Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.12.mlp	Russian	31110	The highlighted tokens are predominantly Russian morphemes, roots, or affixes that form the core semantic or grammatical structure of words, often marking verbs, adjectives, or nouns. These segments frequently appear at the beginning or end of words and are essential for conveying meaning, tense, aspect, or other grammatical features in Russian. The activations focus on these meaningful subword units that are central to word formation and comprehension in the language.	0.76	0.684	1	0.52	0.87	0.854	0.974	0.76
model.layers.12.mlp	Russian	48228	The highlighted tokens are primarily Russian morphemes, function words, and endings that mark grammatical relationships, verb conjugations, noun/adjective declensions, and common syntactic connectors. These tokens are crucial for sentence structure, tense, case, agreement, and linking clauses, reflecting the morphological richness and syntactic dependencies of Russian language.	0.67	0.507	1	0.34	0.79	0.747	0.939	0.62
model.layers.13.mlp	Russian	29244	The highlighted tokens are primarily morphemes, roots, prefixes, and suffixes within Russian words, often marking word formation, inflection, or derivation. These segments frequently correspond to meaningful subword units that contribute to the grammatical or semantic structure of the word, such as case endings, diminutives, verb aspects, or noun/adjective formation.	0.75	0.675	0.963	0.52	0.88	0.875	0.913	0.84
model.layers.13.mlp	Russian	40052	The highlighted tokens are predominantly Russian (and some other Slavic) word endings, suffixes, and inflections that indicate grammatical case, number, gender, verb tense, or derivational morphology. These endings are crucial for understanding syntactic roles and semantic relationships in the text, as well as for identifying noun, adjective, and verb forms.	0.7	0.615	0.857	0.48	0.75	0.713	0.838	0.62
model.layers.13.mlp	Russian	50197	The prefix \"от\" and its variants are frequently activated, indicating a focus on Russian morphological patterns where \"от\" serves as a prefix in verbs and nouns to convey meanings related to separation, removal, or origin. Other common activations include prefixes and prepositions, highlighting the importance of word formation and grammatical function in Russian text.	0.6	0.355	0.917	0.22	0.6	0.333	1	0.2
model.layers.13.mlp	Russian	68491	The highlighted tokens are predominantly Russian morphemes, roots, and suffixes that form or modify nouns and adjectives, often marking case, number, gender, or forming compound words, especially in contexts describing locations, organizations, or abstract concepts.	0.7	0.583	0.955	0.42	0.71	0.651	0.818	0.54
model.layers.13.mlp	Russian	91930	The highlighted tokens are predominantly Russian (with some Bulgarian) morphemes, word endings, and function words that are essential for grammatical structure, word formation, and meaning in context. These include case endings, verb conjugations, participles, prepositions, conjunctions, and suffixes that indicate tense, number, gender, or aspect. The activations focus on elements that define syntactic roles, connect phrases, or modify the meaning of nouns and verbs, reflecting the importance of morphology and function words in Slavic languages for sentence construction and semantic clarity.	0.76	0.692	0.964	0.54	0.7	0.634	0.812	0.52
model.layers.13.mlp	Russian	97322	The highlighted tokens are predominantly Russian morphological suffixes, prepositions, conjunctions, and noun/adjective endings that indicate grammatical relationships, case, number, gender, and aspect. These elements are essential for syntactic structure and meaning in Russian, often marking possession, agency, time, and other core grammatical functions.	0.61	0.4	0.867	0.26	0.62	0.5	0.731	0.38
model.layers.13.mlp	Russian	103291	The highlighted tokens are predominantly suffixes and endings in Slavic and some other languages, marking grammatical features such as case, number, gender, tense, aspect, or participle forms. These endings are crucial for conveying syntactic and semantic relationships within sentences.	0.85	0.848	0.857	0.84	0.68	0.719	0.641	0.82
model.layers.13.mlp	Russian	115738	The important tokens are predominantly Russian morphemes, roots, and affixes that form the core semantic or grammatical structure of words, often marking noun, verb, or adjective stems, as well as common derivational and inflectional endings. These tokens frequently appear at the beginning or within the main body of words, highlighting the morphological building blocks essential for meaning and word formation in Russian text.	0.7	0.571	1	0.4	0.77	0.716	0.935	0.58
model.layers.13.mlp	Russian	120158	The highlighted tokens are predominantly Russian morphemes, especially suffixes, roots, and stems that form or modify nouns, adjectives, and verbs. These morphemes often indicate grammatical features such as case, number, gender, aspect, or degree, and are essential for word formation and meaning in Russian. The activations focus on these subword units that carry core semantic or grammatical information.	0.75	0.667	1	0.5	0.85	0.824	1	0.7