Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.14.mlp	Thai	62719	The highlighted tokens are primarily Thai morphemes, syllables, or short words that function as grammatical markers, noun or verb roots, or affixes. They often appear at the boundaries of phrases or as key components in compound words, and are frequently associated with high-importance content such as actions, conditions, or objects within sentences. The pattern reflects a focus on semantically or syntactically significant units in Thai text.	0.7	0.571	1	0.4	0.69	0.551	1	0.38
model.layers.14.mlp	Thai	73972	The pattern highlights the importance of single or paired consonant-vowel tokens, especially those involving \"ต\" (Thai), \"ท\" (Thai), and \"त\" (Hindi), which frequently appear at morpheme or word boundaries, often as part of grammatical constructions, inflections, or compound words in Thai and Hindi text. These tokens are crucial for forming or modifying meaning within words and phrases.	0.78	0.756	0.85	0.68	0.81	0.804	0.83	0.78
model.layers.14.mlp	Thai	101719	The highlighted tokens are primarily function words, particles, and morphemes that serve grammatical roles in Thai, such as marking tense, aspect, conjunctions, pronouns, and case. There is a strong emphasis on connectors, sentence structure markers, and suffixes that modify meaning or indicate relationships between clauses and entities. These tokens are essential for the syntactic and semantic cohesion of Thai sentences.	0.85	0.824	1	0.7	0.82	0.791	0.944	0.68
model.layers.15.mlp	Thai	42804	The highlighted tokens are primarily morphemes, syllables, or short word fragments in Thai and English, often marking key semantic units such as nouns, verbs, locations, or grammatical particles. These segments frequently appear at the boundaries of words or phrases, and are important for identifying meaning, structure, or named entities within multilingual or code-mixed text.	0.7	0.643	0.794	0.54	0.54	0.603	0.53	0.7
model.layers.15.mlp	Thai	50967	The highlighted tokens are primarily Thai morphemes, syllables, or short words that often serve as key semantic units within phrases, proper nouns, or grammatical structures. These tokens frequently appear at the boundaries of compound words, names, or important content words, and are often associated with high informational value or serve as connectors in the sentence structure.	0.9	0.889	1	0.8	0.89	0.879	0.976	0.8
model.layers.15.mlp	Thai	58977	The highlighted tokens are primarily morphemes, syllables, or short word fragments in various languages, often marking grammatical functions, word formation, or key semantic units. These include suffixes, prefixes, and root components that are essential for constructing meaning, indicating tense, plurality, comparison, or other grammatical relationships. The activations focus on these subword units as they are crucial for understanding and generating morphologically rich or agglutinative languages, as well as for tokenization in multilingual contexts.	0.5	0.667	0.5	1	0.52	0.676	0.51	1
model.layers.15.mlp	Thai	66481	The highlighted tokens are primarily Thai morphemes, syllables, or short words that serve as key semantic or grammatical units within sentences. These tokens often mark the beginnings or important parts of compound words, function words, or content words, and are frequently found at the start of phrases, within compound constructions, or as part of inflectional or derivational morphology. The pattern reflects the segmentation of Thai text into meaningful subword units that are crucial for understanding sentence structure and meaning.	0.63	0.413	1	0.26	0.69	0.551	1	0.38
model.layers.15.mlp	Thai	81190	The highlighted tokens are primarily Thai syllables, morphemes, or short word fragments, often at the beginning or end of words, including prefixes, suffixes, and root components. These tokens frequently appear in proper nouns, compound words, and key content words, reflecting the agglutinative and syllabic structure of Thai, where meaning is built from combining such elements. The activations suggest importance for semantic, grammatical, or named-entity recognition within the language.	0.69	0.551	1	0.38	0.68	0.556	0.909	0.4