Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.0.mlp	Hindi	17276	The most prominent pattern is the frequent activation of the Hindi character \"ह\" in verb forms, especially as part of auxiliary verbs indicating tense, aspect, or existence (such as है, हैं, होता, होती, होना, हो, etc.), which are essential for sentence structure in Hindi. Other activated characters like \"र\", \"ल\", \"द\", \"भ\", and \"श\" appear less frequently and are typically part of nouns or adjectives, but \"ह\" consistently marks grammatical constructions related to being, happening, or possession.	0.77	0.71	0.97	0.56	0.78	0.72	1.00	0.56
model.layers.0.mlp	Hindi	26755	The vowel diacritics \"े\" and \"ा\" in Hindi, frequently attached to nouns and verbs, indicating grammatical relationships such as possession, case, or postpositions.	0.87	0.85	1.00	0.74	0.86	0.84	1.00	0.72
model.layers.0.mlp	Hindi	33080	The highlighted tokens are common Hindi consonants or syllables, often appearing as prefixes, suffixes, or inflections within words, and are frequently used in grammatical constructions or to form compound words. Their high activation suggests a focus on morphemes that play a key role in word formation and meaning in Hindi text.	0.83	0.80	1.00	0.66	0.81	0.78	0.94	0.66
model.layers.0.mlp	Hindi	53040	The highlighted tokens are frequent, short function words or morphemes in French (\"va\") and Hindi (\"क\", \"म\", \"ह\", etc.), often serving as grammatical connectors, verb forms, or case markers, and are highly activated due to their structural importance in sentence construction.	0.96	0.96	0.98	0.94	0.94	0.94	0.92	0.96
model.layers.0.mlp	Hindi	55180	The most important tokens are Hindi postpositions, case markers, and grammatical particles such as \"क\", \"के\", \"की\", \"का\", and related conjuncts, which are essential for indicating possession, relation, or grammatical structure in Hindi sentences. These tokens frequently appear at the boundaries of noun phrases or as connectors, reflecting their central role in sentence construction and meaning.	0.65	0.46	1.00	0.30	0.65	0.48	0.94	0.32
model.layers.0.mlp	Hindi	56263	The tokens correspond to common Hindi postpositions and case markers, such as \"के\", \"की\", \"का\", \"को\", \"में\", and \"की\", which attach to nouns and pronouns to indicate grammatical relationships like possession, location, and object marking. These markers are highly frequent and essential for sentence structure in Hindi.	0.87	0.85	1.00	0.74	0.88	0.86	1.00	0.76
model.layers.0.mlp	Hindi	76485	Suffixes and endings in Hindi, especially \"ों\", \"ा\", and \"े\", are highlighted, often marking verb conjugations, gender, number, or case, and are crucial for grammatical structure and meaning in sentences.	0.80	0.75	1.00	0.60	0.83	0.80	1.00	0.66
model.layers.0.mlp	Hindi	101208	The most salient pattern is the high activation of Hindi vowel diacritics and suffixes such as \"ा\", \"े\", \"ि\", \"ो\", and \"ी\", which are frequently attached to nouns, verbs, and adjectives to indicate case, possession, gender, number, or tense. These morphemes are essential for grammatical structure and meaning in Hindi sentences.	0.72	0.62	0.96	0.46	0.74	0.66	0.96	0.50
model.layers.0.mlp	Hindi	122265	The highlighted tokens are primarily Hindi morphemes, root words, and grammatical particles that are essential for sentence structure and meaning, including common nouns, verbs, pronouns, and connectors, often marking key semantic or syntactic roles within the sentence.	0.71	0.60	0.96	0.44	0.69	0.63	0.79	0.52
model.layers.0.mlp	Hindi	128987	The nasalized vowel suffix \"ेंं\" is highly activated, typically marking the oblique plural or locative case in Hindi nouns, and frequently appears at the end of words to indicate grammatical relationships such as location or plurality. Other nasalized or feminine suffixes like \"ीं\", \"ैं\", and \"ी\" also show activation, reflecting their role in inflectional morphology.	0.69	0.55	1.00	0.38	0.75	0.67	1.00	0.50
model.layers.1.mlp	Hindi	9891	The highlighted tokens are common morphemes or suffixes in German (\"ue\") and Hindi (\"क\", \"की\", \"के\", \"को\", \"से\", etc.), often marking grammatical relationships such as possession, case, or plurality, and are frequently found at word endings or as postpositions.	0.88	0.88	0.91	0.84	0.87	0.87	0.89	0.84
model.layers.1.mlp	Hindi	49724	The tokens correspond to common Hindi postpositions, suffixes, or inflections (such as \"में\", \"से\", \"का\", \"की\", \"दा\", \"ला\", \"पा\", etc.), which are frequently attached to nouns or verbs to indicate grammatical relationships like location, possession, or agency. These morphemes are highly salient for understanding sentence structure and meaning in Hindi text.	0.75	0.68	0.96	0.52	0.75	0.68	0.96	0.52