Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.11.mlp	Chinese	240	The highlighted tokens are predominantly Chinese nouns and morphemes that denote concrete objects, locations, roles, or abstract concepts, often forming the core of compound words or phrases central to the sentence’s meaning. These tokens frequently appear as the main subject, object, or key element in descriptive or informative contexts.	0.71	0.681	0.756	0.62	0.79	0.783	0.818	0.75
model.layers.11.mlp	Chinese	38263	The highlighted tokens are primarily Chinese characters with high semantic content, often representing nouns, adjectives, or verbs central to the meaning of the sentence, such as objects, actions, or descriptive qualities. There is also frequent activation on corrupted or unknown tokens, likely due to encoding errors or missing data, which are treated as significant by the model.	0.68	0.579	0.846	0.44	0.79	0.807	0.746	0.88
model.layers.11.mlp	Chinese	40026	The highlighted tokens are primarily function words, grammatical particles, and common nouns that serve as structural connectors or key semantic units in Chinese sentences, often marking relationships, actions, time, or entities essential for sentence meaning and coherence.	0.7	0.571	1	0.4	0.65	0.673	0.632	0.72
model.layers.11.mlp	Chinese	65660	The highlighted tokens are primarily pronouns and common nouns referring to people, groups, or entities (such as \"you\", \"we\", \"they\", \"he\", \"she\", \"person\", \"child\", \"everyone\", \"people\"), as well as words indicating possession, time, or location. These tokens often serve as subjects or objects in sentences, anchoring statements to participants, agents, or relevant entities, and are central to the structure and meaning of the text.	0.47	0.539	0.477	0.62	0.73	0.675	0.848	0.56
model.layers.11.mlp	Chinese	73417	The highlighted tokens are primarily nouns, noun phrases, and key modifiers that denote entities, objects, roles, or important concepts within a sentence. These tokens often serve as the main subject or object, or as essential descriptive elements that convey the core meaning or function of the sentence.	0.53	0.676	0.516	0.98	0.73	0.71	0.767	0.66
model.layers.12.mlp	Chinese	104224	The highlighted tokens are predominantly Chinese nouns, verbs, and adjectives that denote key concepts, actions, or attributes central to the meaning of each sentence, often marking important entities, processes, or relationships within scientific, medical, or factual contexts.	0.65	0.557	0.759	0.44	0.68	0.59	0.821	0.46
model.layers.13.mlp	Chinese	113658	The highlighted segments are predominantly noun phrases, event descriptions, or organizational terms, often marking key subjects, activities, or entities within sentences. These segments frequently denote the main topic, object, or action, and are central to the informational structure of the text.	0.6	0.643	0.581	0.72	0.66	0.585	0.75	0.48
model.layers.13.mlp	Chinese	118430	The highlighted tokens are primarily function words, connectors, and common morphemes that structure sentences, indicate relationships, or form compound meanings in Chinese. These include possessives, conjunctions, prepositions, pronouns, and frequently used descriptive or comparative elements, as well as key morphemes in idiomatic or compound expressions. Their importance lies in linking ideas, clarifying relationships, and forming the grammatical backbone of the text.	0.57	0.394	0.667	0.28	0.62	0.721	0.57	0.98