Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.0.mlp	Vietnamese	11601	The highlighted tokens are predominantly function words, pronouns, and common connectors in Vietnamese, such as possessive and relational markers, as well as frequently used nouns and adjectives. These tokens often serve to indicate relationships between entities, possession, inclusion, or specification, and are essential for structuring complex noun phrases and clauses.	0.56	0.24	0.88	0.14	0.63	0.45	0.88	0.30
model.layers.0.mlp	Vietnamese	102922	The highlighted tokens are primarily Vietnamese noun and verb phrases, often denoting roles, actions, or attributes related to people, objects, or abstract concepts. There is a strong emphasis on compound nouns (such as \"bản thân\", \"gia đình\", \"âm thanh\") and action phrases, as well as words indicating possession, agency, or descriptive qualities. Many tokens are part of formal, informational, or instructional contexts, and often appear in collocations or set expressions.	0.62	0.39	1.00	0.24	0.63	0.41	1.00	0.26
model.layers.3.mlp	Vietnamese	14622	The highlighted tokens are predominantly morphemes, syllables, or short word fragments that are common in Vietnamese and other languages, often marking the start or end of words, or forming meaningful units within compound words or names. These fragments frequently appear in proper nouns, place names, and common vocabulary, and are sometimes associated with grammatical or semantic roles in the sentence.	0.75	0.68	0.96	0.52	0.76	0.75	0.80	0.70
model.layers.4.mlp	Vietnamese	110587	The highlighted tokens are predominantly Vietnamese morphemes, syllables, or word parts, often at the beginning or end of words, including common affixes, roots, and particles. Many are components of compound words, proper nouns, or grammatical markers, and often correspond to meaningful subword units in Vietnamese text.	0.87	0.85	0.97	0.76	0.82	0.82	0.83	0.80
model.layers.5.mlp	Vietnamese	49501	The highlighted tokens are characteristic morphemes, suffixes, or letter clusters commonly found in Romanian and related Eastern European place names, river names, and surnames, often marking grammatical or regional features.	0.52	0.23	0.58	0.14	0.55	0.40	0.60	0.30
model.layers.7.mlp	Vietnamese	28785	The highlighted tokens are predominantly Vietnamese morphemes, syllables, or word stems, often at the beginning of compound words or phrases. These tokens frequently represent meaningful units that combine to form nouns, verbs, or adjectives, and are commonly used in word formation, compounding, or as grammatical markers in Vietnamese text.	0.70	0.57	1.00	0.40	0.74	0.67	0.93	0.52
model.layers.8.mlp	Vietnamese	79242	The highlighted tokens are often parts of proper nouns, country names, or compound words in Vietnamese and other languages, as well as morphemes or syllables that form meaningful units within words, especially in multi-syllabic or compound constructions. These tokens frequently appear at the beginning or within words that denote places, people, or key concepts, reflecting the morphological structure and semantic importance in the context.	0.70	0.63	0.83	0.50	0.53	0.64	0.52	0.82
model.layers.9.mlp	Vietnamese	61383	The highlighted tokens correspond to names of Cambodian places and landmarks, especially those related to Tonle Sap, Phnom Krom, Angkor, and Siem Reap, often appearing in multiple languages and scripts, with activations on both full names and their constituent parts.	0.51	0.04	1.00	0.02	0.51	0.04	1.00	0.02
model.layers.10.mlp	Vietnamese	42004	The highlighted tokens are predominantly Vietnamese words and phrases that serve as key content elements in sentences, such as nouns, verbs, and modifiers, often marking important entities, actions, or temporal and locational references. These tokens frequently appear at the start or within core syntactic units, indicating their role in conveying the main informational content of each sentence.	0.74	0.67	0.93	0.52	0.76	0.68	1.00	0.52
model.layers.10.mlp	Vietnamese	42167	The highlighted tokens are primarily Vietnamese morphemes, syllables, or word segments, often marking the start of words or compounds, and include both native and Sino-Vietnamese roots. Many are high-frequency function words, affixes, or common noun/adjective stems, and several are associated with grammatical or semantic roles such as denoting people, places, actions, or qualities. There is a notable emphasis on tokens with the \"ê\"/\"ệ\"/\"é\"/\"è\" vowel, as well as on tokens that form part of compound nouns, proper names, or technical terms.	0.66	0.51	0.90	0.36	0.74	0.65	1.00	0.48