Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.0.mlp	Portuguese	128645	The preposition \"no\" is frequently activated when indicating location, time, or context within a noun phrase, often specifying where or when something occurs in Portuguese text.	0.59	0.328	0.909	0.2	0.56	0.241	0.875	0.14
model.layers.2.mlp	Portuguese	51039	The highlighted tokens are primarily function words, common suffixes, and morphemes in Portuguese and related languages, including articles, conjunctions, prepositions, and endings that form nouns, adjectives, or verb conjugations. These elements are essential for grammatical structure and meaning, often marking gender, number, tense, or connecting phrases.	0.65	0.557	0.759	0.44	0.52	0.5	0.522	0.48
model.layers.3.mlp	Portuguese	129688	The highlighted tokens are primarily morphemes, suffixes, or short word fragments in Portuguese (and some in other languages), often marking grammatical features such as gender, number, tense, or forming part of proper nouns and common words. Many are accented characters or endings that are important for word formation and meaning in the context of Romance languages.	0.69	0.644	0.757	0.56	0.67	0.713	0.631	0.82
model.layers.4.mlp	Portuguese	51358	The highlighted tokens are morphemes, word roots, or affixes in multiple languages, often marking grammatical features such as tense, case, gender, or plurality, or forming part of proper nouns and place names. These tokens frequently appear at word boundaries or within compound words, reflecting their importance in word formation and meaning across diverse linguistic contexts.	0.56	0.627	0.544	0.74	0.53	0.671	0.516	0.96
model.layers.4.mlp	Portuguese	63167	The highlighted tokens are primarily morphemes, suffixes, or roots within words in Portuguese (and some Spanish, Russian, and English), often marking verb conjugations, noun/adjective endings, or forming part of proper names and place names. These segments are linguistically significant for inflection, derivation, or identification of entities.	0.72	0.736	0.696	0.78	0.59	0.687	0.556	0.9
model.layers.5.mlp	Portuguese	23602	The highlighted tokens are common morphemes, suffixes, prefixes, or inflectional endings in Portuguese, often marking verb conjugations, noun/adjective forms, or grammatical gender and number. These elements are crucial for word formation and meaning in the language.	0.61	0.418	0.824	0.28	0.56	0.45	0.6	0.36
model.layers.7.mlp	Portuguese	118682	The highlighted tokens consistently correspond to the morphemes or substrings forming the country name \"Brazil\" and its derivatives across multiple languages and scripts, often marking the root or core segment of the word regardless of linguistic context.	0.56	0.241	0.875	0.14	0.6	0.333	1	0.2
model.layers.8.mlp	Portuguese	27987	The highlighted tokens are predominantly morphemes, affixes, or root segments within Portuguese words, often marking grammatical, semantic, or syntactic boundaries such as verb conjugations, noun/adjective endings, and common prefixes or suffixes. These segments frequently appear at the start, middle, or end of words and are crucial for word formation and meaning in the language.	0.72	0.611	1	0.44	0.62	0.537	0.688	0.44
model.layers.9.mlp	Portuguese	77250	The highlighted tokens are predominantly prefixes, suffixes, or stems within Portuguese words, often marking verb conjugations, noun/adjective forms, or common morphemes. These segments frequently appear at the start or end of words, reflecting morphological structure and inflectional patterns in the language.	0.73	0.649	0.926	0.5	0.69	0.652	0.744	0.58
model.layers.9.mlp	Portuguese	99643	The highlighted tokens are primarily stems, roots, or affixes of words in Portuguese (and some in Spanish, French, German, and Hindi), often marking verb conjugations, noun/adjective forms, or place names. These tokens frequently appear at morpheme boundaries, within inflected or derived forms, and in multiword proper nouns or locations, reflecting the morphological structure and multilingual context of the text.	0.7	0.746	0.647	0.88	0.67	0.744	0.608	0.96
model.layers.10.mlp	Portuguese	3658	The highlighted tokens are predominantly morphemes, roots, prefixes, and suffixes within Portuguese words, often marking word formation, inflection, or derivation. These segments frequently appear at the start, middle, or end of words and are crucial for constructing meaning, tense, plurality, or grammatical function in the language.	0.94	0.939	0.958	0.92	0.93	0.931	0.922	0.94
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	0.676	0.51	1	0.51	0.671	0.505	1