Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.14.mlp	Portuguese	87187	The highlighted tokens are primarily function words, common suffixes, and short morphemes that serve as grammatical connectors or modifiers in Portuguese, such as prepositions, conjunctions, pronouns, and common verb or noun endings. These elements are essential for sentence structure, cohesion, and meaning, often marking relationships between phrases, indicating possession, plurality, tense, or degree, and facilitating the flow of information in the text.	0.96	0.958	1	0.92	0.95	0.949	0.959	0.94
model.layers.14.mlp	Portuguese	99713	The text frequently highlights pronouns and conjunctions, especially those forming common Portuguese phrases such as \"que você\" (that you), \"se você\" (if you), and negative constructions like \"não\" (not). There is a strong focus on verbal inflections, adverbs, and connectors that structure conditional, temporal, or explanatory clauses, as well as endings that form adverbs or comparatives. These patterns reflect the importance of grammatical connectors and personal references in constructing meaning and flow in Portuguese sentences.	0.9	0.889	1	0.8	0.87	0.851	1	0.74
model.layers.14.mlp	Portuguese	114684	The highlighted tokens are predominantly Portuguese suffixes and word endings that indicate verb conjugations, noun and adjective forms, and grammatical inflections, such as tense, number, gender, and person. These include common verb endings (-ar, -ado, -ando, -am, -ou, -ia, -iu, -ava, -ando, -endo, -indo), noun/adjective endings (-ção, -dade, -idade, -mento, -s, -es, -os, -as, -is, -ais, -eira, -eiro), and other morphological markers that are essential for syntactic and semantic structure in the language.	0.92	0.917	0.957	0.88	0.92	0.915	0.977	0.86
model.layers.14.mlp	Portuguese	117527	The highlighted tokens are predominantly Portuguese suffixes and word endings that indicate grammatical features such as gender, number, tense, and part of speech, as well as common noun and adjective forms. These endings are essential for word formation and meaning in Portuguese text.	0.87	0.854	0.974	0.76	0.87	0.857	0.951	0.78
model.layers.14.mlp	Portuguese	127951	The highlighted tokens are predominantly Portuguese morphemes, suffixes, and word endings that indicate grammatical features such as tense, number, gender, and part of speech, as well as common noun and verb roots. These patterns reflect the morphological structure of Portuguese, with frequent emphasis on inflectional endings and affixes that modify meaning and grammatical function.	0.84	0.81	1	0.68	0.8	0.767	0.917	0.66
model.layers.15.mlp	Portuguese	5133	The highlighted tokens are primarily morphemes, suffixes, verb endings, and short function words in Portuguese, often marking grammatical features such as tense, person, number, or gender, as well as common connectors and prepositions. These elements are crucial for sentence structure and meaning, frequently appearing at word endings or as standalone short words, and are essential for the grammatical cohesion and flow of the text.	0.94	0.936	1	0.88	0.87	0.869	0.878	0.86
model.layers.15.mlp	Portuguese	13973	High activations occur on common Portuguese articles, prepositions, and pronouns such as \"o\", \"a\", \"os\", \"um\", \"no\", \"do\", \"ao\", and on adjectives or nouns immediately following them, reflecting the importance of grammatical structure and noun phrase boundaries in the language.	0.79	0.753	0.914	0.64	0.76	0.7	0.933	0.56
model.layers.15.mlp	Portuguese	14772	The highlighted tokens are predominantly prefixes, suffixes, and stems within Portuguese words, often marking grammatical or semantic units such as verb conjugations, noun/adjective endings, or morphemes that contribute to word formation and meaning. These segments frequently appear at the boundaries of words or as part of longer, morphologically complex terms.	0.89	0.884	0.933	0.84	0.83	0.838	0.8	0.88
model.layers.15.mlp	Portuguese	16451	The highlighted tokens are primarily function words, conjunctions, prepositions, pronouns, and common verb endings in Portuguese, as well as frequent noun and adjective suffixes. These tokens are essential for sentence structure, grammatical agreement, and the formation of complex phrases, indicating a focus on the connective and morphological elements that underpin fluent, coherent text in Portuguese.	0.97	0.969	1	0.94	0.94	0.94	0.94	0.94
model.layers.15.mlp	Portuguese	22802	The highlighted tokens are primarily function words, common suffixes, and connectors in Portuguese, such as conjunctions, pronouns, and endings that form adjectives, adverbs, or plurals. These elements are essential for sentence structure, linking clauses, and expressing relationships between ideas, as well as for forming derived or inflected word forms.	0.94	0.939	0.958	0.92	0.86	0.865	0.833	0.9
model.layers.15.mlp	Portuguese	23259	High activations are found on common Portuguese articles and pronouns such as \"uma\", \"a\", \"na\", \"da\", \"as\", and related forms, which function as determiners or refer to feminine nouns, as well as on adjectives and possessives that modify or specify nouns. These tokens are essential for grammatical structure and meaning in Portuguese sentences.	0.96	0.959	0.979	0.94	0.87	0.854	0.974	0.76