Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.10.mlp	Turkish	14995	The highlighted tokens are predominantly Turkish morphemes, roots, and suffixes that form the core of verbs, nouns, and adjectives, often marking tense, person, plurality, or derivation. These segments are crucial for word formation and meaning, reflecting the agglutinative structure of Turkish where words are built from a sequence of meaningful units.	0.67	0.507	1	0.34	0.71	0.603	0.957	0.44
model.layers.10.mlp	Turkish	23876	The highlighted tokens are predominantly suffixes, inflectional endings, or morphemes in Turkish, Hindi, and related languages, often marking grammatical features such as case, possession, plurality, tense, or forming nouns and adjectives. These segments frequently appear at the end of words and are crucial for word formation and meaning in agglutinative and inflectional languages.	0.88	0.867	0.975	0.78	0.78	0.792	0.75	0.84
model.layers.10.mlp	Turkish	30160	The highlighted tokens are predominantly parts of proper nouns, place names, and personal names, often in various languages and scripts. These tokens frequently appear as meaningful subword units within longer names or terms, especially in contexts involving transliteration, multilingual text, or named entities.	0.5	0.537	0.5	0.58	0.6	0.697	0.561	0.92
model.layers.10.mlp	Turkish	38256	The highlighted tokens are predominantly Turkish morphemes, roots, and suffixes that form the core of nouns, verbs, and adjectives, often marking derivational or inflectional boundaries. These segments frequently appear at the start or end of words, indicating their importance in Turkish word formation and grammatical structure.	0.95	0.947	1	0.9	0.97	0.97	0.98	0.96
model.layers.10.mlp	Turkish	65847	The highlighted tokens frequently correspond to named entities, especially place names, personal names, and cultural terms, often in Turkish or related to Turkey, as well as key nouns and morphemes in multiple languages. These tokens often appear in contexts describing locations, people, or culturally significant items, and sometimes include inflectional or derivational suffixes.	0.74	0.768	0.694	0.86	0.58	0.65	0.557	0.78
model.layers.10.mlp	Turkish	81107	The highlighted tokens are predominantly Turkish suffixes, conjunctions, and common noun or verb endings that serve grammatical functions such as possession, plurality, tense, case, and coordination. These morphemes are essential for sentence structure, meaning, and cohesion in Turkish, often attaching to root words to modify or clarify their roles within the sentence.	0.61	0.361	1	0.22	0.62	0.406	0.929	0.26
model.layers.10.mlp	Turkish	97688	The highlighted tokens are verb endings or suffixes in Romance languages, often marking tense, person, or number, and are typically found at the end of verbs in various conjugated forms.	0.55	0.366	0.619	0.26	0.69	0.627	0.788	0.52
model.layers.11.mlp	Turkish	16555	The highlighted tokens are predominantly Turkish morphemes, roots, and suffixes that form the core of words, especially those indicating tense, plurality, possession, or case. Many activations occur at the boundaries of word stems and suffixes, or at the start of compound or derived words, reflecting the agglutinative structure of Turkish. There is also a focus on function words and key content words that carry the main semantic load in a sentence.	0.94	0.936	1	0.88	0.93	0.929	0.939	0.92
model.layers.11.mlp	Turkish	19579	The highlighted tokens are predominantly Turkish suffixes or stems, often marking noun or adjective forms, verb conjugations, or pluralization. These morphemes typically appear at the end of words and are essential for grammatical structure and meaning in Turkish.	0.66	0.485	1	0.32	0.68	0.556	0.909	0.4
model.layers.11.mlp	Turkish	37895	The highlighted tokens are primarily suffixes, inflections, or short function words in Turkish and German, often marking grammatical relationships such as possession, plurality, tense, or case, as well as prepositions and names in German. These tokens are crucial for the syntactic and semantic structure of sentences in agglutinative and inflectional languages.	0.6	0.429	0.75	0.3	0.54	0.574	0.534	0.62
model.layers.11.mlp	Turkish	38442	The highlighted tokens are predominantly Turkish suffixes and inflectional endings that modify word meaning, indicate possession, plurality, case, tense, or person, as well as some noun and verb roots. These morphological markers are essential for grammatical structure and meaning in Turkish sentences.	0.63	0.413	1	0.26	0.68	0.543	0.95	0.38
model.layers.11.mlp	Turkish	55507	The highlighted tokens are predominantly Turkish morphemes, noun and verb roots, and suffixes that form key parts of words, especially those marking tense, plurality, possession, or case. Many are used in constructing complex noun phrases, verb conjugations, or expressing relationships such as agency, time, and location. The activations focus on linguistically significant segments that contribute to the grammatical structure and meaning of sentences.	0.74	0.649	1	0.48	0.78	0.732	0.938	0.6
model.layers.11.mlp	Turkish	56369	The highlighted tokens are Turkish verb and noun roots, suffixes, and inflectional endings, often marking tense, person, plurality, or case, as well as forming compound words and participles. These morphemes are essential for word formation and grammatical structure in Turkish.	0.69	0.551	1	0.38	0.69	0.563	0.952	0.4