Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
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	0.35	0.52	0.26	0.53	0.28	0.60	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	1.00	0.68	0.88	0.87	0.98	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\", \"ยื\", \"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	0.41	1.00	0.26	0.67	0.51	1.00	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	0.38	0.72	0.26	0.77	0.73	0.89	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	0.65	0.53	0.82	0.54	0.68	0.52	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	0.61	1.00	0.44	0.71	0.63	0.86	0.50
model.layers.13.mlp	Thai	38988	The important tokens are primarily Thai morphemes, syllables, or word fragments, often marking the core of nouns, verbs, or key modifiers. These tokens frequently appear at the start or within compound words, proper nouns, or technical terms, and are often associated with semantic content or grammatical function, such as denoting objects, actions, or attributes. The activations highlight meaningful subword units that contribute to the overall meaning and structure of the sentence.	0.71	0.59	1.00	0.42	0.76	0.70	0.93	0.56
model.layers.13.mlp	Thai	40808	The highlighted tokens are word fragments, often suffixes or inflections, that appear at the end of words across multiple languages, indicating morphological boundaries or grammatical modifications.	0.51	0.52	0.51	0.52	0.57	0.68	0.54	0.92
model.layers.13.mlp	Thai	74014	The highlighted tokens are primarily function words, affixes, and common morphemes in Thai, as well as high-frequency content words and syllables. These tokens often serve as grammatical connectors, markers of tense, aspect, or possession, and are essential for sentence structure and meaning. The pattern reflects the importance of short, frequent, and semantically central elements in Thai text, including pronouns, particles, conjunctions, and key noun or verb roots.	0.80	0.75	1.00	0.60	0.78	0.73	0.97	0.58
model.layers.13.mlp	Thai	85051	The highlighted tokens are predominantly morphemes, suffixes, or short function words across multiple languages, often marking grammatical relationships, inflections, or forming parts of compound words. These elements are crucial for the syntactic and morphological structure of sentences, indicating tense, case, possession, comparison, or serving as connectors.	0.53	0.67	0.52	0.96	0.55	0.68	0.53	0.96
model.layers.13.mlp	Thai	125546	The Devanagari character \"ब\" (and its variants) is frequently activated, often as the initial consonant in Hindi words, especially at the start of syllables or morphemes. This pattern also appears in Thai script with the character \"บ\", indicating a focus on the initial consonant in words across different Indic and Southeast Asian languages.	0.75	0.72	0.82	0.64	0.73	0.64	0.96	0.48