Layer	Lang	Feature ID	Interpretation	Detection				Fuzzing			
				Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
model.layers.5.mlp	French	63694	The highlighted tokens are predominantly French (with some other languages), often marking inflectional or derivational morphemes such as verb endings, noun/adjective suffixes, and contractions, as well as common function words and parts of compound words, reflecting morphological and syntactic structure.	0.85	0.842	0.889	0.8	0.74	0.768	0.694	0.86
model.layers.8.mlp	French	86519	The highlighted tokens are primarily French verb forms, noun and adjective endings, and common morphemes that mark tense, aspect, plurality, or gender. These include verb conjugations (especially past participles and third-person forms), noun/adjective suffixes, and frequent function words or prefixes, reflecting the morphological structure and grammatical markers typical in French text.	0.86	0.841	0.974	0.74	0.9	0.891	0.976	0.82
model.layers.9.mlp	French	16701	The highlighted tokens are predominantly French morphological suffixes and inflections (such as verb endings, noun/adjective pluralizations, and gender/tense markers), as well as common function words and phrase boundaries. These elements are crucial for grammatical structure, word formation, and meaning in French text.	0.89	0.876	1	0.78	0.83	0.828	0.837	0.82
model.layers.9.mlp	French	44983	The highlighted tokens are predominantly French verb and noun stems, prefixes, and suffixes, often marking inflectional or derivational morphology, especially in verbs (e.g., marking tense, person, or participle forms), as well as common noun and adjective roots, and some frequent affixes or word fragments that contribute to word formation and grammatical structure.	0.65	0.533	0.8	0.4	0.7	0.659	0.763	0.58
model.layers.11.mlp	French	25732	The highlighted tokens are predominantly French nouns, adjectives, and verb forms, especially those with common French suffixes such as -tion, -ment, -age, -ance, -ité, -eur, -eux, -able, -aire, -ier, -eur, -ien, -ique, -eux, -er, -é, -es, -ant, -ent, -ons, -ais, -ais, -ée, -ées, -aux, -elle,	0.85	0.831	0.949	0.74	0.84	0.826	0.905	0.76
model.layers.11.mlp	French	110940	The highlighted tokens are predominantly French verb and noun stems, prefixes, and suffixes, often marking inflectional or derivational morphology (such as verb conjugations, participles, or noun/adjective forms). These segments frequently appear at morpheme boundaries, indicating their importance in identifying word structure and grammatical function in French text.	0.88	0.864	1	0.76	0.86	0.851	0.909	0.8
model.layers.12.mlp	French	22084	The highlighted tokens are predominantly French morphemes, pronouns, prepositions, verb endings, and noun/adjective suffixes, as well as common function words and inflections. These tokens are essential for grammatical structure, agreement, and meaning in French sentences, often marking tense, number, gender, or syntactic relationships.	0.82	0.78	1	0.64	0.8	0.756	0.969	0.62
model.layers.12.mlp	French	41725	The highlighted tokens are predominantly French suffixes and word endings, often marking grammatical features such as gender, number, tense, or part of speech, as well as common noun, adjective, and verb endings. These endings are crucial for morphological structure and meaning in French text.	0.71	0.603	0.957	0.44	0.83	0.809	0.923	0.72
model.layers.13.mlp	French	6455	Frequent use of French prepositions and conjunctions such as \"dans,\" \"de,\" \"sur,\" \"et,\" \"par,\" and suffixes like \"ant,\" \"ment,\" \"er,\" \"é,\" marking grammatical relationships, locations, and actions within sentences.	0.93	0.928	0.957	0.9	0.84	0.83	0.886	0.78
model.layers.13.mlp	French	77011	The highlighted tokens are primarily French conjunctions, pronouns, and discourse markers (such as \"que\", \"parce que\", \"si\", \"mais\", \"et\", \"donc\", \"alors\", \"ou\", \"comme\", \"quand\", \"où\", \"dont\", \"que\", \"qu'\", \"si\", \"mais\", \"alors\", \"donc\", \"bien\", \"fait\", \"vois\", \"wow\", \"Non\", \"Oh\", \"accord\") as well as punctuation and suffixes. These elements are essential for structuring sentences, expressing relationships between clauses, and managing the flow of conversation in French text.	0.81	0.796	0.86	0.74	0.85	0.828	0.973	0.72
model.layers.14.mlp	French	45764	The highlighted tokens are primarily French morphemes, function words, and noun or adjective endings, including plural and gendered suffixes, as well as common prepositions, articles, and pronouns. These tokens often mark grammatical structure, agreement, and relationships between words, and are frequently found at the ends of words or as connectors within phrases.	0.79	0.734	1	0.58	0.85	0.828	0.973	0.72