Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.10.mlp	Italian	38479	The highlighted tokens are predominantly Italian morphological suffixes and inflections, including verb endings, noun and adjective endings, and common word fragments that indicate tense, number, gender, or part of speech. These patterns reflect the agglutinative and inflectional nature of Italian grammar, where meaning and grammatical function are often encoded in word endings.	0.65	0.462	1	0.3	0.73	0.649	0.926	0.5
model.layers.10.mlp	Italian	108570	The highlighted tokens are predominantly Italian word stems, prefixes, and suffixes, often marking verb conjugations, noun/adjective derivations, or forming compound words. These segments are crucial for morphological construction and meaning in Italian, frequently appearing at the start or end of words to indicate tense, plurality, or word family.	0.74	0.649	1	0.48	0.84	0.818	0.947	0.72
model.layers.11.mlp	Italian	14702	The highlighted segments are common morphemes, roots, or affixes in Italian, often marking verb conjugations, noun/adjective derivations, or forming part of compound words. These segments frequently appear at the end or within words, reflecting productive morphological patterns in Italian word formation.	0.82	0.786	0.971	0.66	0.88	0.878	0.896	0.86
model.layers.11.mlp	Italian	35463	The highlighted tokens are predominantly Italian word endings and suffixes, often marking grammatical features such as gender, number, tense, or part of speech (e.g., -zione, -mente, -ato, -ità, -are, -ente, -ale, -ico, -ista, -zione, -aggio, -ore, -enza, -ità, -zione, -mento, -ante, -ato, -ivo, -ente, -ale, -oso, -ista, -ore, -zione, -ità,	0.53	0.254	0.615	0.16	0.56	0.267	0.8	0.16
model.layers.12.mlp	Italian	3234	The highlighted tokens are predominantly Italian word endings, especially common suffixes such as -zione, -mente, -ità, -zione,	0.55	0.182	1	0.1	0.53	0.113	1	0.06
model.layers.13.mlp	Italian	30409	The highlighted tokens are predominantly Italian noun and adjective suffixes, verb endings, and common word endings that indicate grammatical categories such as gender, number, tense, and degree, as well as some high-frequency nouns and adverbs. These endings are essential for word formation and inflection in Italian, marking parts of speech and syntactic roles.	0.9	0.889	1	0.8	0.91	0.903	0.977	0.84
model.layers.13.mlp	Italian	44286	The highlighted tokens are primarily Italian verb and noun suffixes, conjunctions, and prepositions, often marking grammatical tense, aspect, or function within sentences. These include verb endings for infinitive, gerund, participle, and past tense, as well as common connectors and punctuation, reflecting the morphological and syntactic structure of Italian text.	0.92	0.913	1	0.84	0.86	0.848	0.929	0.78
model.layers.14.mlp	Italian	90522	The highlighted tokens are predominantly suffixes or endings of words in Italian and other languages, often marking grammatical features such as gender, number, tense, or case, as well as forming nouns, adjectives, and participles. These endings are crucial for word formation and meaning in morphologically rich languages.	0.6	0.655	0.576	0.76	0.7	0.754	0.639	0.92