Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.15.mlp	Korean	17864	The highlighted tokens are primarily Korean morphemes, words, or short phrases that serve as grammatical particles, noun or verb endings, or key content words within sentences. These tokens often mark syntactic boundaries, indicate relationships between sentence elements, or represent important semantic units such as subjects, objects, or actions. The activations tend to focus on tokens that are functionally or semantically significant for understanding sentence structure and meaning in Korean text.	0.64	0.438	1	0.28	0.69	0.551	1	0.38
model.layers.15.mlp	Korean	41829	The highlighted tokens are primarily proper nouns, place names, and specialized terms in Korean text, often denoting people, locations, or unique concepts. Many are multi-syllabic and appear in contexts introducing or describing entities, sometimes accompanied by quotation marks or parenthetical explanations.	0.63	0.448	0.882	0.3	0.61	0.4	0.867	0.26
model.layers.15.mlp	Korean	57880	The highlighted tokens are primarily Korean grammatical particles, verb endings, and noun suffixes, as well as some content words and punctuation. These tokens often serve as key syntactic or morphological markers in Korean, indicating case, tense, subject, object, or other grammatical relationships, and are essential for sentence structure and meaning. Some non-Korean examples show similar emphasis on function words, endings, or connectors in other languages.	0.63	0.507	0.76	0.38	0.57	0.65	0.548	0.8
model.layers.15.mlp	Korean	60936	The highlighted tokens are primarily proper nouns, place names, and key content words in Korean and multilingual text, often marking named entities, important objects, or significant actions within sentences. These tokens frequently appear at the start or end of phrases, and include both native and foreign names, as well as terms central to the sentence's meaning.	0.71	0.734	0.678	0.8	0.53	0.447	0.543	0.38
model.layers.15.mlp	Korean	114116	The highlighted tokens are predominantly Korean nouns, noun phrases, and compound words, often denoting objects, concepts, activities, or roles relevant to travel, information, media, and communication. Many are domain-specific terms or collocations, and several are used in formal, instructional, or informational contexts. There is a strong emphasis on key content words that convey the main subject or function within each sentence.	0.78	0.744	0.889	0.64	0.87	0.851	1	0.74
model.layers.15.mlp	Korean	131027	The highlighted tokens are primarily Korean morphemes, words, and grammatical endings, often marking key semantic or syntactic units such as nouns, verbs, particles, and endings that define sentence structure or meaning. There is also frequent highlighting of proper nouns, numbers, and technical terms, as well as some corrupted or non-standard characters, indicating a focus on linguistically significant elements and possibly on tokens that are informative for parsing or translation tasks.	0.74	0.658	0.962	0.5	0.74	0.667	0.929	0.52