Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.0.mlp	Korean	87423	The highlighted tokens are primarily Korean morphemes, words, or short phrases that serve as key semantic units, often marking nouns, noun phrases, or grammatical particles that define objects, locations, roles, or actions within sentences. There is a strong emphasis on tokens that denote specific entities, locations, or concepts, as well as grammatical markers that structure meaning, such as subject or object particles, and elements that form compound nouns or descriptive phrases. The pattern reflects the importance of these units in conveying core information and relationships in Korean text.	0.54	0.148	1	0.08	0.58	0.323	0.833	0.2
model.layers.4.mlp	Korean	99278	The highlighted tokens are predominantly Korean morphemes, particles, or short words, often at the beginning of words or phrases, as well as parenthesis and some punctuation. These tokens frequently serve grammatical or structural roles in Korean text, such as marking possession, subject, object, or other syntactic functions, and are often found at word boundaries or as part of compound words.	0.88	0.872	0.932	0.82	0.79	0.792	0.784	0.8
model.layers.6.mlp	Korean	109962	The highlighted tokens are primarily high-activation morphemes, word stems, or grammatical particles in Korean and Turkish, as well as some corrupted or unknown characters, often marking key syntactic or semantic roles within sentences. These tokens frequently appear at word boundaries, as suffixes, or as part of compound words, and are crucial for sentence structure or meaning.	0.78	0.75	0.868	0.66	0.55	0.628	0.535	0.76
model.layers.7.mlp	Korean	115377	The highlighted tokens are primarily morphemes, syllables, or words related to the Korean language, Korean names, or references to Korea in various languages, as well as some corrupted or non-standard characters. These tokens often appear in multilingual contexts, especially when discussing Korean culture, language, or proper nouns.	0.95	0.947	1	0.9	0.926	0.921	0.976	0.872
model.layers.8.mlp	Korean	94922	The highlighted tokens are primarily Korean morphemes, syllables, or words, often marking grammatical functions, noun or verb stems, and endings. There is frequent activation on common content and function morphemes, as well as on foreign or borrowed terms (such as Hangeul-related names). Additionally, corrupted or unreadable characters are sometimes marked, likely due to encoding issues. The pattern reflects a focus on meaningful linguistic units in Korean text, especially those that contribute to sentence structure or key content.	0.87	0.851	1	0.74	0.86	0.841	0.974	0.74
model.layers.9.mlp	Korean	31611	The highlighted tokens are primarily Korean morphemes and words that serve as grammatical markers, verb endings, or key content words (such as nouns and verbs) essential for sentence structure and meaning. There is a strong emphasis on verb and adjective endings that indicate tense, aspect, or politeness, as well as on nouns and particles that define relationships and actions within the sentence. Additionally, some tokens are corrupted or incomplete, likely due to encoding issues, but the overall pattern centers on the linguistic building blocks that are critical for understanding and generating Korean sentences.	0.85	0.824	1	0.7	0.87	0.851	1	0.74
model.layers.9.mlp	Korean	53931	The highlighted tokens are primarily Korean nouns, verbs, and grammatical endings that denote key entities, actions, or states within sentences, often marking subjects, objects, or important predicates. Many tokens are part of compound words or phrases that convey core semantic content, such as people, places, actions, or results. There is a strong emphasis on tokens that contribute to the main informational structure of the sentence, including those that indicate possession, causality, or completion.	0.75	0.667	1	0.5	0.78	0.756	0.85	0.68