Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.14.mlp	Portuguese	660	The highlighted tokens are predominantly Portuguese suffixes, verb endings, and noun/adjective inflections that mark tense, number, gender, or degree, as well as common function words and connectors. These elements are crucial for grammatical structure and meaning in Portuguese, often indicating relationships between words, actions, and descriptions within sentences.	0.9	0.891	0.976	0.82	0.91	0.901	1	0.82
model.layers.14.mlp	Portuguese	1650	The highlighted tokens are primarily Portuguese function words, verb endings, and suffixes that indicate tense, aspect, or grammatical relationships, as well as punctuation and conjunctions that structure sentences. There is a strong focus on verb conjugations, prepositions, and connectors that are essential for sentence cohesion and meaning.	0.91	0.903	0.977	0.84	0.92	0.918	0.938	0.9
model.layers.14.mlp	Portuguese	8795	The highlighted tokens are primarily Portuguese morphemes, verb endings, and function words that are essential for grammatical structure and meaning. These include verb conjugations, noun and adjective endings, and common connectors, which are crucial for tense, agreement, and sentence cohesion in Portuguese.	0.96	0.958	1	0.92	0.96	0.958	1	0.92
model.layers.14.mlp	Portuguese	16854	The highlighted tokens are predominantly verb roots or stems in Portuguese, often marking the beginning or core of verbs before inflectional endings are added. These roots are essential for verb conjugation and carry the main semantic content of the verb, frequently appearing in various tenses and forms throughout the text.	0.76	0.692	0.964	0.54	0.89	0.884	0.933	0.84
model.layers.14.mlp	Portuguese	22918	The highlighted tokens are primarily functional morphemes, suffixes, prepositions, conjunctions, and date or time expressions in Portuguese, often marking grammatical relationships, verb conjugations, and temporal references. There is a strong emphasis on endings that indicate tense, plurality, or comparison, as well as on tokens that structure time, quantity, and sequence within sentences.	0.95	0.949	0.959	0.94	0.96	0.959	0.979	0.94
model.layers.14.mlp	Portuguese	30905	The highlighted tokens are primarily function words, verb endings, and common morphemes in Portuguese, such as articles, prepositions, conjunctions, pronouns, and verb or noun suffixes. These elements are essential for grammatical structure, sentence cohesion, and meaning, often marking relationships between phrases, tense, number, or gender.	0.95	0.948	0.979	0.92	0.86	0.87	0.81	0.94
model.layers.14.mlp	Portuguese	35259	The highlighted tokens are primarily function words, common suffixes, and frequent morphemes in Portuguese, as well as high-frequency connectors and grammatical elements that structure sentences, such as conjunctions, pronouns, prepositions, and verb endings. These elements are essential for sentence cohesion and meaning, and often appear at clause or phrase boundaries.	0.95	0.949	0.959	0.94	0.83	0.841	0.789	0.9
model.layers.14.mlp	Portuguese	39613	The highlighted tokens are primarily function words, verb endings, and common morphemes in Portuguese, such as prepositions, conjunctions, pronouns, and verb suffixes, which are essential for grammatical structure and meaning in sentences.	0.92	0.92	0.92	0.92	0.93	0.931	0.922	0.94
model.layers.14.mlp	Portuguese	52522	The highlighted tokens are primarily function words, verb endings, noun and adjective suffixes, and common connectors in Portuguese, often marking grammatical relationships, verb conjugations, and sentence structure, as well as frequent punctuation. These elements are essential for the cohesion and flow of the language, indicating tense, plurality, possession, subordination, and logical connections within and between sentences.	0.84	0.818	0.947	0.72	0.78	0.771	0.804	0.74
model.layers.14.mlp	Portuguese	61856	The highlighted tokens are primarily suffixes, verb endings, and noun/adjective endings common in Portuguese, such as -ção, -idade, -mente, -ar, -er, -ir, and plural or gender markers. These morphemes signal grammatical categories like tense, aspect, number, gender, and part of speech, and are crucial for understanding word formation and syntactic roles in the language.	0.95	0.948	0.979	0.92	0.93	0.928	0.957	0.9
model.layers.14.mlp	Portuguese	73514	The highlighted tokens are predominantly prefixes, roots, or morphemes that form the basis of longer words, often marking the start of nouns, verbs, or adjectives in Portuguese. These segments frequently correspond to meaningful word parts that contribute to the construction of complex or compound words, reflecting morphological structure and semantic building blocks in the language.	0.92	0.913	1	0.84	0.91	0.914	0.873	0.96
model.layers.14.mlp	Portuguese	81345	The highlighted tokens are primarily word endings, suffixes, and punctuation marks that indicate sentence boundaries, grammatical forms, or transitions in Portuguese text. These include verb and noun suffixes, comparative and adverbial endings, and punctuation such as periods, commas, and quotation marks, all of which play a key role in structuring sentences and conveying meaning.	0.94	0.936	1	0.88	0.94	0.938	0.978	0.9