Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.15.mlp	Hindi	49232	The highlighted tokens are common morphemes, suffixes, and inflectional endings in Hindi, such as vowel and consonant diacritics, plural and case markers, and common noun or verb endings, which are essential for word formation and grammatical structure in the language.	0.99	0.99	1	0.98	0.98	0.98	0.98	0.98
model.layers.15.mlp	Hindi	52158	The highlighted tokens are primarily punctuation marks, quotation marks, and sentence-ending markers in Hindi text, often appearing at the end of reported speech, direct quotations, or as delimiters for parenthetical or quoted content. These tokens help structure sentences, indicate dialogue, and separate clauses or phrases.	0.73	0.658	0.897	0.52	0.84	0.814	0.972	0.7
model.layers.15.mlp	Hindi	53517	High activations are found on common Hindi single-letter tokens, especially those marking grammatical particles, case markers, or inflections such as \"म\", \"ह\", \"ब\", \"व\", \"ल\", \"प\", \"ज\", and \"द\", which frequently appear at morpheme or word boundaries and play key roles in sentence structure.	0.81	0.771	0.97	0.64	0.8	0.778	0.875	0.7
model.layers.15.mlp	Hindi	56951	The highlighted tokens are primarily function words, common morphemes, and noun or verb roots in Hindi and related scripts, often marking grammatical relationships, inflections, or forming the core of named entities and key content words. These tokens are essential for sentence structure, meaning, and the identification of important entities or actions within the text.	0.79	0.747	0.939	0.62	0.71	0.734	0.678	0.8
model.layers.15.mlp	Hindi	76535	The highlighted tokens are primarily Hindi morphemes, suffixes, and inflections that contribute to grammatical structure, word formation, and meaning, such as case markers, verb endings, pluralization, and connectors. These elements are essential for syntactic and semantic cohesion in Hindi sentences.	1	1	1	1	1	1	1	1
model.layers.15.mlp	Hindi	77521	The highlighted tokens are common Hindi suffixes, verb endings, case markers, and particles that play a key role in inflection, agreement, and grammatical structure within sentences. These include markers for tense, number, gender, case, and postpositions, as well as common conjunctions and punctuation, reflecting the morphological richness and syntactic dependencies of Hindi text.	0.92	0.913	1	0.84	0.92	0.913	1	0.84
model.layers.15.mlp	Hindi	79680	The highlighted tokens are primarily suffixes, inflections, and common morphemes in Hindi, such as vowel and consonant endings, postpositions, and grammatical markers. There is also frequent activation on punctuation, conjunctions, and some corrupted or non-standard characters, indicating a focus on morphological structure, word boundaries, and syntactic connectors within Hindi text.	0.96	0.958	1	0.92	0.96	0.959	0.979	0.94
model.layers.15.mlp	Hindi	85279	The highlighted tokens are primarily Hindi and transliterated foreign words, with a focus on noun and proper noun morphemes, suffixes, and inflectional endings. There is a strong emphasis on word parts that denote names, places, titles, and grammatical markers, often at the start or end of words, reflecting the structure and morphology of Hindi and related languages.	0.91	0.903	0.977	0.84	0.9	0.891	0.976	0.82
model.layers.15.mlp	Hindi	94347	The highlighted tokens are primarily Hindi morphemes, suffixes, and function words that serve grammatical roles such as case marking, tense, plurality, and conjunctions. These include postpositions, verb endings, and common connectors, which are essential for sentence structure and meaning in Hindi text.	0.9	0.889	1	0.8	0.89	0.879	0.976	0.8
model.layers.15.mlp	Hindi	99002	The highlighted tokens are primarily Hindi or Devanagari script morphemes, syllables, or short word fragments, often marking the start, end, or important part of proper nouns, place names, or key content words. There is a strong emphasis on tokens that function as grammatical markers, connectors, or are part of compound words, especially in names, locations, and official titles. The pattern reflects the segmentation of Hindi text into meaningful units, with frequent focus on morphemes that contribute to the structure and meaning of complex words or phrases.	0.91	0.901	1	0.82	0.89	0.884	0.933	0.84