| Layer | Lang | Feature ID | Interpretation | Detection | | | | Fuzzing | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Accuracy | F1 score | Precision | Recall | Accuracy | F1 score | Precision | Recall |
| model.layers.13.mlp | Turkish | 89356 | The highlighted tokens are predominantly Turkish morphemes, roots, and suffixes that form or modify nouns, verbs, and adjectives, often marking grammatical features such as possession, plurality, tense, or case. These segments frequently appear at word boundaries or as inflectional/derivational endings, reflecting the agglutinative structure of Turkish. | 0.82 | 0.78 | 1 | 0.64 | 0.85 | 0.824 | 1 | 0.7 |
| model.layers.13.mlp | Turkish | 94117 | The highlighted tokens are primarily Turkish morphemes, compound nouns, and suffixes that form key semantic units such as professions, institutions, services, actions, and descriptors. These tokens often appear at the end of words or as part of compound structures, marking grammatical roles (e.g., possession, plurality, case), or denoting specific domains (e.g., programs, organizations, locations, official documents). The pattern reflects the agglutinative nature of Turkish, where meaning is built up through the addition of suffixes and compound elements. | 0.84 | 0.814 | 0.972 | 0.7 | 0.93 | 0.925 | 1 | 0.86 |
| model.layers.13.mlp | Turkish | 102015 | The highlighted tokens are predominantly Turkish suffixes and inflections that indicate tense, person, plurality, possession, negation, and modality, as well as some common noun and verb roots. These morphological markers are essential for conveying grammatical relationships and meaning in Turkish sentences. | 0.64 | 0.438 | 1 | 0.28 | 0.72 | 0.622 | 0.958 | 0.46 |
| model.layers.13.mlp | Turkish | 119094 | The highlighted tokens are predominantly Turkish suffixes and inflectional endings that attach to word stems to indicate grammatical relationships such as possession, plurality, tense, case, and person. These suffixes are essential for conveying meaning and structure in Turkish sentences, and their activation reflects their importance in determining the function and relationship of words within the text. | 0.66 | 0.485 | 1 | 0.32 | 0.64 | 0.455 | 0.938 | 0.3 |
| model.layers.13.mlp | Turkish | 119659 | The highlighted tokens are predominantly Turkish morphemes, suffixes, and function words that play a key role in sentence structure, verb conjugation, possession, and case marking, as well as common connectors and pronouns. These elements are essential for expressing grammatical relationships and meaning in Turkish sentences. | 0.83 | 0.795 | 1 | 0.66 | 0.83 | 0.795 | 1 | 0.66 |
| model.layers.14.mlp | Turkish | 22302 | The highlighted tokens are predominantly Turkish morphemes, suffixes, and stems that form the core of word meanings, verb conjugations, and noun/adjective derivations. These include tense, person, plurality, possession, and case markers, as well as common roots and affixes that are essential for grammatical structure and semantic content in Turkish sentences. The activations focus on the morphological building blocks that determine the function and meaning of words within the sentence. | 0.74 | 0.649 | 1 | 0.48 | 0.76 | 0.684 | 1 | 0.52 |
| model.layers.14.mlp | Turkish | 30275 | The highlighted tokens are predominantly Turkish suffixes, inflections, and function words that modify meaning, indicate grammatical relationships, or form noun and verb phrases. These include case endings, possessives, plural markers, verb conjugations, and common connectors, reflecting the agglutinative structure of Turkish and the importance of morphological units in sentence construction. | 0.76 | 0.684 | 1 | 0.52 | 0.78 | 0.718 | 1 | 0.56 |
| model.layers.14.mlp | Turkish | 33836 | The highlighted tokens are predominantly Turkish suffixes, case endings, and common word stems, as well as function words and frequently used noun or verb forms. These elements are essential for grammatical structure, word formation, and meaning in Turkish sentences, often marking tense, possession, plurality, or case, and sometimes indicating proper nouns or key content words. | 0.91 | 0.903 | 0.977 | 0.84 | 0.89 | 0.887 | 0.915 | 0.86 |
| model.layers.14.mlp | Turkish | 43402 | The highlighted tokens are predominantly suffixes, inflections, and morphemes in Turkish and related languages, as well as proper nouns, dates, and grammatical particles. These tokens often mark case, possession, plurality, tense, or are part of compound words and names, reflecting the agglutinative structure of the language and the importance of morphological boundaries and named entities in text processing. | 0.94 | 0.941 | 0.923 | 0.96 | 0.91 | 0.913 | 0.887 | 0.94 |
| model.layers.14.mlp | Turkish | 51854 | The highlighted tokens are predominantly Turkish suffixes, inflections, and root morphemes, especially those involving \"gö\", \"kü\", \"ü\", \"ö\", and other vowel-rich or agglutinative elements, often marking grammatical features like possession, plurality, tense, or forming nouns and adjectives. These tokens are central to Turkish word formation and meaning. | 0.93 | 0.928 | 0.957 | 0.9 | 0.9 | 0.896 | 0.935 | 0.86 |
| model.layers.14.mlp | Turkish | 65657 | The highlighted tokens are predominantly Turkish morphemes, suffixes, and function words that play key roles in sentence structure, verb conjugation, possession, and meaning. These include verb endings, possessive and case suffixes, conjunctions, and common auxiliary words, reflecting the agglutinative nature of Turkish and the importance of these elements in constructing grammatical and semantically complete sentences. | 0.84 | 0.81 | 1 | 0.68 | 0.82 | 0.791 | 0.944 | 0.68 |