Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.11.mlp	Hindi	129484	The highlighted tokens frequently correspond to named entities, locations, organizations, and key connecting words in multiple languages, often marking the start or continuation of proper nouns, place names, or important phrases within travel routes, political contexts, or descriptive sequences. These tokens are often found at the boundaries of multi-word expressions or in the context of enumerating places, people, or organizations, and they include both content words and function words that are crucial for linking or specifying entities in multilingual text.	0.48	0.509	0.482	0.54	0.41	0.504	0.435	0.6
model.layers.12.mlp	Hindi	5795	The highlighted tokens are common Hindi morphemes and pronouns such as \"हम\", \"िस\", \"ुझ\", and plural or case markers like \"ों\", which frequently appear as grammatical components in Hindi sentences, often forming parts of pronouns, possessives, or verb conjugations.	0.85	0.824	1	0.7	0.86	0.837	1	0.72
model.layers.12.mlp	Hindi	9045	The highlighted tokens are primarily Hindi grammatical markers, case endings, and common suffixes (such as those denoting possession, plurality, comparison, or verb tense), as well as frequently used pronouns, conjunctions, and numerals. These elements are essential for sentence structure and meaning, often appearing at word endings or as short function words.	0.66	0.485	1	0.32	0.75	0.713	0.838	0.62
model.layers.12.mlp	Hindi	10675	The important tokens are primarily Hindi suffixes, inflections, and endings that indicate grammatical case, number, tense, or form, as well as noun and verb roots, which are essential for meaning and sentence structure in Hindi text.	0.74	0.649	1	0.48	0.74	0.675	0.9	0.54
model.layers.12.mlp	Hindi	13954	Suffixes and endings in Hindi, such as े, ं, ों, ी, ै, ों, and ें, are frequently activated, reflecting their grammatical role in verb conjugation, noun/adjective inflection, and pluralization. These endings are crucial for indicating tense, number, gender, and case in Hindi sentences.	0.7	0.583	0.955	0.42	0.69	0.597	0.852	0.46
model.layers.12.mlp	Hindi	25328	The highlighted tokens are primarily Hindi morphemes, pronouns, postpositions, and verb endings that serve as grammatical connectors, markers of tense, person, or case, and are essential for sentence structure and meaning in Hindi text.	0.93	0.925	1	0.86	0.93	0.925	1	0.86
model.layers.12.mlp	Hindi	46246	The text frequently highlights pronouns and possessive forms, especially those related to the first person (such as \"मुझे\", \"मैं\", \"मेरे\"), as well as case markers and connectors that are essential for sentence structure and meaning in Hindi.	0.67	0.507	1	0.34	0.66	0.5	0.944	0.34
model.layers.12.mlp	Hindi	46338	The token \"क\" frequently appears as a grammatical marker in Hindi, often as part of postpositions or possessive constructions, and is highly activated when attached to or preceding other morphemes to indicate relationships such as possession, association, or purpose.	0.71	0.603	0.957	0.44	0.76	0.684	1	0.52
model.layers.12.mlp	Hindi	49871	The highlighted tokens are primarily verb roots and suffixes in Hindi, often forming verbs related to actions, states, or processes such as making, doing, taking, finding, or moving. These roots are frequently combined with auxiliary or inflectional suffixes to indicate tense, aspect, or modality.	0.7	0.571	1	0.4	0.81	0.771	0.97	0.64
model.layers.12.mlp	Hindi	57856	The highlighted tokens are primarily Hindi verb endings and auxiliary verbs that indicate tense, aspect, mood, or agreement, such as forms of \"है\", \"हैं\", \"था\", \"हुआ\", \"किया\", and various participial or infinitive endings. These are essential for constructing grammatical sentences and conveying actions, states, or conditions in Hindi.	0.67	0.507	1	0.34	0.72	0.611	1	0.44
model.layers.12.mlp	Hindi	72462	The highlighted tokens are primarily suffixes, roots, or morphemes that form the core of Hindi nouns, adjectives, and verbs, often marking grammatical categories such as case, number, gender, or forming abstract and compound words. These elements are central to word formation and meaning in Hindi, frequently appearing in formal, literary, or technical vocabulary.	0.73	0.64	0.96	0.48	0.75	0.719	0.821	0.64