Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.0.mlp	Bulgarian	123218	The text highlights the frequent use of single Cyrillic letters, especially at the beginning of sentences or phrases, functioning as prepositions, conjunctions, or initials in Russian and Bulgarian, often with high activation when capitalized and sentence-initial.	0.76	0.7	0.933	0.56	0.84	0.814	0.972	0.7
model.layers.1.mlp	Bulgarian	32578	The highlighted tokens are common Bulgarian suffixes used to form adjectives, nouns, and participles, indicating grammatical features such as gender, number, definiteness, and case. These suffixes are essential for word formation and inflection in Bulgarian morphology.	0.73	0.64	0.96	0.48	0.83	0.8	0.971	0.68
model.layers.3.mlp	Bulgarian	98476	The highlighted tokens are predominantly Bulgarian suffixes, pronoun and conjunction forms, and inflectional endings that mark grammatical relationships such as gender, number, case, and verb tense, as well as function words that connect clauses or indicate possession and agency. These elements are essential for the syntactic structure and meaning in Bulgarian sentences.	0.82	0.795	0.921	0.7	0.77	0.753	0.814	0.7
model.layers.4.mlp	Bulgarian	3143	The highlighted tokens are predominantly Bulgarian morphemes and suffixes that serve grammatical functions, such as forming participles, plurals, or indicating gender, number, and case. There is a strong focus on the relative pronoun \"който\" (who/which/that), as well as common noun and verb endings, and function words that are essential for sentence structure and meaning in Bulgarian.	0.9	0.889	1	0.8	0.93	0.925	1	0.86
model.layers.4.mlp	Bulgarian	82561	The highlighted tokens are morphemes, word stems, or suffixes that form parts of proper nouns, place names, or grammatical endings across multiple languages, often marking nationality, location, or grammatical case.	0.57	0.517	0.59	0.46	0.49	0.585	0.493	0.72
model.layers.4.mlp	Bulgarian	92607	The highlighted tokens are morphemes, suffixes, or name fragments that are significant in identifying proper nouns, place names, and grammatical forms across multiple languages, especially in Slavic, Romance, and Turkic contexts. These tokens often mark inflections, diminutives, or are parts of multi-token named entities, and are important for language identification, morphological analysis, and named entity recognition.	0.64	0.667	0.621	0.72	0.54	0.681	0.521	0.98
model.layers.5.mlp	Bulgarian	949	The highlighted tokens are predominantly Bulgarian morphemes, word stems, and suffixes that form the core of nouns, verbs, and adjectives, often marking grammatical features such as tense, number, gender, or case. These segments frequently appear at the beginning or end of words, indicating their role in word formation and inflection, and are essential for the syntactic and semantic structure of Bulgarian sentences.	0.68	0.543	0.95	0.38	0.71	0.613	0.92	0.46
model.layers.5.mlp	Bulgarian	12333	The highlighted tokens are predominantly suffixes, prefixes, or inflectional endings in Bulgarian (and occasionally other languages), marking grammatical features such as tense, number, gender, case, or aspect, as well as forming participles, adjectives, and nouns. These morphemes are essential for word formation and grammatical agreement in the language.	0.82	0.795	0.921	0.7	0.69	0.693	0.686	0.7
model.layers.5.mlp	Bulgarian	30265	The highlighted tokens are predominantly suffixes, inflections, or short morphemes in Bulgarian (and some other languages), such as grammatical endings for tense, case, gender, number, or diminutives, as well as common function words and particles. These elements are crucial for the grammatical structure and meaning of words and sentences.	0.87	0.854	0.974	0.76	0.73	0.761	0.683	0.86
model.layers.6.mlp	Bulgarian	99108	The highlighted tokens are common Bulgarian suffixes, prepositions, conjunctions, and pronouns, often marking grammatical relationships such as possession, case, verb tense, or plurality, as well as forming parts of function words and endings that are essential for sentence structure and meaning.	0.94	0.938	0.978	0.9	0.88	0.885	0.852	0.92
model.layers.6.mlp	Bulgarian	112351	The highlighted tokens are common Bulgarian suffixes, pronouns, and function words, often marking grammatical features such as case, gender, number, tense, or forming parts of verbs and nouns. These elements are essential for the structure and meaning of Bulgarian sentences.	0.93	0.926	0.978	0.88	0.89	0.893	0.868	0.92
model.layers.6.mlp	Bulgarian	126200	The highlighted tokens are predominantly suffixes, prefixes, or inflectional endings in various languages, especially Bulgarian, marking grammatical features such as tense, case, number, gender, or aspect, as well as function words and morphemes that contribute to word formation and syntactic structure.	0.76	0.774	0.732	0.82	0.62	0.708	0.575	0.92