Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.12.mlp	Hindi	80456	The vowel \"े\" in Hindi frequently appears as a grammatical marker, often as part of postpositions or verb forms, and is highly activated when used in conjunction with other postpositional or case-marking tokens, especially in phrases indicating purpose, possession, or relation.	0.73	0.64	0.96	0.48	0.67	0.507	1	0.34
model.layers.12.mlp	Hindi	92233	The highlighted tokens are primarily Hindi verb suffixes, auxiliary verbs, and noun/adjective endings that indicate tense, aspect, number, gender, or degree. These include common verb endings (such as those for present, past, and future tense), comparative and superlative markers, and participial forms, as well as auxiliary verbs that support the main verb. The pattern reflects the morphological structure of Hindi, where such suffixes and auxiliaries are crucial for grammatical meaning.	0.68	0.529	1	0.36	0.73	0.63	1	0.46
model.layers.12.mlp	Hindi	96938	The highlighted tokens are primarily suffixes, inflections, or morphemes in Hindi and related languages, often marking grammatical features such as tense, plurality, or case, as well as common noun and verb roots. There is a strong emphasis on the token \"व\", which frequently appears as a morpheme or part of compound words, and on other short, high-frequency morphemes that contribute to word formation and meaning in Hindi text.	0.81	0.765	1	0.62	0.85	0.839	0.907	0.78
model.layers.12.mlp	Hindi	108252	The token \"ज\" is highly activated when it appears as part of Hindi verbs related to going, knowing, or actions (e.g., जाना, जानना, जाता, जाते), often in various conjugated forms and contexts.	0.68	0.529	1	0.36	0.62	0.387	1	0.24
model.layers.12.mlp	Hindi	108339	The highlighted tokens are common Hindi verb suffixes and endings (such as \"ता\", \"ते\", \"ता है\", \"ते हैं\", \"करते\", \"होता\", \"देते\") that indicate tense, aspect, and agreement in verbs, marking present, habitual, or continuous actions.	0.65	0.462	1	0.3	0.69	0.551	1	0.38
model.layers.12.mlp	Hindi	116711	The highlighted tokens are primarily Hindi (and some transliterated or foreign) morphemes and words, often marking diminutives, adjectives, or nouns describing size, quantity, or relation (such as \"छोटा\" for \"small\", \"बहुत\" for \"very\", \"अपनी\" for \"own\", \"लंबा\" for \"long\", etc.), as well as suffixes and inflections that modify meaning or grammatical role. The activations focus on meaningful morphemes that contribute to descriptive or relational semantics in the sentence.	0.92	0.917	0.957	0.88	0.92	0.917	0.957	0.88
model.layers.13.mlp	Hindi	17614	The highlighted tokens are primarily suffixes, inflections, or morphemes in Hindi and related scripts, often marking grammatical features such as case, number, tense, or forming part of compound words. These tokens frequently appear at word boundaries or as part of agglutinative constructions, reflecting the morphological richness of the language.	0.82	0.786	0.971	0.66	0.75	0.737	0.778	0.7
model.layers.13.mlp	Hindi	23180	The highlighted tokens are common Hindi morphemes, suffixes, and inflections that form or modify verbs, adjectives, and nouns, often indicating tense, plurality, possession, or degree, and are essential for grammatical structure and meaning in Hindi sentences.	0.75	0.667	1	0.5	0.76	0.692	0.964	0.54
model.layers.13.mlp	Hindi	39402	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, and pluralization. These elements are essential for syntactic and semantic coherence in Hindi sentences.	0.89	0.876	1	0.78	0.87	0.854	0.974	0.76
model.layers.13.mlp	Hindi	57919	The tokens correspond to common Hindi postpositions and case markers, such as those indicating location, means, or possession, which frequently appear attached to nouns or pronouns and are essential for grammatical relationships in Hindi sentences.	0.71	0.603	0.957	0.44	0.72	0.632	0.923	0.48
model.layers.13.mlp	Hindi	69531	The highlighted tokens are common Hindi verb and noun suffixes, especially those forming infinitives, participles, comparatives, or case endings, such as \"ने\", \"के\", \"ने के\", \"ाने\", \"ाने के\", \"करने\", \"आने\", \"बनने\", \"रहने\", \"होने\", and plural or oblique markers like \"ों\", \"ी\", \"ा\", \"े\". These suffixes are crucial for indicating grammatical relationships, actions, and states in Hindi sentences.	0.87	0.854	0.974	0.76	0.87	0.851	1	0.74