Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.13.mlp	Portuguese	61518	The highlighted tokens are predominantly Portuguese suffixes and word endings that indicate grammatical features such as tense, number, gender, and part of speech, as well as common noun and adjective forms, and some function words. These patterns reflect morphological markers and frequent word constructions in Portuguese text.	0.87	0.85	0.97	0.76	0.86	0.84	0.95	0.76
model.layers.13.mlp	Portuguese	71681	The highlighted tokens frequently correspond to morphemes, suffixes, or short function words in Portuguese, often marking verb conjugations, noun/adjective endings, or linking words. There is a strong emphasis on grammatical markers, sentence boundaries, and common connectors, reflecting the structure and flow of natural Portuguese text.	0.88	0.86	1.00	0.76	0.81	0.80	0.84	0.76
model.layers.13.mlp	Portuguese	81450	The highlighted tokens are primarily function words, verb endings, and common morphemes in Portuguese, such as verb conjugations, prepositions, conjunctions, and pronouns. These elements are essential for grammatical structure, sentence cohesion, and conveying relationships between ideas, actions, and entities within the text.	0.91	0.90	1.00	0.82	0.79	0.79	0.78	0.80
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.	0.90	0.89	0.98	0.82	0.91	0.90	1.00	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	0.90	0.98	0.84	0.92	0.92	0.94	0.90
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	0.96	1.00	0.92	0.96	0.96	1.00	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	0.69	0.96	0.54	0.89	0.88	0.93	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	0.95	0.96	0.94	0.96	0.96	0.98	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	0.95	0.98	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	0.95	0.96	0.94	0.83	0.84	0.79	0.90
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	0.93	0.92	0.94