Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.2.mlp	Spanish	17631	The highlighted tokens are predominantly Spanish (and some Portuguese) word endings and morphemes, such as verb and noun suffixes (-ar, -ado, -ción, -ión, -er, -os, -es, -ía, -ando, -iendo), as well as common function words (el, la, de, en, y, su, los, las). These patterns reflect grammatical inflections, gender/number agreement, and frequent connectors in Romance languages.	0.67	0.602	0.758	0.5	0.74	0.711	0.8	0.64
model.layers.3.mlp	Spanish	4518	The highlighted tokens are often morphemes, word stems, or affixes from various languages, especially Spanish, marking grammatical or semantic units such as verb endings, noun forms, or prepositions, and sometimes appear in named entities or set phrases.	0.6	0.592	0.604	0.58	0.53	0.657	0.517	0.9
model.layers.5.mlp	Spanish	32141	The highlighted tokens are common Spanish morphemes, suffixes, and inflections that form parts of verbs, nouns, and adjectives, often marking tense, number, gender, or person, and are frequently found at the beginning, middle, or end of words.	0.78	0.725	0.967	0.58	0.62	0.648	0.603	0.7
model.layers.11.mlp	Spanish	94207	The highlighted tokens are predominantly function words, verb forms, pronouns, and common connectors in Spanish, as well as frequent phrase fragments and endings. These elements are essential for sentence structure, tense, and meaning, often marking grammatical relationships, subject/object references, and common idiomatic or formulaic expressions. The activations focus on the connective tissue of the language, reflecting the importance of these tokens in maintaining coherence and fluency in Spanish text.	0.92	0.917	0.957	0.88	0.89	0.884	0.933	0.84
model.layers.13.mlp	Spanish	53088	The highlighted tokens are primarily Spanish morphemes, suffixes, and function words, including endings for adjectives, nouns, and adverbs, as well as conjunctions, prepositions, and common connectors. There is also emphasis on punctuation and proper nouns, reflecting structural and grammatical elements essential for sentence construction and meaning in Spanish text.	0.73	0.64	0.96	0.48	0.71	0.651	0.818	0.54