Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.14.mlp	Japanese	17056	The highlighted tokens are primarily Japanese grammatical particles, verb endings, and function words that play key roles in sentence structure, such as marking subjects, objects, possession, and verb conjugations. These elements are essential for conveying relationships, tense, and meaning within Japanese sentences.	0.76	0.684	1	0.52	0.73	0.64	0.96	0.48
model.layers.14.mlp	Japanese	17850	The highlighted tokens are primarily content words—nouns, verbs, and adjectives—that carry core semantic meaning in Japanese sentences, often relating to scientific, technical, or descriptive contexts. These tokens frequently denote key concepts, actions, or properties, and are central to the informational structure of the text.	0.72	0.689	0.775	0.62	0.67	0.667	0.673	0.66
model.layers.14.mlp	Japanese	34692	The highlighted tokens are primarily Japanese content words, such as nouns, verbs, and adjectives, as well as some grammatical particles and endings that are important for sentence structure and meaning. There is also a notable presence of corrupted or placeholder tokens, likely representing unreadable or missing characters. The activations focus on semantically significant morphemes and key syntactic elements within Japanese sentences.	0.68	0.529	1	0.36	0.64	0.526	0.769	0.4
model.layers.14.mlp	Japanese	43240	The highlighted tokens are predominantly Japanese grammatical particles, auxiliary verbs, and function words that serve as syntactic connectors or markers of case, topic, or relation within sentences. These include particles like \"は\", \"の\", \"に\", \"を\", \"で\", \"が\", as well as common auxiliary forms and punctuation. Their frequent activation reflects their essential role in structuring Japanese sentences and conveying grammatical relationships.	0.62	0.406	0.929	0.26	0.71	0.592	1	0.42
model.layers.14.mlp	Japanese	44808	The highlighted tokens are primarily 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, or conjunctions, as well as verb endings and polite forms, which are essential for sentence structure and meaning in Japanese.	0.68	0.543	0.95	0.38	0.67	0.571	0.815	0.44
model.layers.14.mlp	Japanese	77483	The highlighted tokens are primarily nouns, noun phrases, and compound words that denote key concepts, roles, institutions, or technical terms, often central to the informational content of the sentence. These include terms related to education, medicine, law, government, science, and other formal or specialized domains, as well as modifiers and suffixes that form or specify such terms.	0.55	0.628	0.535	0.76	0.73	0.697	0.795	0.62
model.layers.14.mlp	Japanese	85299	The highlighted tokens are morphemes or short words in Italian and Japanese that frequently indicate proximity, location, or grammatical relationships, such as \"vic\" (near) in Italian and particles or connectors like \"の\", \"に\", \"から\", \"で\", \"を\" in Japanese, which serve to link nouns, indicate possession, direction, or other syntactic roles. These tokens are essential for constructing meaning and structure in their respective languages.	0.83	0.809	0.923	0.72	0.81	0.782	0.919	0.68
model.layers.14.mlp	Japanese	97905	The highlighted tokens are primarily functional and grammatical elements in Japanese, such as auxiliary verbs, verb endings, and common particles, which are essential for expressing tense, aspect, modality, negation, and politeness. These tokens often appear at the end of clauses or sentences and are crucial for constructing natural, contextually appropriate Japanese sentences.	0.59	0.305	1	0.18	0.75	0.667	1	0.5
model.layers.14.mlp	Japanese	98700	The most prominent pattern is the frequent highlighting of Japanese grammatical particles such as \"を\", \"に\", \"が\", \"で\", and \"の\", which function as case markers or connectors in sentences. These tokens are essential for indicating grammatical relationships, object marking, and topic or subject identification, and are consistently activated at high importance, reflecting their central role in Japanese sentence structure.	0.65	0.462	1	0.3	0.69	0.551	1	0.38
model.layers.14.mlp	Japanese	119697	The highlighted tokens are primarily functional morphemes, particles, and common suffixes in Japanese, as well as high-frequency content words and grammatical connectors. These elements are essential for sentence structure, meaning, and cohesion, often marking relationships between clauses, indicating grammatical roles, or forming key parts of compound words and set phrases.	0.8	0.762	0.941	0.64	0.66	0.707	0.621	0.82