic or functional expressions.	0.95	0.947	1	0.9	0.92	0.915	0.977	0.86
model.layers.8.mlp	Hindi	48133	The highlighted tokens are primarily morphemes, suffixes, and root components in Hindi and related languages, often marking grammatical roles, inflections, or forming compound words, especially in technical, scientific, or formal contexts. These elements are crucial for word formation and meaning in complex or compound terms.	0.8	0.773	0.895	0.68	0.82	0.83	0.786	0.88
model.layers.8.mlp	Hindi	88795	The highlighted tokens are primarily Hindi morphemes, suffixes, and root words that contribute to grammatical structure, verb formation, and meaning, often marking actions, states, or relationships within sentences. These tokens are essential for constructing and understanding the core semantics and syntax of Hindi text.	0.68	0.529	1	0.36	0.68	0.529	1	0.36
model.layers.9.mlp	Hindi	12012	The highlighted tokens are verb forms and auxiliary constructions in Hindi, especially those ending with \"ते हैं\", \"ता है\", \"ती है\", \"ता हूँ\", \"ते हुए\", and similar, which are used to indicate habitual actions, ongoing processes, or states of being. These forms are central to expressing tense, aspect, and agreement in Hindi sentences.	0.64	0.438	1	0.28	0.66	0.485	1	0.32
model.layers.9.mlp	Hindi	26002	The highlighted tokens are primarily Hindi morphemes, suffixes, and function words that contribute to grammatical structure, verb conjugation, and meaning, such as markers for tense, aspect, possession, negation, and case. These elements are essential for sentence construction and semantic clarity in Hindi text.	0.83	0.795	1	0.66	0.83	0.805	0.946	0.7