Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.14.mlp	Vietnamese	73696	The highlighted tokens are primarily Vietnamese words and phrases that represent key semantic units such as nouns, verbs, adjectives, and named entities, often marking important information, actions, or attributes within sentences. These tokens frequently correspond to the main subject, object, or action, and sometimes to quantifiers, time expressions, or locations, reflecting the core informational structure of Vietnamese text.	0.88	0.867	0.975	0.78	0.87	0.851	1	0.74
model.layers.14.mlp	Vietnamese	111142	The highlighted tokens are primarily function words, grammatical particles, and common morphemes from multiple languages, including Vietnamese, Thai, Hindi, Russian, French, and Spanish. These tokens often serve as connectors, case markers, prepositions, conjunctions, pronouns, or inflectional endings, and are essential for sentence structure and meaning across diverse linguistic contexts.	0.55	0.656	0.531	0.86	0.51	0.642	0.506	0.88
model.layers.14.mlp	Vietnamese	121534	The important tokens are primarily sentence-ending punctuation, quotation marks, and common Vietnamese word endings or particles, often marking the boundaries of sentences, clauses, or quoted speech, as well as frequent grammatical or functional morphemes.	0.85	0.842	0.889	0.8	0.92	0.925	0.875	0.98
model.layers.14.mlp	Vietnamese	130249	The highlighted tokens are primarily Vietnamese words and phrases, often marking key nouns, verbs, or descriptive elements within sentences. These tokens frequently denote important actions, objects, or attributes, and sometimes include grammatical particles or function words that are essential for sentence structure. The activations tend to focus on semantically significant content words and occasionally on connectors or modifiers that clarify relationships or provide context within the sentence.	0.95	0.948	0.979	0.92	0.96	0.958	1	0.92
model.layers.15.mlp	Vietnamese	489	The highlighted tokens are predominantly Vietnamese function words, noun phrases, and key content words that structure sentences, indicate relationships, or specify important entities and actions, often marking organizational, procedural, or descriptive information within formal or informational contexts.	1	1	1	1	0.98	0.98	0.962	1
model.layers.15.mlp	Vietnamese	56098	The highlighted tokens are primarily Vietnamese morphemes, syllables, or word segments, often marking the start, end, or core of words and phrases. They frequently correspond to meaningful units in Vietnamese grammar or vocabulary, such as nouns, verbs, adjectives, or function words, and sometimes appear at points of syntactic or semantic importance within sentences. The activations also include some English morphemes and punctuation, but the dominant pattern is the focus on Vietnamese linguistic units.	0.64	0.438	1	0.28	0.7	0.583	0.955	0.42
model.layers.15.mlp	Vietnamese	92076	The highlighted tokens are primarily Vietnamese words and morphemes that form key components of noun phrases, time expressions, geographic locations, and institutional or organizational names. These tokens often mark the boundaries or heads of important entities, such as countries, cities, months, organizations, or significant events, and are frequently used in formal or encyclopedic contexts to convey factual or structural information.	0.69	0.563	0.952	0.4	0.73	0.64	0.96	0.48
model.layers.15.mlp	Vietnamese	114319	The highlighted tokens are primarily Vietnamese words and morphemes that serve as key components in forming noun phrases, verb phrases, and expressing grammatical relationships. These tokens often include function words, affixes, and high-frequency content words that are essential for sentence structure, topic indication, or semantic emphasis within conversational or descriptive contexts.	0.72	0.611	1	0.44	0.74	0.649	1	0.48