Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.13.mlp	Vietnamese	7310	The highlighted tokens are primarily Vietnamese and some foreign proper nouns, place names, and common nouns, as well as morphemes and suffixes that form meaningful units in Vietnamese. These include country names, city names, administrative regions, and frequently used words or affixes that denote locations, people, or abstract concepts. The pattern reflects a focus on semantic units that are important for understanding context, especially in geographical, cultural, or institutional references.	0.9	0.9	0.9	0.9	0.95	0.95	0.941	0.96
model.layers.13.mlp	Vietnamese	19743	The highlighted tokens are primarily Vietnamese words and morphemes that serve as key content carriers, including nouns, verbs, adjectives, and numerals, as well as grammatical markers and affixes. These tokens often appear at the start or end of words, or as standalone function words, and are crucial for conveying the main meaning, structure, or relationships within sentences. The activations tend to focus on semantically significant elements, such as entities, actions, quantities, and modifiers, reflecting their importance in sentence comprehension and information extraction.	0.78	0.718	1	0.56	0.81	0.765	1	0.62
model.layers.13.mlp	Vietnamese	65213	The highlighted tokens are predominantly Vietnamese morphemes, often forming parts of compound nouns, place names, or descriptive phrases. These tokens frequently appear at the beginning or end of words and are commonly used in proper nouns, geographic locations, and formal or literary expressions. The pattern reflects the morphological structure of Vietnamese, where meaning is built from combining short, meaningful syllables.	0.94	0.936	1	0.88	0.96	0.959	0.979	0.94
model.layers.13.mlp	Vietnamese	75072	The highlighted tokens are primarily Vietnamese morphemes, syllables, or words, often forming meaningful units such as nouns, verbs, adjectives, or grammatical markers. Many are roots or affixes that combine to create compound words or phrases, and several are parts of common collocations or idiomatic expressions. The activations frequently correspond to semantically significant or content-bearing elements within Vietnamese sentences.	0.63	0.431	0.933	0.28	0.67	0.522	0.947	0.36
model.layers.13.mlp	Vietnamese	77971	The highlighted tokens are primarily function words, pronouns, common verbs, and particles in Vietnamese, often marking grammatical structure, subject/object relationships, or indicating tense, aspect, and modality. These tokens are essential for sentence cohesion and meaning, frequently appearing at clause boundaries or as connectors within conversational or narrative contexts.	0.91	0.901	1	0.82	0.87	0.851	1	0.74
model.layers.13.mlp	Vietnamese	98734	The highlighted tokens are primarily function words, common nouns, and grammatical particles in Vietnamese, often marking relationships between entities, locations, time, or actions. These tokens frequently appear in prepositional phrases, noun phrases, and as connectors, reflecting their importance in structuring sentences and conveying meaning in Vietnamese text.	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
model.layers.13.mlp	Vietnamese	110071	The highlighted spans are predominantly noun phrases, verb phrases, or collocations that convey key actions, attributes, or relationships within a sentence. These often include subjects, objects, or descriptive elements central to the meaning, such as people, places, activities, emotions, or qualities. The selections frequently capture the core semantic content or the main event of the sentence, often involving personal pronouns, verbs of being or action, and their direct complements or modifiers.	0.55	0.667	0.529	0.9	0.67	0.602	0.758	0.5
model.layers.14.mlp	Vietnamese	54719	The highlighted tokens are predominantly Vietnamese morphemes, syllables, or word stems, often at the beginning of words or as standalone syllables, including both lowercase and uppercase forms. These tokens frequently represent meaningful units in Vietnamese, such as prefixes, roots, or grammatical markers, and are often used to construct or modify words. The pattern reflects the tokenization of Vietnamese text into short, meaningful segments that align with the language's syllabic and morphological structure.	0.67	0.522	0.947	0.36	0.62	0.525	0.7	0.42
model.layers.14.mlp	Vietnamese	59851	The highlighted tokens are primarily Vietnamese nouns, noun phrases, and key modifiers that denote entities, locations, roles, or important actions and attributes within sentences. These tokens often serve as the main subjects, objects, or descriptive elements, and are frequently used to convey core information, context, or relationships in the text.	0.66	0.564	0.786	0.44	0.66	0.528	0.864	0.38
model.layers.14.mlp	Vietnamese	71834	The highlighted tokens are primarily function words, common particles, conjunctions, pronouns, and frequently used morphemes in Vietnamese, as well as punctuation and quotation markers. These tokens are essential for sentence structure, grammatical relationships, and discourse flow, often marking beginnings, endings, or transitions in sentences and direct speech.	0.83	0.805	0.946	0.7	0.78	0.771	0.804	0.74