Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.14.mlp	Bulgarian	11396	The highlighted tokens are predominantly function words, verb forms, pronouns, conjunctions, and common suffixes or prefixes in Bulgarian, often marking grammatical relationships, tense, negation, or person. These tokens are essential for sentence structure, coherence, and meaning, frequently appearing at clause boundaries or as connectors within and between phrases.	0.89	0.876	1	0.78	0.91	0.901	1	0.82
model.layers.14.mlp	Bulgarian	14651	The highlighted tokens are predominantly Bulgarian morphemes, suffixes, and inflections that mark grammatical features such as tense, number, gender, case, and aspect, as well as common function words and short stems. These elements are crucial for the syntactic and semantic structure of Bulgarian sentences, often appearing at the ends of words or as short connecting words, reflecting the language's rich inflectional morphology.	0.85	0.831	0.949	0.74	0.85	0.831	0.949	0.74
model.layers.14.mlp	Bulgarian	17319	The highlighted tokens are predominantly morphemes, roots, prefixes, and suffixes within Bulgarian (and some other Slavic) words, as well as some in other languages, that contribute to the grammatical structure and meaning of verbs, nouns, and adjectives. These include verb stems, inflectional endings, and function words that are essential for forming questions, commands, and statements, as well as expressing tense, aspect, and modality. The pattern reflects a focus on the internal structure of words and the functional elements that drive sentence construction and meaning in morphologically rich languages.	0.96	0.958	1	0.92	0.7	0.766	0.628	0.98
model.layers.14.mlp	Bulgarian	68900	The highlighted tokens are predominantly Bulgarian function words, suffixes, and short morphemes that serve grammatical roles such as case, number, tense, and prepositions, as well as common noun and adjective endings. These elements are essential for sentence structure and meaning, reflecting the importance of morphology and syntactic connectors in Bulgarian text.	0.93	0.928	0.957	0.9	0.93	0.929	0.939	0.92
model.layers.14.mlp	Bulgarian	112671	The highlighted tokens are predominantly Bulgarian morphemes, function words, and common inflections, including verb endings, pronouns, conjunctions, and prepositions. These tokens often appear at clause or sentence boundaries, within idiomatic expressions, or as part of frequently used grammatical constructions, reflecting their high utility in structuring and connecting ideas in Bulgarian text.	0.73	0.63	1	0.46	0.75	0.667	1	0.5
model.layers.14.mlp	Bulgarian	117211	The highlighted tokens are predominantly Bulgarian morphemes, suffixes, and inflections that mark grammatical features such as tense, number, gender, case, and aspect, as well as common noun and verb endings. These elements are crucial for the syntactic and semantic structure of Bulgarian sentences, often appearing at the ends of words to indicate relationships, roles, or actions within the sentence.	0.92	0.913	1	0.84	0.87	0.857	0.951	0.78
model.layers.15.mlp	Bulgarian	119682	The text contains frequent use of Bulgarian function words and particles such as conjunctions, pronouns, and auxiliary verbs, especially forms of \"да\" (to), \"се\" (oneself), \"не\" (not), and prepositions, which are essential for constructing clauses, expressing modality, and forming complex verb phrases. These tokens are highly activated due to their grammatical importance in sentence structure and meaning.	0.75	0.675	0.963	0.52	0.77	0.716	0.935	0.58