Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.8.mlp	Thai	113053	The highlighted tokens are primarily function words, affixes, or short morphemes in Thai and Bulgarian, often marking grammatical relationships, conjunctions, pronouns, or forming part of common expressions. These tokens are frequently used to construct meaning, indicate possession, connect clauses, or modify verbs and nouns, reflecting their high utility and importance in sentence structure.	0.73	0.675	0.848	0.56	0.63	0.641	0.623	0.66
model.layers.9.mlp	Thai	13432	The highlighted tokens are primarily morphemes, syllables, or word segments in Thai, Hindi, and other languages, often marking key semantic units such as time periods (e.g., \"century\"), proper nouns, or important grammatical structures. There is a strong emphasis on tokens that form or contribute to compound nouns, temporal expressions, and institutional or historical references, especially those denoting centuries or significant eras.	0.83	0.825	0.851	0.8	0.81	0.8	0.844	0.76
model.layers.9.mlp	Thai	14922	The highlighted tokens are primarily proper nouns, time expressions, and function words in both Latin and Thai scripts, often marking names, time references, or grammatical connectors. In Thai, there is frequent emphasis on words or morphemes indicating time, quantity, or conjunctions, as well as on proper names and technical terms.	0.69	0.71	0.667	0.76	0.6	0.524	0.647	0.44
model.layers.9.mlp	Thai	51329	The highlighted tokens are primarily Thai morphemes, prefixes, or roots that form the core of verbs, nouns, and function words, often marking tense, aspect, negation, or agency. These tokens frequently appear at the start or within compound words and are essential for constructing meaning in Thai sentences.	0.61	0.361	1	0.22	0.68	0.529	1	0.36
model.layers.9.mlp	Thai	62908	The highlighted tokens are predominantly function words, affixes, or short morphemes in Thai, Hindi, and English, often marking grammatical relationships, conjunctions, pronouns, or forming part of set phrases and idiomatic expressions. These tokens are crucial for sentence structure, meaning, and cohesion, frequently appearing at clause boundaries or as connectors within and between sentences.	0.67	0.673	0.667	0.68	0.65	0.696	0.615	0.8
model.layers.9.mlp	Thai	103134	The highlighted tokens are primarily common Thai morphemes, function words, and affixes such as particles, pronouns, prepositions, conjunctions, and comparative or plural markers. These tokens are essential for grammatical structure, sentence cohesion, and meaning, often appearing at word boundaries or as part of compound words. Their frequent activation reflects their foundational role in Thai syntax and morphology.	0.75	0.667	1	0.5	0.74	0.667	0.929	0.52
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	0.901	1	0.82	0.92	0.913	1	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.	0.5	0.658	0.5	0.96	0.51	0.662	0.505	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	0.868	0.66	0.64	0.673	0.617	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	0.551	1	0.38	0.7	0.571	1	0.4
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	0.652	0.524	0.86	0.61	0.711	0.565	0.96