Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.4.mlp	Russian	48343	The highlighted tokens are primarily short morphemes, prefixes, or function words in Slavic and related languages, often marking grammatical relationships, word formation, or serving as connectors within or between words and phrases.	0.75	0.771	0.712	0.84	0.55	0.667	0.529	0.9
model.layers.6.mlp	Russian	121507	The highlighted tokens are predominantly morphemes, roots, prefixes, suffixes, or short function words within Russian, Bulgarian, and related Slavic texts. These segments often mark grammatical features, word formation, or serve as connectors, and are frequently found at the beginning, middle, or end of words, reflecting the agglutinative and inflectional nature of these languages.	0.91	0.909	0.918	0.9	0.83	0.838	0.8	0.88
model.layers.7.mlp	Russian	110894	The highlighted tokens are predominantly morphemes, word stems, suffixes, and function words in Russian (and some other Slavic languages), often marking grammatical features such as case, number, tense, aspect, or forming participles and adjectives. There is a strong focus on inflectional and derivational morphology, as well as on common connectors and particles that structure sentences.	0.94	0.938	0.978	0.9	0.86	0.865	0.833	0.9
model.layers.8.mlp	Russian	69357	The highlighted tokens are predominantly Russian word endings, suffixes, and inflections that indicate grammatical case, number, gender, or part of speech, as well as some noun and verb roots. These morphological elements are essential for conveying syntactic and semantic relationships in Russian text.	0.82	0.78	1	0.64	0.86	0.848	0.929	0.78
model.layers.8.mlp	Russian	109317	The highlighted tokens are primarily morphemes, word stems, suffixes, and grammatical endings in Russian (and some in English), often marking case, number, tense, or forming nouns, adjectives, and verbs. There is a focus on functional elements that contribute to word formation, inflection, and syntactic structure, as well as on common connective words and phrase boundaries.	0.88	0.875	0.913	0.84	0.59	0.692	0.554	0.92
model.layers.9.mlp	Russian	30251	The highlighted tokens are predominantly Russian suffixes and verb endings that indicate tense, aspect, person, number, or case, as well as nominal and adjectival endings. These morphological markers are essential for grammatical structure and meaning in Russian sentences.	0.76	0.684	1	0.52	0.82	0.795	0.921	0.7
model.layers.9.mlp	Russian	59767	The highlighted tokens are predominantly Russian word stems, prefixes, and suffixes that form the core of verbs, nouns, and adjectives, often marking grammatical or semantic roles such as actions, objects, or qualities. These morphemes are crucial for word formation and meaning in Russian, frequently appearing at the beginning or within words to indicate tense, aspect, subject, or object, and are central to the structure and interpretation of Russian sentences.	0.79	0.734	1	0.58	0.89	0.882	0.953	0.82
model.layers.10.mlp	Russian	56594	The highlighted tokens are predominantly proper nouns, technical terms, and named entities—such as personal names, place names, and institutional or scientific terms—often in multiple languages. These tokens frequently appear in contexts involving formal identification, attribution, or description of people, locations, organizations, or specialized concepts.	0.51	0.505	0.51	0.5	0.58	0.4	0.7	0.28
model.layers.11.mlp	Russian	8452	The highlighted tokens are predominantly Russian morphemes, words, or short phrases that serve as key semantic or syntactic units within sentences. They include verb roots, noun endings, pronouns, conjunctions, and common collocations, often marking the core meaning, grammatical structure, or transitions in the text. The activations focus on elements that define actions, states, relationships, or important contextual information, reflecting the building blocks of Russian sentence construction and meaning.	0.78	0.718	1	0.56	0.81	0.765	1	0.62
model.layers.11.mlp	Russian	74675	The highlighted tokens are primarily Russian morphemes, roots, and affixes that form the core semantic or grammatical structure of words, often marking verbs, nouns, or adjectives, and are crucial for conveying meaning, tense, aspect, or function within the sentence.	0.79	0.734	1	0.58	0.84	0.814	0.972	0.7