Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.2.mlp	Chinese	18693	The highlighted tokens are primarily Chinese, Japanese, and some corrupted or missing characters, often marking key nouns, verbs, or morphemes that carry core semantic meaning in a sentence, such as names, actions, or important objects. These tokens frequently appear at the start of compound words or phrases, and are often associated with high informational content or serve as grammatical anchors in the text.	0.70	0.60	0.92	0.44	0.89	0.89	0.88	0.90
model.layers.6.mlp	Chinese	126429	The highlighted segments are primarily noun phrases, verb phrases, or set expressions that convey key information, actions, or attributions within a sentence. They often encapsulate the main subject, object, or predicate, and are central to the meaning or structure of the sentence. These segments frequently include proper nouns, technical terms, or idiomatic expressions, and are often used to introduce, describe, or summarize important events, entities, or concepts.	0.50	0.46	0.50	0.42	0.59	0.39	0.77	0.26
model.layers.7.mlp	Chinese	82671	The highlighted tokens are primarily function words, particles, and common connectors in Chinese (and some in Vietnamese and English), such as grammatical markers, conjunctions, and punctuation. These tokens are essential for sentence structure, logical flow, and connecting clauses or ideas, often marking relationships, transitions, or the boundaries of statements.	0.74	0.72	0.79	0.66	0.78	0.79	0.76	0.82
model.layers.9.mlp	Chinese	41091	The highlighted tokens are primarily nouns, noun phrases, or key descriptive terms that denote entities, objects, people, places, or important concepts within the sentence. These tokens often serve as the main subject or object, or otherwise carry the core informational content of the sentence.	0.51	0.66	0.51	0.96	0.62	0.51	0.71	0.40
model.layers.10.mlp	Chinese	86769	The highlighted tokens are predominantly nouns, noun phrases, and key modifiers that denote concrete objects, abstract concepts, time spans, roles, and relationships. Many are compound or multi-character words central to the sentence's meaning, often marking scientific, technical, or descriptive content. There is a strong emphasis on terms that specify entities, durations, properties, and causal or functional relationships, reflecting a focus on information-rich, content-bearing elements in Chinese text.	0.65	0.60	0.70	0.52	0.65	0.65	0.65	0.66
model.layers.10.mlp	Chinese	119196	The highlighted tokens are primarily Chinese idioms, set phrases, or compound words, often conveying abstract, figurative, or formal meanings. Many are four-character expressions or fixed collocations, and several appear at sentence boundaries or in contexts emphasizing notable events, qualities, or actions.	0.66	0.49	1.00	0.32	0.62	0.39	1.00	0.24