Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.0.mlp	Thai	38602	The highlighted tokens are primarily Thai suffixes, particles, and function words that serve grammatical roles such as marking location, possession, comparison, or forming questions and relative clauses. These tokens are essential for sentence structure and meaning in Thai, often appearing at the end of phrases or as connectors within sentences.	0.72	0.622	0.958	0.46	0.74	0.658	0.962	0.5
model.layers.0.mlp	Thai	48180	The highlighted tokens are predominantly function words, prefixes, or short morphemes that serve as grammatical connectors or building blocks in Thai sentences. These include pronouns, conjunctions, prepositions, classifiers, and common verb or noun prefixes/suffixes. Their importance lies in structuring sentences, indicating relationships, and forming compound or derived words, which are essential for the coherence and meaning of Thai text.	0.69	0.575	0.913	0.42	0.72	0.622	0.958	0.46
model.layers.0.mlp	Thai	53064	The highlighted tokens are primarily single or compound morphemes in Thai (and some in other languages), often marking key semantic units such as nouns, classifiers, or function words, especially those denoting groups, professions, or locations. These tokens frequently appear in compound words or as part of noun phrases, and are often central to the meaning or grammatical structure of the sentence.	0.69	0.627	0.788	0.52	0.59	0.65	0.567	0.76
model.layers.0.mlp	Thai	60676	The highlighted tokens are primarily function words, pronouns, or common morphemes in Hindi and Thai, such as demonstratives, possessives, or suffixes, which play a key role in sentence structure and meaning.	0.74	0.658	0.962	0.5	0.77	0.768	0.776	0.76
model.layers.1.mlp	Thai	9059	The highlighted tokens are morphemes, syllables, or short words in various languages (such as Turkish, Thai, and Bulgarian) that often serve as grammatical markers, roots, or affixes, and are frequently found within or at the boundaries of words, indicating their importance in word formation and meaning.	0.68	0.714	0.645	0.8	0.57	0.695	0.538	0.98
model.layers.1.mlp	Thai	73197	The highlighted tokens are primarily non-English characters or syllables from various languages, often representing morphemes, suffixes, or inflections that are important for grammatical structure or meaning within their respective languages. These tokens frequently appear in the middle or end of words, indicating their role in word formation or modification.	0.71	0.772	0.636	0.98	0.59	0.692	0.554	0.92
model.layers.2.mlp	Thai	45311	The highlighted tokens are primarily Thai morphemes, syllables, or word stems that form key parts of nouns, verbs, and adjectives, often marking important semantic content such as professions, objects, actions, or abstract concepts. These tokens frequently appear in compound words or as affixes, and are often associated with high-importance words in the sentence, especially those conveying the main subject, action, or descriptive quality.	0.7	0.595	0.917	0.44	0.75	0.675	0.963	0.52
model.layers.2.mlp	Thai	100192	The highlighted tokens are often function words, affixes, or short morphemes in various languages, including prepositions, conjunctions, pronouns, and grammatical particles, as well as some high-activation non-standard or corrupted characters. These tokens are typically important for sentence structure, meaning, or language identification.	0.52	0.676	0.51	1	0.47	0.629	0.484	0.9
model.layers.2.mlp	Thai	105874	The highlighted tokens are often morphemes, syllables, or short word fragments that serve as meaningful units in various languages, including Thai, Hindi, and others. These units frequently appear at the beginning, middle, or end of words and are important for word formation, inflection, or conveying grammatical and semantic information.	0.76	0.714	0.882	0.6	0.6	0.701	0.56	0.94
model.layers.2.mlp	Thai	110194	The highlighted tokens are predominantly short function words, pronouns, prepositions, conjunctions, and common morphemes in Bulgarian and Thai, often appearing at the start or end of words. These tokens are essential for grammatical structure, sentence cohesion, and meaning, frequently marking subject, object, possession, or verb forms. Their high activation suggests a focus on syntactic roles and morphological boundaries in multilingual contexts.	0.78	0.776	0.792	0.76	0.68	0.742	0.622	0.92
model.layers.3.mlp	Thai	71756	The highlighted segments are primarily Thai morphemes, syllables, or short words that serve as key grammatical or semantic units within phrases, often marking important nouns, verbs, or connectors that structure meaning in the sentence. These tokens frequently appear at the boundaries of compound words or phrases, and are often associated with core content or function words essential for understanding the main idea.	0.7	0.571	1	0.4	0.69	0.587	0.88	0.44
model.layers.3.mlp	Thai	130933	The highlighted tokens are primarily single or multi-character morphemes in Thai, Bulgarian, and related scripts, often marking key semantic units such as roots, affixes, or important syllables within words. These tokens frequently appear in positions of morphological or syntactic significance, such as forming the core meaning of a word, indicating grammatical relationships, or serving as part of compound or derived forms. The activations suggest a focus on linguistically meaningful subword units across multiple languages.	0.77	0.753	0.814	0.7	0.66	0.738	0.6	0.96