Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
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	0.82	0.95	0.72	0.78	0.77	0.80	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	0.95	0.98	0.92	0.93	0.93	0.96	0.90
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	0.91	1.00	0.84	0.91	0.91	0.87	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	0.94	1.00	0.88	0.94	0.94	0.98	0.90
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	0.96	1.00	0.92	0.95	0.95	0.96	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.	0.90	0.89	1.00	0.80	0.87	0.85	1.00	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	0.92	0.96	0.88	0.92	0.92	0.98	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	0.85	0.97	0.76	0.87	0.86	0.95	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	1.00	0.68	0.80	0.77	0.92	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	0.94	1.00	0.88	0.87	0.87	0.88	0.86