Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.14.mlp	Hindi	70559	The highlighted tokens are common Hindi suffixes and verb endings that indicate tense, aspect, plurality, or grammatical agreement, such as markers for ability, completion, plurality, and participles. These morphemes are essential for conveying grammatical relationships and meaning in Hindi sentences.	0.98	0.98	1	0.96	0.99	0.99	1	0.98
model.layers.14.mlp	Hindi	77614	The highlighted tokens are primarily morphemes, word stems, or suffixes in Hindi, often marking grammatical, semantic, or inflectional features such as case, number, gender, or verb tense. Many are parts of compound words, technical terms, or proper nouns, and some are high-activation due to their role in forming or modifying the core meaning of words within complex or formal contexts.	0.75	0.684	0.931	0.54	0.8	0.787	0.841	0.74
model.layers.14.mlp	Hindi	79133	The pattern highlights Hindi plural markers and oblique case endings, especially the nasalized \"ों\" suffix, as well as other common inflectional morphemes and case markers that modify nouns and pronouns to indicate plurality, possession, or grammatical relationships.	0.72	0.611	1	0.44	0.71	0.592	1	0.42
model.layers.14.mlp	Hindi	102271	The highlighted tokens are primarily Hindi verb forms, suffixes, and auxiliary markers that indicate tense, aspect, mood, or agreement, as well as common noun and adjective endings. These tokens are essential for constructing grammatical and meaningful sentences in Hindi, often appearing at the end of words to convey actions, states, or relationships.	0.68	0.529	1	0.36	0.71	0.613	0.92	0.46
model.layers.14.mlp	Hindi	108281	The highlighted tokens are primarily Hindi script morphemes, suffixes, and inflections, including vowel and consonant diacritics, case markers, and common word endings. These elements are crucial for grammatical structure, word formation, and meaning in Hindi, often marking tense, number, gender, case, or forming compound words. The activations focus on these subword units and inflectional endings, reflecting their importance in the morphological and syntactic construction of Hindi sentences.	0.83	0.795	1	0.66	0.82	0.791	0.944	0.68
model.layers.14.mlp	Hindi	108954	The highlighted tokens are common Hindi suffixes and vowel diacritics, such as \"ा\", \"ी\", \"े\", \"ै\", and \"ें\", which are essential for inflection, tense, plurality, and grammatical agreement in Hindi words. These suffixes frequently appear at the end of verbs, nouns, and adjectives, marking grammatical features and contributing to the overall meaning and structure of sentences.	0.78	0.718	1	0.56	0.77	0.709	0.966	0.56
model.layers.14.mlp	Hindi	125706	The highlighted tokens are primarily Hindi verb roots, suffixes, and inflections that form or modify verbs, often indicating tense, aspect, or agreement. These include common verb endings, auxiliary markers, and participial forms, as well as noun or adjective roots that combine with these endings to create meaningful phrases in context. The activations focus on morphemes essential for constructing or modifying actions and states in Hindi sentences.	0.89	0.876	1	0.78	0.96	0.958	1	0.92
model.layers.15.mlp	Hindi	3696	The highlighted tokens are primarily Hindi grammatical particles, case markers, postpositions, conjunctions, and inflectional suffixes that indicate relationships between words, such as possession, comparison, agency, and coordination. These tokens are essential for sentence structure, meaning, and cohesion in Hindi text.	0.83	0.795	1	0.66	0.8	0.767	0.917	0.66
model.layers.15.mlp	Hindi	16877	The highlighted tokens are primarily Hindi morphemes, suffixes, and inflections, as well as common connective words and noun/adjective endings. There is frequent emphasis on grammatical markers, case endings, and conjunct forms, along with some corrupted or non-standard characters. The pattern reflects a focus on the morphological structure and syntactic connectors in Hindi text, especially those that signal relationships between words or modify meaning.	0.8	0.75	1	0.6	0.83	0.805	0.946	0.7
model.layers.15.mlp	Hindi	28301	The highlighted tokens are primarily suffixes, inflections, and short morphemes in Hindi, as well as common conjunctions, postpositions, and endings that contribute to grammatical structure and meaning. There is also frequent emphasis on diacritics, conjuncts, and short function words, reflecting the importance of morphological and syntactic markers in Hindi text.	0.9	0.889	1	0.8	0.85	0.839	0.907	0.78