Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.4.mlp	Thai	33160	The highlighted tokens are often single characters, syllables, or short morphemes from multiple languages, especially Thai, as well as fragments of words in other languages. These tokens tend to be linguistically meaningful units such as prefixes, suffixes, or roots, and are frequently found at the beginning, middle, or end of words, often marking grammatical or semantic boundaries.	0.62	0.712	0.573	0.94	0.56	0.694	0.532	1
model.layers.5.mlp	Thai	118169	The highlighted tokens correspond to the word \"Thailand\" and its demonyms or derivatives, as well as related country names, written in various languages and scripts. The pattern is the consistent identification of the country \"Thailand\" or its linguistic equivalents across multilingual contexts.	0.54	0.148	1	0.08	0.54	0.148	1	0.08
model.layers.5.mlp	Thai	122535	The highlighted tokens are primarily Thai consonants, vowels, and syllables, often appearing at the beginning or within words, and are frequently found in grammatical markers, word stems, or affixes. These tokens tend to have high activation when they form meaningful morphemes, serve as connectors, or are part of common word constructions in Thai text. Occasionally, non-Thai tokens are highlighted when they appear as significant morphemes or name components in other languages.	0.85	0.828	0.973	0.72	0.91	0.903	0.977	0.84
model.layers.6.mlp	Thai	61889	The highlighted tokens correspond to the word \"Thailand\" and its variants across multiple languages and scripts, as well as frequent high-activation function words or morphemes in Thai and other Asian languages, often marking country names, locations, or grammatical particles.	0.78	0.718	1	0.56	0.71	0.592	1	0.42
model.layers.6.mlp	Thai	68498	The highlighted tokens are morphemes or syllables at word boundaries across multiple languages, often marking inflections, derivations, or meaningful subword units. These include suffixes, prefixes, or roots that contribute to the grammatical or semantic structure of words.	0.54	0.671	0.522	0.94	0.56	0.676	0.535	0.92
model.layers.6.mlp	Thai	77505	The highlighted tokens are predominantly function words, particles, and grammatical markers in Thai (and some in Hindi), such as those indicating possession, location, comparison, or subordination, as well as common pronouns and auxiliary verbs. These tokens play a crucial role in sentence structure, linking, and meaning, often marking relationships between clauses, objects, and actions. Their high activation suggests their importance in parsing and understanding the syntactic and semantic framework of the sentences.	0.77	0.742	0.846	0.66	0.78	0.771	0.804	0.74
model.layers.6.mlp	Thai	86299	The highlighted tokens are primarily Thai morphemes, syllables, or short words that function as key semantic units, including prefixes, classifiers, pronouns, and common noun or verb roots. These tokens often appear at the start or within compound words, and are frequently used in forming grammatical structures, proper nouns, or technical terms. The pattern reflects the agglutinative and compounding nature of Thai, where meaning is built from short, high-frequency morphemes.	0.87	0.851	1	0.74	0.88	0.867	0.975	0.78
model.layers.7.mlp	Thai	70835	The highlighted tokens are primarily morphemes, syllables, or short word fragments from multiple languages, especially Thai, Russian, and Bulgarian, often marking the start or core of content words such as nouns, verbs, or adjectives. These fragments frequently appear at the beginning or within important words, indicating their role in word formation and semantic content across diverse scripts and languages.	0.82	0.809	0.864	0.76	0.58	0.7	0.544	0.98
model.layers.7.mlp	Thai	128501	The highlighted tokens are often subword units or morphemes within names, words, or phrases across multiple languages, frequently marking meaningful components such as name parts, suffixes, or linguistic roots, and are not limited to a single language or script.	0.46	0.603	0.477	0.82	0.52	0.676	0.51	1
model.layers.8.mlp	Thai	48547	The highlighted tokens are predominantly Thai morphemes, syllables, or short words that frequently serve as functional units in compound words, proper nouns, or technical terms. Many activations correspond to prefixes, suffixes, or core morphemes that contribute to the grammatical structure or meaning of the phrase, such as indicating time, quantity, location, or agency. There is a recurring emphasis on tokens that form part of administrative, temporal, or descriptive expressions, as well as those that are central to the construction of compound nouns and formal terminology.	0.86	0.841	0.974	0.74	0.86	0.844	0.95	0.76
model.layers.8.mlp	Thai	55844	The highlighted tokens are short function words, prepositions, conjunctions, or grammatical particles in various languages (such as Bulgarian, Thai, Russian), often marking relationships between phrases or clauses, or serving as connectors within sentences.	0.62	0.703	0.577	0.9	0.64	0.727	0.585	0.96
model.layers.8.mlp	Thai	56232	The highlighted tokens are primarily morphemes, syllables, or short word segments that form the core of Thai words, often marking key semantic or grammatical units such as nouns, verbs, or important modifiers. These segments frequently appear at the beginning or within words, and are often associated with the main meaning or function of the word in the sentence.	0.64	0.455	0.938	0.3	0.67	0.522	0.947	0.36