Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.13.mlp	Hindi	74419	The highlighted tokens are primarily verb roots and suffixes in Hindi, often marking verb forms, actions, or participles, especially at the end of words or as part of verb conjugations. These tokens are central to expressing actions, states, or processes in sentences.	0.81	0.776	0.943	0.66	0.86	0.848	0.929	0.78
model.layers.13.mlp	Hindi	100472	The highlighted tokens are common Hindi suffixes, postpositions, and inflectional endings that modify nouns, verbs, and adjectives to indicate grammatical relationships such as case, number, gender, tense, and respect, as well as particles and connectors essential for sentence structure.	0.73	0.64	0.96	0.48	0.73	0.69	0.811	0.6
model.layers.13.mlp	Hindi	109574	The highlighted tokens are Hindi verb endings and auxiliary markers, especially those denoting tense, aspect, and agreement (such as \"है\", \"था\", \"गया\", \"किया\", \"हैं\"), as well as punctuation and conjunctions that structure sentences. These elements are crucial for indicating grammatical relationships and sentence boundaries in Hindi text.	0.66	0.485	1	0.32	0.67	0.507	1	0.34
model.layers.13.mlp	Hindi	111068	The highlighted tokens are primarily Hindi morphemes, suffixes, and grammatical markers that contribute to word formation, inflection, and syntactic structure, such as case endings, verb forms, and connectors. These elements are essential for conveying relationships between words and for the grammatical coherence of Hindi sentences.	0.85	0.824	1	0.7	0.86	0.837	1	0.72
model.layers.13.mlp	Hindi	113614	The highlighted tokens are common Hindi suffixes and word endings that form nouns, adjectives, or verbs, often indicating grammatical inflection, tense, plurality, or case, and are crucial for the structure and meaning of Hindi words.	0.68	0.543	0.95	0.38	0.73	0.649	0.926	0.5
model.layers.13.mlp	Hindi	119344	The highlighted tokens are primarily Hindi and related Indic script morphemes, suffixes, and inflections that form or modify nouns, verbs, and adjectives, often marking case, number, gender, tense, or forming compound words. These tokens are essential for grammatical structure and meaning in Hindi sentences, frequently appearing at word endings or as connectors within compound expressions.	0.83	0.795	1	0.66	0.83	0.805	0.946	0.7
model.layers.13.mlp	Hindi	122646	The highlighted tokens predominantly mark verb forms, especially those indicating actions or states (such as \"करते\", \"कहते\", \"देखते\", etc.), as well as grammatical particles and suffixes that contribute to tense, aspect, or case. There is also emphasis on noun forms, pronouns, and connectors that structure sentences, reflecting a focus on the core syntactic and semantic elements that drive meaning and grammatical relationships in Hindi text.	0.96	0.959	0.979	0.94	0.88	0.872	0.932	0.82
model.layers.14.mlp	Hindi	22707	The highlighted tokens are primarily Hindi suffixes, inflections, and word endings that modify grammatical meaning, such as case, number, gender, tense, or form compound words. These include common noun, verb, and adjective endings, as well as postpositions and particles that are essential for syntactic and semantic structure in Hindi sentences.	0.96	0.958	1	0.92	0.96	0.958	1	0.92
model.layers.14.mlp	Hindi	27401	Single uppercase letters or digraphs, often at the start of a word or name, are highlighted across multiple languages and scripts, frequently marking proper nouns, initials, or abbreviations.	0.51	0.462	0.512	0.42	0.92	0.923	0.889	0.96
model.layers.14.mlp	Hindi	27671	The highlighted tokens are primarily morphemes, syllables, or word segments that are common in Hindi and English loanwords, especially those forming or ending nouns, adjectives, or technical terms. These segments often appear in compound words, transliterations, or borrowed terminology, and are frequently found in formal, technical, or institutional contexts.	0.7	0.643	0.794	0.54	0.74	0.755	0.714	0.8
model.layers.14.mlp	Hindi	38728	High activations are found on single Devanagari characters, especially those that commonly begin or are part of Hindi words, with a strong emphasis on the character for \"म\" (ma), often in positions marking pronouns, verbs, or key sentence elements.	0.76	0.7	0.933	0.56	0.81	0.776	0.943	0.66
model.layers.14.mlp	Hindi	63110	The highlighted tokens are common Hindi vowel and consonant diacritics, matras, and inflections, as well as endings of words, which are essential for correct word formation, pronunciation, and grammatical structure in Hindi text.	0.71	0.592	1	0.42	0.71	0.633	0.862	0.5