Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
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	0.592	1	0.42	0.76	0.7	0.933	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	0.515	0.51	0.52	0.57	0.681	0.541	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.	0.8	0.75	1	0.6	0.78	0.725	0.967	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	0.671	0.516	0.96	0.55	0.681	0.527	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	0.719	0.821	0.64	0.73	0.64	0.96	0.48
model.layers.14.mlp	Thai	11744	The highlighted tokens are primarily function words, pronouns, particles, and common morphemes in Thai, often marking grammatical relationships, sentence structure, or serving as connectors. There is a strong emphasis on words that indicate possession, agency, time, location, and conjunctions, as well as frequent use of polite particles and verb auxiliaries. These tokens are essential for the syntactic and semantic coherence of Thai sentences.	0.73	0.658	0.897	0.52	0.76	0.684	1	0.52
model.layers.14.mlp	Thai	22779	The highlighted segments are primarily Thai words, phrases, or morphemes that serve as meaningful units within sentences, including nouns, verbs, particles, and function words. These segments often correspond to syntactic or semantic boundaries, such as names, actions, objects, or grammatical markers, and sometimes include foreign words or transliterations. The activations tend to focus on morphemes or syllables that carry core meaning or grammatical function within the context.	0.88	0.864	1	0.76	0.86	0.841	0.974	0.74
model.layers.14.mlp	Thai	33861	The highlighted tokens are primarily Thai morphemes, syllables, or short words that serve as grammatical markers, connectors, or key semantic units within sentences. These include function words, affixes, and core content words that are essential for sentence structure and meaning, such as those indicating possession, agency, location, comparison, or action. The pattern reflects the importance of these units in Thai syntax and information flow, often marking relationships between clauses, specifying entities, or denoting actions and attributes.	0.99	0.99	1	0.98	0.97	0.969	1	0.94
model.layers.14.mlp	Thai	34721	The highlighted tokens are overwhelmingly function words, pronouns, and common morphemes or syllables that serve as grammatical connectors in Thai, such as markers for possession, location, time, or subject/object reference. There is a strong emphasis on high-frequency, short tokens that are essential for sentence structure and meaning, including prefixes, suffixes, and particles that modify or clarify the main content words. These tokens are crucial for the cohesion and flow of Thai sentences.	0.66	0.485	1	0.32	0.57	0.469	0.613	0.38
model.layers.14.mlp	Thai	54904	The highlighted tokens are primarily morphemes, affixes, or short words that serve as grammatical connectors, markers of tense, aspect, or case, and components of compound words in Thai. They often appear at the boundaries of words or phrases, contributing to the structure and meaning of sentences by indicating relationships, possession, comparison, or forming part of proper nouns and technical terms.	0.87	0.851	1	0.74	0.87	0.866	0.894	0.84