Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
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	0.75	0.91	0.64	0.76	0.70	0.93	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	0.88	0.93	0.84	0.83	0.84	0.80	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	0.97	1.00	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	0.94	0.96	0.92	0.86	0.87	0.83	0.90
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	0.96	0.98	0.94	0.87	0.85	0.97	0.76
model.layers.15.mlp	Portuguese	26198	The highlighted tokens are predominantly Portuguese suffixes and word endings that form nouns, adjectives, and verb conjugations, as well as some full words and phrases. These include common morphological markers such as -ção, -dade, -mento, -idade, -ado, -ente, -ista, -ção, -são, -ção, -ções, -ais, -ais, -os, -as, -ia, -io, -iva,	0.94	0.94	1.00	0.88	0.94	0.94	0.98	0.90
model.layers.15.mlp	Portuguese	27629	The highlighted tokens are primarily morphemes, word endings, and short function words in Portuguese, often marking grammatical features such as tense, number, gender, or case, as well as common connectors and noun/adjective endings. These elements are crucial for sentence structure and meaning in the language.	0.98	0.98	1.00	0.96	0.97	0.97	0.98	0.96
model.layers.15.mlp	Portuguese	47463	The highlighted tokens are primarily function words, verb forms, and affixes in Portuguese that are essential for expressing modality, obligation, ability, and temporal or conditional relationships. These include modal verbs, pronouns, prepositions, conjunctions, and verb endings, which are crucial for structuring instructions, expressing necessity or possibility, and connecting clauses in informative or directive contexts.	0.95	0.95	1.00	0.90	0.88	0.88	0.91	0.84
model.layers.15.mlp	Portuguese	49673	The highlighted tokens are primarily function words, verb endings, pronouns, and common connectors in Portuguese, often marking grammatical relationships, verb conjugations, and clause boundaries. These tokens are essential for sentence structure, cohesion, and conveying tense, person, or logical connections within and between clauses.	0.81	0.78	0.94	0.66	0.86	0.85	0.89	0.82
model.layers.15.mlp	Portuguese	74006	The highlighted tokens are predominantly prefixes, roots, or stems within Portuguese words, often marking the beginning or core of nouns, verbs, or adjectives, and are frequently associated with technical, formal, or compound terms. These segments are crucial for word formation and meaning, especially in morphologically rich contexts.	0.89	0.88	0.95	0.82	0.93	0.93	0.89	0.98
model.layers.15.mlp	Portuguese	83064	The highlighted tokens frequently mark grammatical morphemes, noun and verb endings, prepositions, conjunctions, and common phrase boundaries in Romance languages, often signaling syntactic structure, tense, plurality, or idiomatic expressions.	0.67	0.70	0.64	0.78	0.57	0.70	0.54	0.98
model.layers.15.mlp	Portuguese	106475	The highlighted tokens are primarily Portuguese suffixes and word endings that form nouns, adjectives, and verb conjugations, as well as common prepositions, conjunctions, and function words. These patterns reflect morphological structures (such as -ção, -idade, -amento, -ar, -ado, -ente, -ista, -izar, -izarão, -ante, -ente, -ivo, -eira, -eiro, -oso, -ável, -ível,	0.94	0.94	0.98	0.90	0.97	0.97	0.98	0.96