Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.7.mlp	Bulgarian	83448	The highlighted tokens are suffixes, inflections, or short morphemes in various languages (notably Bulgarian and related Slavic languages), as well as fragments of country or place names, and abbreviations for years, indicating a focus on linguistically meaningful subword units, grammatical endings, and named entities.	0.64	0.719	0.59	0.92	0.56	0.694	0.532	1
model.layers.7.mlp	Bulgarian	116104	The character \"ъ\" in Bulgarian is highly activated, especially when appearing within or at the end of words, often as a root or inflectional vowel. Other activations include common Bulgarian morphemes and suffixes, such as those forming participles, comparatives, or noun/adjective endings. The pattern centers on the morphological and phonological significance of \"ъ\" and related morphemes in Bulgarian word formation.	0.83	0.809	0.923	0.72	0.77	0.701	1	0.54
model.layers.8.mlp	Bulgarian	99889	The highlighted tokens are primarily Bulgarian morphemes, suffixes, and function words, especially those forming relative pronouns (e.g., \"която\", \"който\"), comparative or grammatical endings (e.g., \"ъ\", \"г\"), and other short connective elements. These tokens are important for marking grammatical relationships, tense, and reference within sentences.	0.94	0.938	0.978	0.9	0.95	0.95	0.941	0.96
model.layers.8.mlp	Bulgarian	119082	The highlighted tokens are predominantly Bulgarian morphemes, suffixes, and function words, especially those forming relative pronouns, verb endings, and noun/adjective inflections. There is a strong focus on morphological boundaries and grammatical markers that are essential for syntactic structure and meaning in Bulgarian text.	0.91	0.901	1	0.82	0.92	0.913	1	0.84
model.layers.9.mlp	Bulgarian	6073	The highlighted tokens are primarily short morphemes, suffixes, or function words in various languages, often marking grammatical features such as case, tense, plurality, or forming part of proper nouns and place names. These tokens frequently appear at word endings or as standalone grammatical markers, reflecting their importance in morphological and syntactic structure.	0.52	0.642	0.512	0.86	0.48	0.649	0.49	0.96
model.layers.9.mlp	Bulgarian	10192	The highlighted tokens are predominantly morphemes, suffixes, prefixes, or short stems within words across multiple languages, often marking grammatical features such as tense, case, plurality, or forming part of proper nouns and place names. These segments are crucial for word formation and meaning, and their activation suggests a focus on subword units that carry significant syntactic or semantic information.	0.53	0.68	0.515	1	0.53	0.68	0.515	1
model.layers.9.mlp	Bulgarian	41496	The highlighted tokens are predominantly Bulgarian morphemes, word stems, and affixes, especially those involving the Cyrillic letter \"ъ\" and other common inflectional or derivational elements. These tokens often mark grammatical features such as tense, aspect, person, or case, and are crucial for word formation and meaning in Bulgarian text.	0.97	0.97	0.98	0.96	0.98	0.98	1	0.96
model.layers.9.mlp	Bulgarian	72400	The highlighted tokens are predominantly Bulgarian morphemes, word stems, and suffixes that form the core of nouns, verbs, and adjectives, as well as grammatical markers for tense, person, and case. These include common inflectional endings, prefixes, and roots that are essential for word formation and meaning in Bulgarian, often marking possession, plurality, aspect, or derivation. The activations focus on linguistically significant subword units that contribute to the syntactic and semantic structure of the language.	0.97	0.969	1	0.94	0.97	0.97	0.961	0.98
model.layers.9.mlp	Bulgarian	102864	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 tokens 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.	0.81	0.765	1	0.62	0.83	0.8	0.971	0.68