| Layer | Lang | Feature ID | Interpretation | Detection | | | | Fuzzing | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Accuracy | F1 score | Precision | Recall | Accuracy | F1 score | Precision | Recall |
| model.layers.14.mlp | Turkish | 87172 | The highlighted tokens are predominantly Turkish morphemes, roots, and affixes, often marking noun and verb stems, derivational or inflectional endings, and common word fragments. These tokens frequently appear at the beginning or end of words, reflecting the agglutinative structure of Turkish, where meaning is built up through the addition of suffixes and prefixes to roots. | 0.9 | 0.889 | 1 | 0.8 | 0.92 | 0.917 | 0.957 | 0.88 |
| model.layers.14.mlp | Turkish | 91273 | The highlighted tokens are predominantly Turkish morphemes, roots, and suffixes that form or modify nouns, adjectives, and verbs, often marking possession, plurality, tense, or case. These segments frequently appear at the end or within words, reflecting Turkish's agglutinative structure where meaning is built up through the addition of affixes. | 0.91 | 0.901 | 1 | 0.82 | 0.93 | 0.926 | 0.978 | 0.88 |
| model.layers.14.mlp | Turkish | 106497 | The highlighted tokens predominantly correspond to Turkish noun and adjective roots, suffixes, and compound structures, often marking key semantic units such as institutions, official terms, services, locations, and formal processes. There is a strong emphasis on morphological components (roots and suffixes) that form the core meaning of words, especially in administrative, legal, travel, and service-related contexts. These patterns reflect the agglutinative nature of Turkish, where meaning is built up through the combination of roots and multiple suffixes. | 0.88 | 0.867 | 0.975 | 0.78 | 0.92 | 0.917 | 0.957 | 0.88 |
| model.layers.14.mlp | Turkish | 108034 | The highlighted tokens are predominantly Turkish morphemes, suffixes, and function words that play a key role in grammatical structure, such as denoting possession, plurality, case, tense, or forming compound words and phrases. These tokens often appear at the end of words or as connectors, reflecting the agglutinative nature of Turkish, where meaning and grammatical relationships are built up through the addition of suffixes and short function words. | 0.65 | 0.462 | 1 | 0.3 | 0.65 | 0.493 | 0.895 | 0.34 |
| model.layers.15.mlp | Turkish | 7049 | The highlighted tokens are predominantly Turkish suffixes, inflections, and function words that modify meaning, indicate possession, plurality, tense, or case, as well as common noun and verb roots. These elements are essential for grammatical structure and semantic relationships in Turkish sentences. | 0.89 | 0.876 | 1 | 0.78 | 0.9 | 0.891 | 0.976 | 0.82 |
| model.layers.15.mlp | Turkish | 12913 | The highlighted tokens are predominantly Turkish morphemes, suffixes, and function words, as well as common noun and verb roots. These tokens often mark grammatical relationships, inflections, or frequently used connectors in Turkish, reflecting the language's agglutinative structure and the importance of suffixes for meaning and syntax. | 0.83 | 0.8 | 0.971 | 0.68 | 0.85 | 0.828 | 0.973 | 0.72 |
| model.layers.15.mlp | Turkish | 29918 | The highlighted tokens are predominantly Turkish morphemes, suffixes, and word stems that play key grammatical or semantic roles, such as indicating tense, plurality, possession, or forming compound words. These tokens often appear at the ends of words or as part of agglutinative constructions, reflecting the morphological structure of Turkish, where meaning and grammatical function are built up through the addition of suffixes and inflections. | 0.75 | 0.667 | 1 | 0.5 | 0.73 | 0.649 | 0.926 | 0.5 |
| model.layers.15.mlp | Turkish | 36457 | The highlighted tokens are predominantly Turkish morphemes, roots, and suffixes that form the core of word meanings or grammatical functions, such as verb stems, noun roots, and common derivational or inflectional endings. These elements are crucial for constructing and understanding words, indicating tense, possession, plurality, or forming new words, and are often the most semantically or syntactically informative parts of the text. | 0.67 | 0.522 | 0.947 | 0.36 | 0.69 | 0.617 | 0.806 | 0.5 |
| model.layers.15.mlp | Turkish | 57867 | The highlighted tokens are predominantly Turkish morphemes, roots, and affixes, often marking noun and verb stems, derivational or inflectional endings, and compound word boundaries. These tokens frequently appear at the start or end of words, indicating their importance in Turkish word formation and morphological structure. | 0.87 | 0.851 | 1 | 0.74 | 0.87 | 0.857 | 0.951 | 0.78 |
| model.layers.15.mlp | Turkish | 67563 | The highlighted tokens are predominantly Turkish morphemes, suffixes, and roots that form or modify nouns, verbs, and adjectives, often marking grammatical features such as possession, plurality, tense, or case. Many activations correspond to common Turkish word endings or inflectional affixes, as well as roots of frequently used words, indicating a focus on morphological structure and word formation in Turkish text. | 0.84 | 0.81 | 1 | 0.68 | 0.85 | 0.824 | 1 | 0.7 |
| model.layers.15.mlp | Turkish | 86432 | The highlighted tokens are predominantly suffixes, inflections, or short morphemes in Turkish and related languages, often marking grammatical features such as possession, plurality, tense, case, or forming nouns and adjectives. These tokens frequently appear at the end of words and are essential for conveying syntactic and semantic relationships in agglutinative languages. | 0.89 | 0.876 | 1 | 0.78 | 0.8 | 0.808 | 0.778 | 0.84 |