Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.1.mlp	Hindi	56943	The English examples highlight the use of the word \"distance\" and its plural \"distances\" in contexts related to measurement, travel, or physical space. The Hindi examples show frequent activation of single-character tokens, especially \"स\" and \"क\", which are common morphemes or syllables in Hindi, often appearing as grammatical markers, prefixes, or within compound words, indicating a focus on morphological or syntactic elements in the language.	0.8	0.762	0.941	0.64	0.88	0.867	0.975	0.78
model.layers.1.mlp	Hindi	65640	The Hindi character \"ह\" is highly activated when used as an auxiliary or copula in verb forms, indicating tense, aspect, or state, and frequently appears at the end of clauses or sentences.	0.57	0.246	1	0.14	0.61	0.361	1	0.22
model.layers.1.mlp	Hindi	106597	The highlighted tokens often correspond to morphemes, syllables, or short word fragments that are significant in named entities, place names, or proper nouns, especially in multilingual or transliterated contexts. These fragments frequently appear in the middle or end of words and are commonly found in Indian names, administrative terms, and other culturally specific vocabulary.	0.73	0.682	0.829	0.58	0.55	0.646	0.532	0.82
model.layers.2.mlp	Hindi	49394	The tokens correspond to common Hindi grammatical particles and suffixes, such as case markers and postpositions (e.g., \"के\", \"की\", \"का\", \"को\", \"कि\", \"कीya\"), which are essential for indicating relationships between nouns, possession, and grammatical roles in sentences.	0.68	0.543	0.95	0.38	0.71	0.603	0.957	0.44
model.layers.2.mlp	Hindi	53504	The Hindi character \"है\" (or its variants with different matras) frequently appears at the end of sentences or clauses, functioning as a copula or auxiliary verb to indicate present tense or state of being.	0.69	0.563	0.952	0.4	0.69	0.551	1	0.38
model.layers.2.mlp	Hindi	70081	The tokens \"े\" and \"ा\" are highly activated as common Hindi grammatical suffixes, frequently marking case, possession, plurality, or verb forms, and often appear at the end of words or as postpositions.	0.79	0.747	0.939	0.62	0.78	0.732	0.938	0.6
model.layers.3.mlp	Hindi	12962	The most prominent pattern is the frequent activation of Hindi postpositions and grammatical markers, especially the token corresponding to \"क\" (ka/ke/ki/ko/ka), which functions as a possessive, case marker, or connector in Hindi grammar. These tokens are highly activated in contexts where they attach to or modify nouns, pronouns, or verbs, reflecting their central role in sentence structure and meaning in Hindi text.	0.67	0.535	0.905	0.38	0.65	0.478	0.941	0.32
model.layers.3.mlp	Hindi	82959	The highlighted tokens are common function words, suffixes, or short morphemes in Turkish and Hindi, such as \"de\", \"da\", and various single-character Hindi syllables, which serve grammatical or connective roles within sentences.	0.68	0.644	0.725	0.58	0.73	0.697	0.795	0.62
model.layers.3.mlp	Hindi	85531	The highlighted tokens are morphemes, syllables, or word fragments from various languages, often appearing in proper nouns, technical terms, or culturally significant words, especially those related to Indic languages, Sanskrit, and related terminology. These fragments frequently occur at word boundaries or within compound words, reflecting their importance in identifying or constructing key terms across multilingual contexts.	0.85	0.835	0.927	0.76	0.7	0.762	0.632	0.96
model.layers.3.mlp	Hindi	106101	The text frequently highlights Hindi tokens related to indefinite pronouns and possessives, such as forms of \"किसी\", \"की\", \"का\", and \"के\", as well as other common grammatical morphemes. These tokens often appear in contexts expressing generality, possession, or relation, and are central to sentence structure and meaning in Hindi.	0.89	0.876	1	0.78	0.92	0.913	1	0.84
model.layers.3.mlp	Hindi	128429	The tokens \"का\", \"के\", and \"की\" are postpositions in Hindi that indicate possession or relation, frequently following nouns or pronouns to form genitive constructions. The high activations on these tokens reflect their grammatical importance in linking entities and expressing relationships in sentences.	0.66	0.5	0.944	0.34	0.66	0.485	1	0.32