Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.10.mlp	Portuguese	97091	The highlighted tokens are often morphemes, roots, or affixes within words across multiple languages, especially in Portuguese, Spanish, and related Romance languages, as well as some Slavic and Asian languages. These tokens frequently appear in named entities (such as people, places, and organizations), verb conjugations, noun/adjective endings, and common functional words. The activations tend to focus on linguistically meaningful subword units, including those marking tense, plurality, gender, or forming part of proper nouns and technical terms.	0.52	0.68	0.51	1.00	0.51	0.67	0.51	1.00
model.layers.11.mlp	Portuguese	19293	The highlighted tokens are predominantly roots, stems, or affixes within Portuguese words, often marking the beginning or internal structure of nouns, verbs, and adjectives. These segments are morphologically significant, frequently corresponding to common derivational or inflectional morphemes, and are central to word formation and meaning in the language.	0.80	0.78	0.86	0.72	0.86	0.86	0.88	0.84
model.layers.11.mlp	Portuguese	88127	The highlighted tokens are primarily verb stems, suffixes, and noun roots in Portuguese, often marking key semantic content such as actions, states, or important objects. These tokens frequently appear at the start or within verbs and nouns, indicating morphological boundaries or inflectional changes, and are central to the meaning and structure of the sentence.	0.85	0.82	1.00	0.70	0.77	0.72	0.94	0.58
model.layers.12.mlp	Portuguese	92579	The highlighted tokens are predominantly Portuguese morphemes, suffixes, and function words that form or modify verbs, nouns, and adjectives, as well as common connectors and endings. These elements are essential for grammatical structure, tense, plurality, and meaning in Portuguese sentences.	0.85	0.82	1.00	0.70	0.83	0.81	0.90	0.74
model.layers.12.mlp	Portuguese	125313	The highlighted tokens are predominantly prefixes, roots, or stems of words in Portuguese, often marking the beginning or core of verbs, nouns, and adjectives. These segments are crucial for word formation and meaning, frequently appearing in inflected or derived forms, and are central to the morphological structure of the language.	0.94	0.94	0.96	0.92	0.93	0.93	0.94	0.92
model.layers.13.mlp	Portuguese	18811	The highlighted tokens are frequent function words, pronouns, prepositions, conjunctions, and common noun/adjective endings in Portuguese, often marking grammatical relationships, possession, plurality, or forming part of set phrases and collocations. These tokens are essential for sentence structure and meaning, especially in connecting and specifying relationships between entities.	0.92	0.92	0.98	0.86	0.88	0.88	0.87	0.90
model.layers.13.mlp	Portuguese	34046	The highlighted tokens are predominantly prefixes, roots, and suffixes within Portuguese words, often marking morphological boundaries or derivational elements that contribute to word formation and meaning. These segments frequently appear at the start or end of words, indicating their role in constructing complex words from smaller morphemes.	0.89	0.88	0.95	0.82	0.87	0.87	0.88	0.86
model.layers.13.mlp	Portuguese	34791	The highlighted tokens are predominantly function words, affixes, and common morphemes in Portuguese, such as articles, prepositions, conjunctions, pronouns, and verb endings, as well as frequent noun and adjective suffixes. These elements are essential for grammatical structure, sentence cohesion, and the formation of meaning in the language, often marking relationships between words, tense, number, and gender.	0.92	0.92	0.94	0.90	0.91	0.91	0.90	0.92
model.layers.13.mlp	Portuguese	40526	The highlighted tokens are predominantly Portuguese suffixes, verb endings, and noun/adjective endings that indicate tense, plurality, gender, or word class, as well as some high-frequency function words and punctuation. These patterns reflect morphological markers and grammatical structures central to Portuguese language formation and meaning.	0.90	0.89	1.00	0.80	0.93	0.93	1.00	0.86
model.layers.13.mlp	Portuguese	57410	The highlighted tokens predominantly mark common Portuguese function words, verb forms, and suffixes that construct conditional, temporal, and causal clauses, as well as personal references (especially to \"você\" and \"que\"). There is a strong focus on grammatical connectors, pronouns, and verb endings that are essential for sentence structure and meaning, particularly in instructions, explanations, and descriptions.	0.81	0.77	0.97	0.64	0.83	0.80	1.00	0.66