Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.7.mlp	Thai	128501	The highlighted tokens are often subword units or morphemes within names, words, or phrases across multiple languages, frequently marking meaningful components such as name parts, suffixes, or linguistic roots, and are not limited to a single language or script.	0.46	0.60	0.48	0.82	0.52	0.68	0.51	1.00
model.layers.8.mlp	Thai	48547	The highlighted tokens are predominantly Thai morphemes, syllables, or short words that frequently serve as functional units in compound words, proper nouns, or technical terms. Many activations correspond to prefixes, suffixes, or core morphemes that contribute to the grammatical structure or meaning of the phrase, such as indicating time, quantity, location, or agency. There is a recurring emphasis on tokens that form part of administrative, temporal, or descriptive expressions, as well as those that are central to the construction of compound nouns and formal terminology.	0.86	0.84	0.97	0.74	0.86	0.84	0.95	0.76
model.layers.8.mlp	Thai	55844	The highlighted tokens are short function words, prepositions, conjunctions, or grammatical particles in various languages (such as Bulgarian, Thai, Russian), often marking relationships between phrases or clauses, or serving as connectors within sentences.	0.62	0.70	0.58	0.90	0.64	0.73	0.59	0.96
model.layers.8.mlp	Thai	56232	The highlighted tokens are primarily morphemes, syllables, or short word segments that form the core of Thai words, often marking key semantic or grammatical units such as nouns, verbs, or important modifiers. These segments frequently appear at the beginning or within words, and are often associated with the main meaning or function of the word in the sentence.	0.64	0.46	0.94	0.30	0.67	0.52	0.95	0.36
model.layers.8.mlp	Thai	113053	The highlighted tokens are primarily function words, affixes, or short morphemes in Thai and Bulgarian, often marking grammatical relationships, conjunctions, pronouns, or forming part of common expressions. These tokens are frequently used to construct meaning, indicate possession, connect clauses, or modify verbs and nouns, reflecting their high utility and importance in sentence structure.	0.73	0.68	0.85	0.56	0.63	0.64	0.62	0.66
model.layers.9.mlp	Thai	13432	The highlighted tokens are primarily morphemes, syllables, or word segments in Thai, Hindi, and other languages, often marking key semantic units such as time periods (e.g., \"century\"), proper nouns, or important grammatical structures. There is a strong emphasis on tokens that form or contribute to compound nouns, temporal expressions, and institutional or historical references, especially those denoting centuries or significant eras.	0.83	0.83	0.85	0.80	0.81	0.80	0.84	0.76
model.layers.9.mlp	Thai	14922	The highlighted tokens are primarily proper nouns, time expressions, and function words in both Latin and Thai scripts, often marking names, time references, or grammatical connectors. In Thai, there is frequent emphasis on words or morphemes indicating time, quantity, or conjunctions, as well as on proper names and technical terms.	0.69	0.71	0.67	0.76	0.60	0.52	0.65	0.44
model.layers.9.mlp	Thai	51329	The highlighted tokens are primarily Thai morphemes, prefixes, or roots that form the core of verbs, nouns, and function words, often marking tense, aspect, negation, or agency. These tokens frequently appear at the start or within compound words and are essential for constructing meaning in Thai sentences.	0.61	0.36	1.00	0.22	0.68	0.53	1.00	0.36
model.layers.9.mlp	Thai	62908	The highlighted tokens are predominantly function words, affixes, or short morphemes in Thai, Hindi, and English, often marking grammatical relationships, conjunctions, pronouns, or forming part of set phrases and idiomatic expressions. These tokens are crucial for sentence structure, meaning, and cohesion, frequently appearing at clause boundaries or as connectors within and between sentences.	0.67	0.67	0.67	0.68	0.65	0.70	0.62	0.80
model.layers.9.mlp	Thai	103134	The highlighted tokens are primarily common Thai morphemes, function words, and affixes such as particles, pronouns, prepositions, conjunctions, and comparative or plural markers. These tokens are essential for grammatical structure, sentence cohesion, and meaning, often appearing at word boundaries or as part of compound words. Their frequent activation reflects their foundational role in Thai syntax and morphology.	0.75	0.67	1.00	0.50	0.74	0.67	0.93	0.52