Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
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.	0.70	0.57	1.00	0.40	0.69	0.59	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	0.75	0.81	0.70	0.66	0.74	0.60	0.96
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	0.71	0.57	0.94	0.56	0.69	0.53	1.00
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	0.15	1.00	0.08	0.54	0.15	1.00	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	0.83	0.97	0.72	0.91	0.90	0.98	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	0.72	1.00	0.56	0.71	0.59	1.00	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	0.67	0.52	0.94	0.56	0.68	0.54	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	0.74	0.85	0.66	0.78	0.77	0.80	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	0.85	1.00	0.74	0.88	0.87	0.98	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	0.81	0.86	0.76	0.58	0.70	0.54	0.98