Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.13.mlp	Japanese	44429	The highlighted segments are primarily Japanese noun and verb phrases, often including particles and auxiliary verbs, that form meaningful units within sentences. These segments frequently encapsulate actions, states, or descriptive attributes, and often include grammatical constructions that connect or modify other elements in the sentence. The pattern reflects the agglutinative and phrase-based structure of Japanese, where meaning is built up through the combination of content words and function words.	0.64	0.46	0.94	0.30	0.71	0.63	0.86	0.50
model.layers.13.mlp	Japanese	72538	The highlighted segments are primarily Japanese noun phrases, compound nouns, or set expressions, often denoting entities, roles, locations, or abstract concepts. These segments frequently appear as important information units within sentences, such as names of organizations, places, activities, or key objects, and are often followed or preceded by particles or punctuation that mark their syntactic or semantic boundaries.	0.92	0.92	0.96	0.88	0.77	0.77	0.77	0.78
model.layers.13.mlp	Japanese	77268	The highlighted tokens are primarily Japanese content words, such as nouns, verbs, and adjectives, as well as grammatical particles and function words that mark relationships or structure within sentences. These tokens often appear at or near phrase or clause boundaries, and include both standalone words and morphemes that contribute to meaning or grammatical function. The activations also frequently emphasize key information, actions, or entities relevant to the sentence's main topic.	0.77	0.71	0.97	0.56	0.73	0.68	0.85	0.56
model.layers.13.mlp	Japanese	80366	The highlighted tokens are primarily nouns, noun phrases, and key verbal expressions that denote actions, states, or important entities within sentences. These tokens often represent core semantic content such as people, places, actions, events, and conditions, as well as grammatical markers that indicate relationships or necessity (e.g., \"必要\", \"場合\", \"方法\", \"情報\", \"参加\", \"確認\", \"利用\", \"予約\"). The activations also frequently emphasize compound words, set phrases, and collocations that are central to the meaning or structure of the sentence, including those that express requirements, comparisons, or results.	0.64	0.70	0.60	0.84	0.67	0.70	0.64	0.76
model.layers.13.mlp	Japanese	84606	The highlighted tokens are primarily Japanese nouns, verbs, and grammatical particles that denote key actions, states, or entities, often marking important legal, administrative, or procedural concepts, as well as essential components of compound words and formal expressions. These tokens frequently appear in contexts involving official processes, regulations, or institutional terminology.	0.60	0.33	1.00	0.20	0.78	0.73	0.97	0.58
model.layers.13.mlp	Japanese	101925	The highlighted tokens are primarily Japanese content words, including nouns, verbs, and adjectives, as well as compound words and grammatical particles that contribute to the core meaning of sentences. There is a focus on tokens that form key semantic units, such as objects, actions, and attributes, often at the end or middle of words, and sometimes include numerals or foreign loanwords. These tokens are essential for conveying the main informational content in Japanese text.	0.62	0.39	1.00	0.24	0.73	0.65	0.93	0.50
model.layers.13.mlp	Japanese	121228	The highlighted tokens are primarily Japanese morphemes, particles, and content words that serve as grammatical markers, connectors, or key semantic units within sentences. They include noun and verb stems, particles indicating case or topic, and common suffixes or endings, often marking boundaries of phrases or important syntactic roles. Some tokens are also punctuation or function words that structure or segment the text.	0.80	0.76	0.97	0.62	0.70	0.67	0.75	0.60
model.layers.13.mlp	Japanese	125863	The highlighted tokens are common Japanese grammatical particles, auxiliary verbs, and function words that structure sentences, indicate relationships, and mark objects, topics, or actions. These tokens are essential for the syntactic and semantic coherence of Japanese text.	0.71	0.60	0.96	0.44	0.85	0.84	0.91	0.78