Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.10.mlp	Hindi	115455	The highlighted tokens are verb forms and auxiliary constructions in Hindi, often marking present, past, or continuous tense, and commonly ending with \"ते\", \"ते हैं\", \"ते थे\", \"ता है\", \"ता हूँ\", \"ते हुए\", \"करते\", \"रहते\", \"कहते\", etc. These forms indicate habitual, ongoing, or reported actions and are central to Hindi verb conjugation and sentence structure.	0.65	0.462	1	0.3	0.64	0.438	1	0.28
model.layers.10.mlp	Hindi	119016	The highlighted tokens are primarily Hindi morphemes, suffixes, and inflections that contribute to grammatical structure, tense, comparison, negation, and case marking, as well as common noun and verb roots. These elements are essential for forming meaning, relationships, and emphasis in Hindi sentences.	0.75	0.667	1	0.5	0.79	0.741	0.968	0.6
model.layers.10.mlp	Hindi	126020	The token \"क\" is highly activated when used as a grammatical marker in Hindi, especially as a postposition or inflectional suffix attached to nouns and pronouns to indicate possession, relation, or case, and frequently appears at the end of or within words in various syntactic contexts.	0.77	0.701	1	0.54	0.82	0.78	1	0.64
model.layers.11.mlp	Hindi	16927	High activations are found on common Hindi morphemes, especially verb endings, pronouns, and postpositions, as well as frequent function words and inflectional suffixes, reflecting their grammatical importance in sentence structure.	0.71	0.592	1	0.42	0.73	0.682	0.829	0.58
model.layers.11.mlp	Hindi	26219	The highlighted tokens are primarily Hindi morphemes, suffixes, and inflections that form or modify nouns, verbs, and adjectives, often marking case, number, gender, tense, or aspect, and are essential for grammatical structure and meaning in Hindi sentences.	0.73	0.63	1	0.46	0.75	0.667	1	0.5
model.layers.11.mlp	Hindi	52559	The highlighted tokens are primarily Hindi verb roots, suffixes, and inflections that form or modify verbs, indicating tense, aspect, or action, often appearing at the end or within verbs to convey ongoing, completed, or potential actions.	0.83	0.8	0.971	0.68	0.88	0.867	0.975	0.78
model.layers.11.mlp	Hindi	56085	The most salient pattern is the frequent activation of common Hindi grammatical morphemes and postpositions, especially those involving the \"क\" (ka/ke/ki) forms, which function as possessive markers, connectors, or case markers, as well as other inflectional endings like \"ों\", \"ी\", \"ा\", and \"े\". These tokens are essential for syntactic structure and meaning in Hindi sentences, often linking nouns, indicating possession, plurality, or case, and forming compound expressions.	0.75	0.667	1	0.5	0.74	0.649	1	0.48
model.layers.11.mlp	Hindi	63085	The text contains frequent use of Hindi conjuncts and half-letters, especially the \"्\" (virama) and combinations with \"व\", \"र\", \"य\", and other consonants, which are characteristic of Hindi morphology and word formation, particularly in formal or technical contexts.	0.69	0.597	0.852	0.46	0.68	0.556	0.909	0.4
model.layers.11.mlp	Hindi	89219	The highlighted tokens are common Hindi suffixes and pronouns, such as those marking plurality (ों), comparative or infinitive verb forms (ने, करने, ाने), and pronouns (उन, इन, उस), as well as noun and verb endings. These morphemes are essential for grammatical structure, indicating actions, possession, plurality, and subject-object relationships in Hindi sentences.	0.77	0.701	1	0.54	0.78	0.718	1	0.56
model.layers.11.mlp	Hindi	100951	The highlighted tokens are primarily Hindi morphemes or word stems that form the core of verbs, nouns, or adjectives, often marking tense, aspect, or grammatical function. These tokens frequently appear at the end or within words, contributing to the inflection or derivation of meaning, and are essential for constructing or modifying actions, states, or qualities in the sentence.	0.74	0.658	0.962	0.5	0.73	0.667	0.871	0.54
model.layers.11.mlp	Hindi	114722	The highlighted tokens are primarily Hindi grammatical suffixes, verb endings, and particles that indicate tense, number, gender, and case, as well as common function words and inflections essential for sentence structure and meaning.	0.86	0.837	1	0.72	0.85	0.839	0.907	0.78