Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.14.mlp	Chinese	12944	The highlighted tokens are primarily nouns, proper nouns, and key morphemes that denote people, organizations, places, time expressions, and important actions or states. These tokens often serve as anchors for entities, events, or relationships within sentences, and are frequently found in news, official, or factual reporting contexts. The activations focus on tokens that carry core semantic information, such as names, roles, locations, temporal markers, and institutional or organizational terms, which are essential for understanding the main content and structure of informational Chinese text.	0.59	0.506	0.636	0.42	0.56	0.522	0.571	0.48
model.layers.14.mlp	Chinese	12987	The highlighted tokens are primarily function words, grammatical particles, and common morphemes that structure Chinese sentences, such as possessives, prepositions, conjunctions, and markers of time, place, or sequence. These tokens are essential for indicating relationships between entities, actions, and events, and for connecting clauses or phrases within complex informational or narrative contexts.	0.65	0.533	0.8	0.4	0.73	0.765	0.677	0.88
model.layers.14.mlp	Chinese	49167	The highlighted tokens are predominantly nouns, noun phrases, or compound words that denote concrete objects, locations, roles, or abstract concepts, often serving as key subjects or objects within sentences. Many tokens are also associated with actions, states, or attributes relevant to the main informational content, and frequently appear in contexts describing possession, function, or relationships.	0.48	0.644	0.49	0.94	0.72	0.667	0.824	0.56
model.layers.14.mlp	Chinese	66584	The highlighted tokens are primarily function words, common nouns, and morphemes that serve as grammatical connectors or denote relationships, objects, or actions. These tokens are essential for sentence structure, topic continuity, and the expression of key concepts in technical, scientific, and descriptive contexts.	0.48	0.639	0.489	0.92	0.54	0.676	0.522	0.96
model.layers.14.mlp	Chinese	88535	The highlighted tokens are primarily punctuation marks, conjunctions, and grammatical particles that serve as structural or connective elements in various languages, often marking boundaries, relationships, or transitions within sentences.	0.46	0.591	0.476	0.78	0.59	0.65	0.567	0.76
model.layers.14.mlp	Chinese	103432	The highlighted tokens are primarily Chinese nouns, verbs, and function words that serve as key elements in sentence structure, often marking subjects, objects, actions, or important modifiers. These tokens frequently appear at the start or end of phrases, and are essential for conveying the main meaning or grammatical relationships within the sentence.	0.62	0.472	0.773	0.34	0.66	0.638	0.682	0.6
model.layers.14.mlp	Chinese	112457	The highlighted tokens are primarily Chinese nouns, noun phrases, and descriptive terms that denote objects, locations, people, events, or measurable attributes, often serving as key informational elements or subjects within sentences. There is a focus on concrete entities, roles, and quantifiable features, as well as some action or state descriptors that are central to the meaning of the sentence.	0.68	0.709	0.65	0.78	0.67	0.612	0.743	0.52
model.layers.15.mlp	Chinese	11994	The highlighted tokens are primarily function words, particles, punctuation, and common morphemes that structure Chinese sentences, indicate relationships, or mark boundaries between phrases, clauses, or items in lists. These elements are essential for grammatical cohesion, logical flow, and segmentation in Chinese text.	0.63	0.493	0.783	0.36	0.69	0.716	0.661	0.78
model.layers.15.mlp	Chinese	38094	The highlighted tokens are primarily nouns and noun phrases denoting objects, locations, organizations, features, or abstract concepts, as well as verbs and adjectives related to actions, states, or qualities. These tokens often represent key entities, actions, or attributes central to the meaning of each sentence, and are frequently found in contexts involving technology, communication, services, or descriptive information.	0.58	0.7	0.544	0.98	0.48	0.458	0.478	0.44
model.layers.15.mlp	Chinese	114546	The highlighted tokens are predominantly Chinese characters or short phrases, often representing proper nouns such as personal names, place names, and institutional titles, as well as key content words like roles, actions, or descriptors. There is a strong emphasis on tokens that are semantically significant within a sentence, especially those that denote entities, locations, or important actions, and these often appear in contexts involving lists, attributions, or descriptions.	0.78	0.75	0.868	0.66	0.77	0.729	0.886	0.62