Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.12.mlp	Bulgarian	53983	The highlighted tokens are predominantly Bulgarian morphemes, word stems, and affixes that form the core meaning or grammatical structure of words. These include verb roots, noun and adjective endings, prefixes, and suffixes, often marking tense, aspect, plurality, definiteness, or case. The activations focus on the most semantically or syntactically informative segments within words or short phrases, reflecting the morphological richness and agglutinative nature of Bulgarian, where meaning is built up from smaller meaningful units.	0.98	0.98	1	0.96	0.95	0.949	0.959	0.94
model.layers.12.mlp	Bulgarian	80482	The highlighted tokens are predominantly short function words, suffixes, and inflections common in Bulgarian (and some in other languages), such as pronouns, prepositions, conjunctions, verb endings, and noun/adjective suffixes. These elements are essential for grammatical structure, agreement, and meaning in morphologically rich languages, often marking case, number, gender, tense, or syntactic relationships.	0.92	0.918	0.938	0.9	0.83	0.847	0.77	0.94
model.layers.12.mlp	Bulgarian	90069	The prefix \"вы\" in Russian is highly activated, typically marking the beginning of verbs or nouns, often forming words related to actions, results, or benefits, and is productive across a wide range of derivations.	0.57	0.246	1	0.14	0.54	0.148	1	0.08
model.layers.13.mlp	Bulgarian	47240	The highlighted tokens are predominantly function words, verb forms, pronouns, conjunctions, and common morphemes in Bulgarian, often marking grammatical relationships, verb conjugations, and question or relative clauses. These tokens are essential for sentence structure, tense, aspect, and the formation of complex or subordinate clauses.	0.96	0.96	0.96	0.96	0.93	0.931	0.922	0.94
model.layers.13.mlp	Bulgarian	68656	The highlighted tokens are predominantly inflectional suffixes and short morphemes in Bulgarian and other languages, marking tense, aspect, person, number, or case, as well as short function words and punctuation, which are essential for grammatical structure and meaning.	0.58	0.704	0.543	1	0.55	0.69	0.526	1
model.layers.13.mlp	Bulgarian	77032	The highlighted tokens are predominantly morphemes, word stems, suffixes, and short function words in Bulgarian, often marking grammatical relationships, inflections, or forming parts of compound or derived words. These elements are crucial for constructing meaning, indicating tense, case, number, or connecting phrases, and are frequently found at word boundaries or as parts of longer words.	0.91	0.903	0.977	0.84	0.87	0.871	0.863	0.88
model.layers.13.mlp	Bulgarian	90387	The highlighted tokens are predominantly Bulgarian function words, common suffixes, and inflectional endings that mark grammatical relationships such as case, number, gender, tense, and definiteness, as well as frequent prepositions, conjunctions, and pronouns. These elements are essential for the syntactic structure and meaning in Bulgarian sentences.	0.97	0.969	1	0.94	0.92	0.923	0.889	0.96
model.layers.13.mlp	Bulgarian	91108	The text highlights the importance of common Bulgarian prefixes such as \"из\", \"под\", \"раз\", \"съ\", \"при\", and others, which frequently appear at the beginning of words to form verbs, nouns, or adjectives. These prefixes are key morphological markers that modify the meaning of root words and are highly salient in the structure and semantics of Bulgarian language.	0.63	0.448	0.882	0.3	0.64	0.486	0.85	0.34
model.layers.13.mlp	Bulgarian	103724	The highlighted tokens are predominantly function words, prepositions, conjunctions, pronouns, and common morphemes or affixes in Bulgarian, as well as frequent noun and verb roots. These elements are essential for grammatical structure, sentence cohesion, and the formation of complex word forms, indicating a focus on the connective and structural components of the language.	0.84	0.822	0.925	0.74	0.9	0.898	0.917	0.88
model.layers.13.mlp	Bulgarian	119333	The highlighted tokens are predominantly morphemes, roots, and affixes within Bulgarian words, often marking derivational or inflectional changes (such as forming nouns, adjectives, or verbs). These segments frequently correspond to meaningful subword units that contribute to the grammatical structure or semantic core of the word, including common suffixes, prefixes, and stems.	0.96	0.958	1	0.92	0.93	0.929	0.939	0.92
model.layers.13.mlp	Bulgarian	125029	The text highlights the use of Bulgarian superlative and comparative constructions, especially the prefix for \"most\" or \"best\" attached to adjectives and adverbs, often joined with a hyphen, as well as common suffixes for forming comparatives and superlatives. There is also frequent activation on noun and verb roots, and on morphemes that modify meaning or degree.	0.61	0.381	0.923	0.24	0.67	0.507	1	0.34