Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.15.mlp	Japanese	4131	The highlighted tokens are primarily Japanese grammatical particles, noun and verb suffixes, and function words that structure sentences, indicate relationships, or mark topics, objects, and actions. These tokens are essential for conveying grammatical roles, sentence boundaries, and discourse structure in Japanese text.	0.87	0.851	1	0.74	0.9	0.891	0.976	0.82
model.layers.15.mlp	Japanese	12716	The highlighted tokens are predominantly Japanese grammatical particles, auxiliary verbs, and common function words that serve to connect, modify, or clarify relationships between content words in sentences. These include particles marking case, topic, possession, conjunctions, and suffixes for adjectives and verbs, as well as common noun and verb endings. Their frequent activation reflects their essential role in Japanese sentence structure and meaning.	0.81	0.771	0.97	0.64	0.82	0.786	0.971	0.66
model.layers.15.mlp	Japanese	25135	The highlighted tokens are primarily Japanese grammatical endings, auxiliary verbs, particles, and polite or formal sentence endings. These elements are essential for expressing tense, aspect, negation, modality, and politeness, and for connecting clauses or marking sentence boundaries. They play a crucial role in structuring Japanese sentences and conveying nuanced meaning.	0.73	0.63	1	0.46	0.83	0.795	1	0.66
model.layers.15.mlp	Japanese	35007	The highlighted tokens are primarily Japanese nouns, noun phrases, and compound words, often denoting institutions, official documents, locations, services, or key concepts relevant to travel, administration, and procedures. Many are used in formal or informational contexts, and often appear as part of set expressions or collocations.	0.74	0.649	1	0.48	0.85	0.824	1	0.7
model.layers.15.mlp	Japanese	95097	The text highlights frequent use of Japanese grammatical particles, especially the object marker, and verb or noun phrases that indicate actions, states, or attributes. There is a strong emphasis on function words and suffixes that connect or modify main content words, reflecting the agglutinative and context-dependent nature of Japanese sentence structure.	0.72	0.632	0.923	0.48	0.74	0.74	0.74	0.74
model.layers.15.mlp	Japanese	95151	The highlighted tokens are primarily Japanese content words, including nouns, verbs, and adjectives, often marking key semantic roles or actions within sentences. There is a focus on tokens that convey core meaning, such as objects, actions, and states, as well as grammatical constructions that indicate causality, necessity, or condition. The activations also frequently emphasize parts of compound words and important inflections, reflecting the morphological structure of Japanese.	0.66	0.485	1	0.32	0.7	0.595	0.917	0.44
model.layers.15.mlp	Japanese	102746	The highlighted tokens are primarily Japanese function words, particles, and common morphemes, as well as frequent kanji and katakana components found in names, titles, and grammatical constructions. There is a strong emphasis on elements that structure sentences, indicate relationships, or form part of proper nouns and set phrases, reflecting the syntactic and morphological backbone of Japanese text.	0.81	0.765	1	0.62	0.8	0.767	0.917	0.66