Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.0.mlp	Turkish	10506	The highlighted tokens are predominantly Turkish morphemes, suffixes, and root words that are essential for grammatical structure and meaning, including noun and verb endings, possessive and case markers, and common stems. There is a focus on morphological components that form or modify words, especially those indicating tense, possession, plurality, or forming adjectives and nouns.	0.57	0.30	0.82	0.18	0.57	0.48	0.61	0.40
model.layers.0.mlp	Turkish	64723	Suffixes such as \"ı\", \"lar\", \"ler\", \"lık\", \"lık\", \"ım\", \"in\", \"isi\", \"lığı\", \"ları\", \"leri\", \"ını\", \"udur\", \"ildi\", \"ım\", \"isi\", \"arı\", \"alı\", \"isi\", \"lık\",	0.57	0.30	0.82	0.18	0.60	0.33	1.00	0.20
model.layers.1.mlp	Turkish	84692	The highlighted tokens are predominantly prefixes, roots, or short morphemes within words, often marking the beginning or core of nouns, verbs, or adjectives across multiple languages. These segments frequently correspond to meaningful subword units, such as roots, affixes, or syllables, and are often found in proper names, technical terms, or compound words. The pattern reflects a focus on linguistically significant subword structures that contribute to word formation and meaning.	0.54	0.65	0.53	0.84	0.55	0.69	0.53	0.98
model.layers.2.mlp	Turkish	34696	The highlighted tokens are predominantly Turkish suffixes and inflectional endings, such as plural, possessive, case, tense, and participle markers, as well as common noun and verb roots. These morphological elements are essential for word formation and grammatical structure in Turkish.	0.75	0.67	1.00	0.50	0.78	0.73	0.94	0.60
model.layers.3.mlp	Turkish	4996	The highlighted tokens are predominantly Turkish morphemes, suffixes, and proper names, as well as some foreign names and borrowed words, often marking inflection, derivation, or named entities within multilingual or code-switched text.	0.77	0.71	0.97	0.56	0.72	0.73	0.70	0.76
model.layers.5.mlp	Turkish	45933	The highlighted tokens are Turkish suffixes, inflections, and common morphemes that modify word meaning, tense, possession, or case, as well as proper noun markers and date suffixes. These elements are essential for grammatical structure and semantic roles in Turkish sentences.	0.79	0.73	1.00	0.58	0.83	0.80	1.00	0.66
model.layers.6.mlp	Turkish	76077	The highlighted tokens are predominantly Turkish morphemes, suffixes, and proper names, as well as grammatical particles and inflections. There is a focus on word endings, possessive or case suffixes, and named entities, reflecting the agglutinative structure of Turkish and the importance of morphological boundaries and named entity recognition in the language.	0.90	0.89	1.00	0.80	0.89	0.88	0.93	0.84
model.layers.7.mlp	Turkish	95221	The highlighted tokens are predominantly Turkish morphemes, suffixes, and root words, as well as proper nouns and function words. They often mark grammatical features such as possession, plurality, tense, or case, and include common noun and verb roots, as well as names and frequently used connectors. These tokens are important for understanding word formation, inflection, and syntactic structure in Turkish text.	0.89	0.88	1.00	0.78	0.86	0.85	0.89	0.82
model.layers.8.mlp	Turkish	7945	The highlighted tokens are predominantly Turkish morphemes, suffixes, and root words that form the core of meaning in sentences, including noun and verb roots, inflectional and derivational suffixes, and common function words. These tokens often mark grammatical relationships, tense, possession, plurality, and case, and are essential for constructing and understanding the syntactic and semantic structure of Turkish text.	0.89	0.88	1.00	0.78	0.90	0.89	0.98	0.82
model.layers.8.mlp	Turkish	113784	The highlighted tokens are predominantly morphemes, syllables, or name fragments within proper nouns, place names, and personal names, often from Turkish but also from other languages. These tokens frequently appear in transliterations, compound words, or as parts of culturally specific terms, indicating a focus on subword units that contribute to the identification and construction of named entities and culturally significant vocabulary.	0.87	0.87	0.85	0.90	0.66	0.74	0.60	0.96
model.layers.9.mlp	Turkish	3919	The highlighted tokens are predominantly Turkish morphemes, roots, and suffixes that form the core of words, especially those indicating tense, plurality, possession, or case. These tokens often appear at the beginning, middle, or end of words and are essential for constructing meaning in Turkish sentences, reflecting the language's agglutinative structure.	0.93	0.93	1.00	0.86	0.94	0.94	1.00	0.88