Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
model.layers.0.mlp	English	93066	Highly frequent activations on punctuation marks, especially commas and spaces, often at the start of sentences or clauses, and across multiple languages, indicating a focus on structural or delimiting tokens in multilingual text.	0.51	0.65	0.51	0.92	0.81	0.81	0.82	0.80
model.layers.0.mlp	English	128132	The highlighted tokens are primarily function words (such as articles, conjunctions, prepositions, and auxiliary verbs) and common noun phrases that serve as structural anchors in sentences. These tokens often mark the boundaries of clauses, introduce or connect ideas, and are essential for grammatical cohesion and meaning flow in both narrative and informational contexts.	0.51	0.65	0.51	0.92	0.75	0.76	0.73	0.80
model.layers.2.mlp	English	14207	The highlighted tokens are often function words, conjunctions, prepositions, or common short phrases that serve as syntactic connectors or modifiers, as well as inflectional or derivational morphemes in various languages. These tokens are crucial for sentence structure, meaning transitions, and grammatical relationships across diverse linguistic contexts.	0.50	0.67	0.50	1.00	0.49	0.64	0.50	0.92
model.layers.4.mlp	English	34138	The highlighted text segments are typically multi-word phrases or clauses that convey specific actions, states, or relationships, often providing key contextual or descriptive information within a sentence. These segments frequently include verbs, objects, and modifiers, forming meaningful units that contribute to the overall narrative or informational structure of the text.	0.55	0.55	0.55	0.56	0.60	0.33	1.00	0.20
model.layers.8.mlp	English	123973	The highlighted spans often represent multi-word phrases or clauses that function as cohesive semantic units, such as prepositional phrases, idiomatic expressions, or descriptive noun phrases. These segments frequently provide contextual detail, specify relationships, or add nuance to the main clause, and are typically composed of function words combined with content words to form meaningful, context-dependent chunks.	0.66	0.71	0.62	0.84	0.73	0.67	0.87	0.54
model.layers.15.mlp	English	5889	The highlighted tokens frequently correspond to structural elements in formal or factual writing, such as numbers, dates, conjunctions, prepositions, punctuation, and proper nouns, which are often used to organize, enumerate, or specify information in lists, references, or descriptive passages.	0.70	0.68	0.73	0.64	0.64	0.66	0.63	0.70
model.layers.15.mlp	English	67343	The text contains a wide variety of conversational and narrative fragments, including idiomatic expressions, filler words, discourse markers, and references to people, places, and actions. There is a frequent use of pronouns, conjunctions, and auxiliary verbs, as well as sequences that reflect spoken language patterns such as repetitions, hesitations, and clarifications. Many segments highlight the structure of dialogue, question-answer formats, and descriptive or explanatory statements, often focusing on actions, states, or relationships between entities.	0.72	0.73	0.71	0.74	0.57	0.47	0.61	0.38