| Layer | Lang | Feature ID | Interpretation | Detection | | | | Fuzzing | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Accuracy | F1 score | Precision | Recall | Accuracy | F1 score | Precision | Recall |
| model.layers.9.mlp | Hindi | 29761 | The highlighted tokens are primarily suffixes, verb forms, case markers, and common inflections in Hindi, often marking tense, aspect, possession, plurality, or grammatical relationships. These tokens are essential for sentence structure, meaning, and cohesion, frequently appearing at the ends of words or as short connecting elements within sentences. | 0.99 | 0.99 | 1 | 0.98 | 0.88 | 0.88 | 0.88 | 0.88 |
| model.layers.9.mlp | Hindi | 69906 | The highlighted tokens are primarily morphemes, syllables, or word fragments in Hindi, Sanskrit, and related scripts, often marking grammatical, semantic, or phonetic units within words. These include suffixes, prefixes, conjuncts, and root components that are essential for word formation and meaning, as well as some full words or names in other Indic and East Asian languages. The activations focus on linguistically significant subword units that contribute to the structure and interpretation of complex words. | 0.91 | 0.903 | 0.977 | 0.84 | 0.75 | 0.79 | 0.681 | 0.94 |
| model.layers.9.mlp | Hindi | 77447 | The highlighted tokens are common Hindi grammatical morphemes, especially verb suffixes and auxiliary verbs such as \"है\", \"हैं\", \"था\", and case markers like \"का\", \"की\", \"को\", \"में\", \"से\", which are essential for sentence structure, tense, and case relationships in Hindi text. | 0.8 | 0.75 | 1 | 0.6 | 0.81 | 0.765 | 1 | 0.62 |
| model.layers.9.mlp | Hindi | 94557 | The most salient pattern is the frequent activation of single-character Hindi tokens, especially those representing common grammatical particles, case markers, and postpositions such as \"क\", \"स\", \"ह\", \"म\", \"ल\", \"न\", \"द\", \"र\", \"ज\", and \"व\". These tokens are highly important for sentence structure and meaning in Hindi, often marking relationships between words, possession, location, agency, and other grammatical functions. Their prominence reflects their central role in the construction and parsing of Hindi sentences. | 0.73 | 0.63 | 1 | 0.46 | 0.78 | 0.725 | 0.967 | 0.58 |
| model.layers.9.mlp | Hindi | 119704 | The highlighted tokens are primarily Hindi suffixes, conjuncts, and inflections, especially those involving \"्य\", \"िय\", \"ज़्य\", and other matras or half-letters, which are common in forming nouns, adjectives, and participles, often marking grammatical relationships or word forms in Hindi text. | 0.65 | 0.478 | 0.941 | 0.32 | 0.65 | 0.493 | 0.895 | 0.34 |
| model.layers.10.mlp | Hindi | 13080 | High activations are found on common Hindi postpositions, conjunctions, and grammatical markers such as \"के\", \"से\", \"की\", \"का\", \"में\", \"और\", and similar short tokens, which function as connectors or case markers in Hindi syntax. | 0.68 | 0.529 | 1 | 0.36 | 0.63 | 0.431 | 0.933 | 0.28 |
| model.layers.10.mlp | Hindi | 44479 | The highlighted tokens are Hindi verb suffixes and auxiliary verbs, especially forms of \"है/हैं\" and verb endings like \"त\", \"ते\", \"ता\", \"ती\", \"ना\", | 0.95 | 0.947 | 1 | 0.9 | 0.92 | 0.913 | 1 | 0.84 |
| model.layers.10.mlp | Hindi | 49310 | The highlighted tokens are primarily morphemes, suffixes, and root components in Hindi words, often marking grammatical features such as case, number, gender, tense, or forming compound and derived words. These segments frequently appear at the end or within words, contributing to word formation, inflection, and meaning in Hindi morphology. | 0.68 | 0.529 | 1 | 0.36 | 0.75 | 0.675 | 0.963 | 0.52 |
| model.layers.10.mlp | Hindi | 61058 | The highlighted tokens are Hindi verb endings and auxiliary verbs that indicate tense, aspect, and agreement, commonly appearing at the end of clauses or sentences to mark actions, states, or habitual occurrences. | 0.65 | 0.478 | 0.941 | 0.32 | 0.69 | 0.575 | 0.913 | 0.42 |
| model.layers.10.mlp | Hindi | 84378 | The text contains references to the year 1947, the country Pakistan, and related political or administrative terms, often in the context of independence or governance, and these patterns appear across multiple languages and scripts. | 0.52 | 0.077 | 1 | 0.04 | 0.51 | 0.039 | 1 | 0.02 |
| model.layers.10.mlp | Hindi | 87650 | The highlighted tokens are predominantly parts of place names, administrative regions, and institutional names across multiple languages, often marking boundaries between morphemes or components within proper nouns, especially in geographic or official contexts. | 0.47 | 0.404 | 0.462 | 0.36 | 0.43 | 0.26 | 0.37 | 0.2 |
| model.layers.10.mlp | Hindi | 101033 | The highlighted tokens are primarily morphemes, suffixes, and root components in Hindi and related scripts, often marking grammatical roles, inflections, or forming compound words. There is a strong emphasis on function words, affixes, and connectors that are essential for syntactic and semantic structure in the language. | 0.7 | 0.571 | 1 | 0.4 | 0.68 | 0.59 | 0.821 | 0.46 |