Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.0.mlp	Japanese	100720	The Japanese particle \"の\" is consistently used as a possessive or attributive marker, linking nouns to indicate possession, belonging, or descriptive relationships. It frequently appears in noun phrases and compound expressions.	0.64	0.455	0.938	0.3	0.6	0.355	0.917	0.22
model.layers.0.mlp	Japanese	112017	The highlighted tokens are primarily nouns, noun compounds, and key morphemes that denote agents, roles, time, location, or abstract concepts central to the sentence's meaning. There is a strong emphasis on words that specify entities (such as organizations, people, or objects), temporal or locational references, and terms that mark focus, comparison, or categorization. These tokens often serve as anchors for the main informational content or structural roles within Japanese sentences.	0.54	0.641	0.526	0.82	0.63	0.634	0.627	0.64
model.layers.8.mlp	Japanese	32154	The highlighted tokens are primarily Bulgarian morphemes or word segments, especially those forming relative pronouns (such as \"която\", \"който\", \"което\", \"кои\"), as well as common suffixes and inflections (like \"ъ\", \"оя\", \"ой\", \"г\", \"точ\", \"еч\", \"еств\", \"ал\", \"ото\"). These segments often appear at the boundaries of words or as part of grammatical constructions, indicating a focus on morphological and syntactic markers in Bulgarian text.	0.73	0.64	0.96	0.48	0.71	0.603	0.957	0.44
model.layers.9.mlp	Japanese	34376	The highlighted tokens are primarily Japanese nouns, noun phrases, and compound words that denote concrete objects, locations, roles, or abstract concepts, as well as verbal nouns and set expressions. Many are used in formal, technical, or informational contexts, often marking key entities, actions, or conditions within sentences. There is a tendency to select tokens that encapsulate the main subject, object, or function in a clause, including terms for facilities, events, permissions, and quantifiable attributes.	0.8	0.796	0.812	0.78	0.81	0.8	0.844	0.76
model.layers.11.mlp	Japanese	29972	The highlighted Japanese text segments frequently correspond to noun phrases, verb phrases, or set expressions that convey key actions, states, or relationships within a sentence. These often include grammatical constructions for explanation, enumeration, or qualification, and commonly appear in contexts providing information, instructions, or descriptions.	0.62	0.406	0.929	0.26	0.6	0.412	0.778	0.28
model.layers.11.mlp	Japanese	126209	The highlighted tokens are primarily Japanese function words, particles, and common suffixes or inflections that mark grammatical relationships, connect clauses, or indicate possession, location, or action. These tokens are essential for sentence structure and meaning, often appearing at phrase or clause boundaries, and are frequently used in both formal and informal contexts.	0.62	0.441	0.833	0.3	0.79	0.769	0.854	0.7
model.layers.12.mlp	Japanese	25492	The highlighted tokens are primarily Japanese grammatical particles, auxiliary verbs, and function words that structure sentences, indicate relationships between clauses, and mark objects, subjects, or purposes. These elements are essential for conveying meaning, connecting ideas, and forming natural, coherent Japanese sentences.	0.59	0.305	1	0.18	0.68	0.529	1	0.36
model.layers.12.mlp	Japanese	121526	The highlighted tokens are primarily Japanese content words—nouns, verbs, and adjectives—often marking key entities, actions, or attributes within phrases or sentences. These tokens tend to be semantically rich, frequently appearing at or near phrase boundaries, and are often accompanied by grammatical particles or function words that clarify their syntactic role. The pattern emphasizes the importance of content-bearing morphemes in conveying the main informational structure of Japanese text.	0.69	0.575	0.913	0.42	0.69	0.597	0.852	0.46