Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.10.mlp	Bulgarian	11345	The highlighted tokens are predominantly Bulgarian morphemes, suffixes, and stems that form grammatical endings, verb conjugations, noun/adjective inflections, and function words. These elements are crucial for syntactic structure, agreement, and meaning in Bulgarian sentences, often marking tense, number, gender, case, or aspect.	0.97	0.97	0.98	0.96	0.96	0.96	0.96	0.96
model.layers.10.mlp	Bulgarian	37904	The highlighted tokens predominantly mark common Bulgarian morphemes, especially suffixes and inflections that form plurals, adjectives, comparatives, and verb conjugations, as well as frequent roots and prefixes. There is a strong focus on the morphological structure of words, particularly endings like -ич, -ен, -ат, -ия, -на, and -ко, which are essential for grammatical agreement and meaning in Bulgarian.	0.67	0.522	0.947	0.36	0.67	0.535	0.905	0.38
model.layers.10.mlp	Bulgarian	41383	The highlighted tokens are predominantly proper nouns, especially names of people, places, and media titles, often appearing in multiple languages and scripts. These tokens frequently occur in contexts involving legal, journalistic, or entertainment references, and are often accompanied by titles, roles, or institutional affiliations.	0.48	0.395	0.472	0.34	0.45	0.068	0.222	0.04
model.layers.10.mlp	Bulgarian	43549	The highlighted tokens are primarily function words, pronouns, conjunctions, prepositions, and common verb forms in Bulgarian (and some in other languages), as well as frequent morphemes and endings. These elements are essential for sentence structure, grammatical relationships, and meaning, often marking tense, negation, possession, or subordination. The activations focus on the connective tissue of language that enables coherent and contextually appropriate communication.	0.63	0.661	0.61	0.72	0.51	0.637	0.506	0.86
model.layers.10.mlp	Bulgarian	116770	The highlighted tokens are primarily Bulgarian suffixes, word endings, and compound forms that indicate grammatical features such as tense, number, gender, and case, as well as units of time, quantities, and object types. These patterns reflect morphological markers for nouns, adjectives, and verbs, and often denote time periods, ownership, professions, or collective/abstract concepts.	0.73	0.64	0.96	0.48	0.76	0.692	0.964	0.54
model.layers.11.mlp	Bulgarian	52975	The highlighted tokens are predominantly short function words, prefixes, or single letters from various languages, often marking grammatical relationships, verb forms, or serving as conjunctions and prepositions. These tokens are crucial for sentence structure and meaning, especially in morphologically rich or agglutinative languages.	0.55	0.685	0.527	0.98	0.57	0.695	0.538	0.98
model.layers.11.mlp	Bulgarian	117002	The highlighted tokens are predominantly Bulgarian morphemes, word stems, suffixes, and inflections that form the core of nouns, verbs, and adjectives, as well as function words and conjunctions. These tokens often mark grammatical relationships, derivational processes, and key semantic units within sentences, reflecting the structure and morphology of Bulgarian language.	0.95	0.948	0.979	0.92	0.95	0.948	0.979	0.92
model.layers.12.mlp	Bulgarian	16648	The highlighted tokens are predominantly suffixes or stems in Bulgarian that form adjectives, nouns, or participles, often indicating grammatical features such as gender, number, or case, and are especially common in forming complex or compound words. These morphemes frequently appear at the end of words and are central to word formation and inflection in the language.	0.7	0.583	0.955	0.42	0.73	0.658	0.897	0.52
model.layers.12.mlp	Bulgarian	41511	The highlighted tokens are predominantly morphemes, roots, or affixes within words in Slavic and some Turkic languages, often marking grammatical features such as case, number, gender, tense, or forming nouns and adjectives. These segments frequently appear at word boundaries or as part of compound or derived words, reflecting the agglutinative and inflectional nature of these languages.	0.87	0.863	0.911	0.82	0.71	0.752	0.657	0.88
model.layers.12.mlp	Bulgarian	42598	The highlighted tokens are predominantly function words, verb forms, noun and adjective endings, and common morphemes in Bulgarian, often marking grammatical relationships, verb conjugations, noun/adjective inflections, and clause boundaries. These tokens are crucial for sentence structure, agreement, and meaning, frequently appearing at the start or end of words, and are essential for parsing and understanding Bulgarian syntax and morphology.	0.9	0.891	0.976	0.82	0.89	0.882	0.953	0.82
model.layers.12.mlp	Bulgarian	43375	The highlighted tokens are predominantly verb roots, stems, or affixes in Bulgarian, often marking present, past, or participle forms. These segments frequently appear at morpheme boundaries, especially at the end of verbs or within verb conjugations, indicating a focus on verbal morphology and inflectional patterns.	0.73	0.63	1	0.46	0.89	0.882	0.953	0.82