Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.10.mlp	Korean	96918	The highlighted tokens are primarily Korean morphemes, word stems, and grammatical endings that are essential for constructing meaning in Korean sentences, including verbs, nouns, particles, and honorific or formal verb endings. These tokens often mark key semantic roles, actions, or attributes, and are frequently found at the end of words or phrases, reflecting the agglutinative structure of the Korean language.	0.70	0.57	1.00	0.40	0.78	0.72	1.00	0.56
model.layers.10.mlp	Korean	125019	The highlighted tokens consistently mark proper nouns, numerals, and key terms related to King Sejong, the Joseon Dynasty, and the invention of the Hangul alphabet, as well as their equivalents in multiple languages. Dates, names, and specific terminology are emphasized, often in the context of historical or factual statements.	0.52	0.08	1.00	0.04	0.51	0.04	1.00	0.02
model.layers.11.mlp	Korean	64252	The highlighted tokens are primarily Korean morphemes, words, or word endings that are semantically or grammatically significant, such as verbs, nouns, adjectives, and particles that convey core meaning, tense, or case. There is a strong focus on tokens that mark actions, states, or attributes, as well as those that indicate relationships or structure within sentences. High activations often correspond to key content words or inflectional endings that are essential for understanding the main information or function of the sentence.	0.60	0.33	1.00	0.20	0.68	0.54	0.95	0.38
model.layers.11.mlp	Korean	102511	The highlighted tokens are predominantly Korean noun and suffix forms, often ending with \"함\", \"것\", \"음\", \"법\", \"형\", \"율\", \"량\", \"군\", \"객\", \"면\", \"성\", \"전\", \"편\", \"심\", \"득\", \"상\", \"시간\", and similar morphemes. These tokens typically denote abstract concepts, objects, people, or states, and are frequently used as nominalizers or to form compound nouns, reflecting a pattern of marking key semantic units or grammatical roles in Korean text.	0.55	0.21	0.86	0.12	0.61	0.38	0.92	0.24
model.layers.12.mlp	Korean	27775	The highlighted tokens are primarily Korean morphemes, words, or short phrases that serve as key semantic units, often marking grammatical functions, core vocabulary, or important content within sentences. These include nouns, verbs, adjectives, and particles, as well as compound words and set expressions. The activations tend to focus on tokens that carry the main informational or structural load in each phrase, such as subjects, objects, predicates, and descriptive elements.	0.60	0.33	1.00	0.20	0.65	0.49	0.90	0.34
model.layers.12.mlp	Korean	67845	The text consistently centers on King Sejong and the Joseon Dynasty, the invention of the Hangeul (Hangul) alphabet, and the name \"Hunmin Jeongeum,\" with key tokens marking royal titles, dynastic references, invention dates (especially 1444 and 1418–1450), and the names of the alphabet and its inventor, across multiple languages.	0.53	0.11	1.00	0.06	0.53	0.11	1.00	0.06
model.layers.13.mlp	Korean	78512	The highlighted tokens are primarily Korean morphemes, words, or short phrases that serve as key semantic units within sentences, often marking grammatical roles, actions, or important nouns. Many activations correspond to travel, location, or procedural contexts, with emphasis on terms related to travel, reservation, information, and official processes. The pattern reflects a focus on core content words and functional morphemes that structure meaning in Korean informational or instructional text.	0.58	0.28	1.00	0.16	0.65	0.46	1.00	0.30
model.layers.14.mlp	Korean	87496	The highlighted tokens are primarily Korean morphemes, words, or short phrases that serve as key semantic units in sentences, often marking grammatical roles, objects, actions, or important nouns. Many activations correspond to noun or verb stems, particles, or endings that are essential for meaning in Korean syntax. Some tokens are also numbers, time expressions, or units, which are important for conveying quantitative or temporal information. Additionally, there are frequent occurrences of corrupted or garbled characters, likely due to encoding issues, which are also marked as important, possibly because they disrupt or alter the expected language pattern.	0.81	0.77	1.00	0.62	0.79	0.75	0.91	0.64
model.layers.14.mlp	Korean	107903	The highlighted tokens are suffixes, particles, or morphemes in Hindi and Korean that commonly appear at the end of words, often marking grammatical features such as plurality, case, or verb tense, or forming part of compound nouns and adjectives.	0.57	0.43	0.64	0.32	0.60	0.62	0.59	0.64