Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.11.mlp	Vietnamese	4065	The highlighted tokens are predominantly Vietnamese content words, including nouns, verbs, adjectives, and function words that are central to sentence meaning. There is a strong emphasis on tokens related to usage, action, and description, as well as frequent marking of high-frequency morphemes and word stems. The activations often correspond to key semantic units or grammatical connectors that structure information, especially in formal, descriptive, or explanatory contexts.	0.76	0.7	0.933	0.56	0.8	0.756	0.969	0.62
model.layers.11.mlp	Vietnamese	58535	The highlighted tokens are primarily Vietnamese function words, common nouns, and morphemes that serve as grammatical connectors, markers of agency, possession, or modality, as well as components of compound words and idiomatic expressions. These tokens are essential for sentence structure, meaning composition, and the formation of complex ideas in Vietnamese text.	0.86	0.837	1	0.72	0.85	0.828	0.973	0.72
model.layers.11.mlp	Vietnamese	66074	The highlighted tokens are predominantly Vietnamese morphemes, words, or word parts that serve as key semantic units, including nouns, verbs, adjectives, and function words. These tokens often represent core concepts, actions, or attributes within a sentence, and include both standalone words and meaningful syllabic components in compound or multi-word expressions. The pattern reflects the analytic, syllable-based structure of Vietnamese, where individual morphemes carry significant meaning and are frequently combined to form compound words or phrases.	0.92	0.913	1	0.84	0.93	0.926	0.978	0.88
model.layers.12.mlp	Vietnamese	44940	The highlighted tokens are predominantly Vietnamese compound nouns and noun phrases, often denoting institutions, services, professions, locations, or abstract concepts. These phrases frequently consist of two or more words forming a semantic unit, and are commonly used in formal, informational, or descriptive contexts. Many relate to administrative, technical, or organizational domains, and often appear as objects, subjects, or key elements within sentences.	0.62	0.424	0.875	0.28	0.62	0.387	1	0.24
model.layers.12.mlp	Vietnamese	56617	The highlighted segments are predominantly Vietnamese noun and verb phrases, often marking subjects, objects, or actions within sentences. These segments frequently include function words (such as pronouns, conjunctions, or prepositions) and are commonly used to introduce, describe, or relate entities and events. The activations tend to focus on core content words and their immediate grammatical context, reflecting the structure and flow of Vietnamese sentences.	0.71	0.592	1	0.42	0.69	0.551	1	0.38
model.layers.12.mlp	Vietnamese	63803	The highlighted tokens are primarily Vietnamese morphemes, often functioning as prefixes, suffixes, or standalone words that form the core of compound nouns, adjectives, or verbs. These tokens frequently appear in formal, technical, or administrative contexts, and are commonly used in constructing terms related to safety, security, organizations, professions, and abstract concepts. The activations tend to focus on meaningful morphemes that contribute significantly to the semantics of the phrase or sentence.	0.76	0.692	0.964	0.54	0.78	0.738	0.912	0.62
model.layers.12.mlp	Vietnamese	92851	The highlighted tokens are primarily Vietnamese morphemes, especially syllables containing the diacritic \"ê\" or \"ệ\", and other common Vietnamese syllables or word parts. These often appear in the middle or end of words, and are frequently found in nouns, verbs, and adjectives, reflecting the syllabic and tonal structure of Vietnamese language.	0.83	0.813	0.902	0.74	0.85	0.828	0.973	0.72
model.layers.12.mlp	Vietnamese	100653	The highlighted tokens are primarily Vietnamese morphemes, words, or short phrases that serve as key semantic units within sentences. They often include nouns, verbs, adjectives, and function words that are essential for conveying the main meaning, describing actions, objects, quantities, or relationships. Many are components of compound words or set expressions, and their importance is context-dependent, frequently marking the core informational content or structural elements of the sentence.	0.97	0.969	1	0.94	0.92	0.922	0.904	0.94
model.layers.12.mlp	Vietnamese	124904	The highlighted tokens are predominantly Vietnamese morphemes, syllables, or word segments, often marking the start or end of compound words, place names, or key nouns. These tokens frequently appear in contexts involving geography, administration, or descriptive attributes, and are often components of multi-syllabic words or phrases central to the sentence meaning.	0.98	0.98	1	0.96	0.95	0.95	0.941	0.96