Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.1.mlp	Italian	28636	The highlighted tokens are primarily Italian morphemes, prefixes, suffixes, and short function words, often marking word boundaries, inflections, or grammatical constructs. These include prepositions, articles, conjunctions, verb endings, and noun/adjective suffixes, as well as common roots and stems, reflecting the morphological structure and syntactic function within Italian sentences.	0.57	0.246	1	0.14	0.54	0.361	0.591	0.26
model.layers.1.mlp	Italian	47594	The highlighted tokens are predominantly prefixes, roots, or suffixes within Italian words, often marking morphological boundaries or meaningful subword units, such as verb endings, noun roots, or common affixes. These segments frequently correspond to points of high linguistic or semantic relevance within the word structure.	0.52	0.2	0.6	0.12	0.59	0.577	0.596	0.56
model.layers.2.mlp	Italian	33390	The highlighted tokens are predominantly suffixes, inflections, or morphemes from Romance and other European languages, often marking grammatical features such as gender, number, tense, or forming part of proper nouns and place names. These tokens frequently appear at the end of words, especially in multilingual contexts, and are important for identifying linguistic structure, word formation, and named entities.	0.65	0.729	0.595	0.94	0.6	0.649	0.578	0.74
model.layers.3.mlp	Italian	120200	The highlighted tokens are often morphemes, suffixes, or name fragments within words, especially in Italian and other Romance languages, marking comparative forms, diminutives, or parts of proper nouns and loanwords. These tokens frequently appear at word endings or within multi-syllabic names and terms.	0.73	0.765	0.677	0.88	0.61	0.698	0.57	0.9
model.layers.4.mlp	Italian	114161	The highlighted tokens are predominantly suffixes, inflections, or name fragments common in Romance and other European languages, often marking person names, place names, or grammatical endings, and sometimes appearing in transliterated or non-Latin scripts.	0.56	0.662	0.537	0.86	0.57	0.681	0.541	0.92
model.layers.6.mlp	Italian	109638	The highlighted tokens are predominantly Italian (with some other Romance and European language examples) and often correspond to morphological affixes, verb endings, noun/adjective suffixes, and function words. These tokens frequently mark grammatical features such as tense, number, gender, or case, and are often found at the end or within words, indicating their role in word formation and inflection.	0.8	0.796	0.812	0.78	0.61	0.667	0.582	0.78
model.layers.7.mlp	Italian	52987	The highlighted tokens are predominantly Italian morphemes, prefixes, suffixes, or root fragments that form or modify words, often marking verb conjugations, noun/adjective forms, or connecting elements within compound words. These fragments are crucial for word formation and grammatical structure in Italian.	0.89	0.879	0.976	0.8	0.81	0.816	0.792	0.84
model.layers.8.mlp	Italian	13547	The highlighted tokens are predominantly Italian morphemes, roots, and affixes that form the core of words, especially those indicating tense, plurality, or word families. These include verb endings, noun and adjective suffixes, and common prefixes, often marking grammatical or semantic relationships within sentences.	0.58	0.276	1	0.16	0.57	0.358	0.706	0.24
model.layers.9.mlp	Italian	5106	The highlighted tokens are predominantly Italian word stems, prefixes, and suffixes, often marking verb conjugations, noun/adjective derivations, or common morphemes. These segments frequently appear at the start or within words, reflecting the agglutinative and inflectional nature of Italian morphology. The activations focus on linguistically meaningful subword units that contribute to word formation and grammatical function.	0.56	0.267	0.8	0.16	0.51	0.246	0.533	0.16
model.layers.9.mlp	Italian	58424	The highlighted tokens are predominantly Italian word endings, suffixes, and inflections that mark verb conjugations, noun/adjective endings, and grammatical agreement, as well as some function words and punctuation. These patterns reflect the morphological structure of Italian, where meaning and grammatical roles are often encoded in word endings.	0.84	0.81	1	0.68	0.87	0.854	0.974	0.76