Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.10.mlp	Thai	72591	The highlighted tokens are predominantly Thai morphemes, syllables, or short words that serve as key semantic units within compound words, proper nouns, or technical terms. These tokens often appear at the beginning or within multi-syllabic constructions, marking important grammatical, nominal, or conceptual boundaries in the text. Their selection reflects the agglutinative and compounding nature of Thai, where meaning is built from smaller, meaningful units.	0.78	0.718	1	0.56	0.81	0.765	1	0.62
model.layers.10.mlp	Thai	79756	The highlighted tokens are primarily function words, affixes, or morphemes in Thai and English that serve grammatical or semantic roles, such as marking causality, negation, comparison, or payment. In Thai, these include frequent use of particles, conjunctions, and affixes that indicate relationships, actions, or states, while in English, the focus is on key content words like \"payment.\" The pattern reflects the importance of structural and connective elements in conveying meaning and relationships within and between clauses.	0.89	0.882	0.953	0.82	0.68	0.754	0.613	0.98
model.layers.10.mlp	Thai	96158	The highlighted tokens are primarily parts of proper nouns, place names, or key content words in multiple languages, often marking named entities, important grammatical particles, or morphemes that contribute to meaning in context. The activations tend to focus on tokens that are semantically or syntactically significant, such as names, locations, and function words that structure information, especially in multilingual or code-switched text.	0.47	0.619	0.483	0.86	0.53	0.671	0.516	0.96
model.layers.10.mlp	Thai	99300	The highlighted tokens are primarily Thai morphemes or syllables that function as prefixes, roots, or grammatical markers, especially those forming verbs, nouns, or indicating actions, states, or relationships. There is a strong emphasis on tokens that begin or form part of compound words, particularly those related to actions, experiences, or states of being.	0.69	0.563	0.952	0.4	0.72	0.632	0.923	0.48
model.layers.11.mlp	Thai	6599	The highlighted tokens are single characters or short syllabic units from Hindi and Thai scripts, often appearing as morphemes or word stems within larger words, and are frequently found at the beginning or within content words, indicating their importance in word formation and meaning in these languages.	0.7	0.634	0.812	0.52	0.76	0.739	0.81	0.68
model.layers.11.mlp	Thai	10258	The highlighted tokens correspond to country names, demonyms, and related geographic or cultural terms, as well as conjunctions and prepositions that connect them, across multiple languages. These tokens often appear in historical or descriptive contexts involving nations, regions, or peoples, and are frequently found in multilingual parallel or comparative text.	0.51	0.347	0.52	0.26	0.53	0.277	0.6	0.18
model.layers.11.mlp	Thai	95738	The highlighted tokens are primarily function words, prefixes, and morphemes that serve as grammatical connectors or structural markers in Thai sentences, such as indicators of time, condition, possession, or subordination. These tokens often appear at the beginning or within compound words and phrases, and are essential for sentence cohesion and meaning, frequently marking relationships between clauses, actions, or participants.	0.84	0.81	1	0.68	0.88	0.867	0.975	0.78
model.layers.11.mlp	Thai	130422	The highlighted tokens are primarily functional morphemes and core verbs in Thai, such as those indicating necessity, ability, causation, or action (e.g., \"ต้อง\", \"จะ\", \"การ\", \"ตระ\", \"สามารถ\", \"กระ\", \"ยิ\", \"แส\", \"เข้า\"), as well as prefixes and stems that form key grammatical or semantic structures. These elements are central to expressing obligation, possibility, agency, and nominalization, and often appear at the start of compound words or as part of verb phrases.	0.63	0.413	1	0.26	0.67	0.507	1	0.34
model.layers.12.mlp	Thai	8775	The highlighted tokens are often found in polite, formal, or explanatory constructions in Japanese and Korean, including suggestions, indirect statements, and expressions of possibility or uncertainty. These tokens frequently appear in verb endings, auxiliary forms, and set phrases that convey nuance, deference, or hypothetical meaning.	0.58	0.382	0.722	0.26	0.77	0.729	0.886	0.62
model.layers.12.mlp	Thai	22929	The highlighted tokens are morphemes, syllables, or short word fragments in various languages, often marking grammatical or semantic units such as prefixes, suffixes, or roots, and are frequently found in compound words or as part of inflectional or derivational morphology.	0.55	0.646	0.532	0.82	0.54	0.676	0.522	0.96
model.layers.12.mlp	Thai	108692	The highlighted tokens are primarily function words, affixes, and key morphemes in Thai that serve to mark grammatical relationships, connect clauses, or indicate important semantic roles within sentences. There is a frequent emphasis on words and morphemes that denote time, agency, possession, and subordination, as well as those that introduce or link descriptive or explanatory clauses.	0.72	0.611	1	0.44	0.71	0.633	0.862	0.5