Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
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	0.67	1.00	0.50	0.85	0.82	1.00	0.70
model.layers.14.mlp	Russian	44860	The highlighted tokens are predominantly Russian morphemes, roots, and affixes that form the core semantic or grammatical structure of words. These include verb and noun roots, derivational and inflectional suffixes, and key word segments that contribute to meaning, tense, aspect, or case. The pattern reflects a focus on the internal morphological structure of Russian words, emphasizing the parts that carry essential lexical or grammatical information.	0.65	0.48	0.94	0.32	0.73	0.66	0.90	0.52
model.layers.14.mlp	Russian	71946	The highlighted tokens are predominantly Russian morphological suffixes and endings that indicate grammatical features such as case, number, gender, tense, and aspect, as well as function words and common inflections. These patterns reflect the importance of morphological structure in Russian, with frequent emphasis on verb conjugations, noun and adjective declensions, and particles that contribute to syntactic and semantic relationships within sentences.	0.68	0.53	1.00	0.36	0.68	0.57	0.88	0.42
model.layers.15.mlp	Russian	56964	The highlighted tokens are primarily Russian morphemes, roots, and affixes that form the semantic and grammatical core of words, often marking key concepts, actions, or relationships within sentences. These include noun, verb, and adjective roots, as well as frequent derivational and inflectional suffixes, reflecting the morphological structure and meaning-bearing elements of the language.	0.65	0.46	1.00	0.30	0.69	0.58	0.91	0.42
model.layers.15.mlp	Russian	65387	The highlighted tokens are predominantly Russian morphemes, word endings, and function words that are essential for grammatical structure, case, and meaning in complex sentences. These include suffixes, inflections, conjunctions, and prepositions, as well as key content words that signal relationships, actions, or abstract concepts. The pattern reflects a focus on the morphological and syntactic elements that construct meaning and coherence in Russian academic or formal text.	0.72	0.61	1.00	0.44	0.74	0.68	0.90	0.54
model.layers.15.mlp	Russian	119968	The highlighted tokens are predominantly Russian verb roots, stems, and affixes, especially those forming past participles, infinitives, or verbal nouns, often marking actions, processes, or results. Many are verb-forming or modifying morphemes, frequently appearing at the end or within verbs, and are central to conveying the core action or state in the sentence.	0.64	0.46	0.94	0.30	0.81	0.78	0.94	0.66