Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.9.mlp	Thai	110801	The highlighted tokens are primarily Thai morphemes, syllables, or words that function as key grammatical or semantic units, such as conjunctions, classifiers, pronouns, and time or quantity expressions. There is a recurring emphasis on tokens that form part of compound words, time expressions, or serve as connectors (e.g., \"ขณะท\" for \"while,\" \"ประมา\" for \" approximately,\" \"หร\" for \"or\"), as well as on morphemes that contribute to the structure and meaning of phrases, especially in contexts involving time, comparison, or enumeration.	0.91	0.90	1.00	0.82	0.92	0.91	1.00	0.84
model.layers.10.mlp	Thai	27357	The highlighted tokens are frequently grammatical particles, affixes, or short function words in various languages, often marking aspects such as negation, comparison, possession, or subordination. These tokens are typically high-frequency morphemes or syllables that play a key role in sentence structure and meaning, especially in agglutinative or analytic languages.	0.50	0.66	0.50	0.96	0.51	0.66	0.51	0.96
model.layers.10.mlp	Thai	34806	The highlighted tokens are primarily function words, affixes, or short morphemes that serve as grammatical connectors, markers of tense, aspect, or case, and components of compound or derived words. They often appear at the boundaries of phrases or as part of multi-token expressions, reflecting their role in structuring sentences and conveying relationships between ideas in Thai text.	0.78	0.75	0.87	0.66	0.64	0.67	0.62	0.74
model.layers.10.mlp	Thai	42320	The highlighted tokens are primarily Thai morphemes, syllables, or short words that serve as key semantic or grammatical units within sentences. Many are high-frequency function words, affixes, or roots (such as those denoting comparison, agency, or time), and often appear at the start or end of compound words or phrases. These tokens are crucial for sentence structure, meaning, and cohesion in Thai text.	0.69	0.55	1.00	0.38	0.70	0.57	1.00	0.40
model.layers.10.mlp	Thai	42706	The highlighted tokens are often initial syllables, morphemes, or characters at the start of words or names, especially in multilingual or non-Latin scripts, and are frequently associated with proper nouns, place names, or key semantic units within a sentence.	0.54	0.65	0.52	0.86	0.61	0.71	0.57	0.96
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	0.72	1.00	0.56	0.81	0.77	1.00	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	0.88	0.95	0.82	0.68	0.75	0.61	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	0.62	0.48	0.86	0.53	0.67	0.52	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	0.56	0.95	0.40	0.72	0.63	0.92	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.	0.70	0.63	0.81	0.52	0.76	0.74	0.81	0.68