layer	lang	feature_id	interpretation	tokens
model.layers.0.mlp	Bulgarian	123218	"The text highlights the frequent use of single Cyrillic letters, especially at the beginning of sentences or phrases, functioning as prepositions, conjunctions, or initials in Russian and Bulgarian, often with high activation when capitalized and sentence-initial."	О, га, на, Г
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."	(, ;, *, , ., , in,), und, , (
model.layers.0.mlp	Turkish	64723	"Suffixes such as \"ı\", \"lar\", \"ler\", \"lık\", \"lık\", \"ım\", \"in\", \"isi\", \"lığı\", \"ları\", \"leri\", \"ını\", \"udur\", \"ildi\", \"ım\", \"isi\", \"arı\", \"alı\", \"isi\", \"lık\"	ир
model.layers.1.mlp	Bulgarian	32578	"The highlighted tokens are common Bulgarian suffixes used to form adjectives, nouns, and participles, indicating grammatical features such as gender, number, definiteness, and case. These suffixes are essential for word formation and inflection in Bulgarian morphology."	па, на
model.layers.1.mlp	Turkish	84692	"The highlighted tokens are predominantly prefixes, roots, or short morphemes within words, often marking the beginning or core of nouns, verbs, or adjectives across multiple languages. These segments frequently correspond to meaningful subword units, such as roots, affixes, or syllables, and are often found in proper names, technical terms, or compound words. The pattern reflects a focus on linguistically significant subword structures that contribute to word formation and meaning."	Gir
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."	<|begin_of_text|>,
model.layers.2.mlp	German	56064	"The highlighted tokens are primarily German morphemes, word stems, or inflections, often marking noun or verb endings, compound word parts, or grammatical function words. Many are common in German-language text, including noun endings (-en, -heit, -keit, -ung), verb forms (-iert, -te, -en), and connecting words (und, von, in, der, die, das, eine). There is also frequent activation on recurring roots in compound nouns and on inflectional suffixes, reflecting the morphological structure of German."	isch, Juni, Lang, ist, ß
model.layers.2.mlp	Thai	100192	"The highlighted tokens are often function words, affixes, or short morphemes in various languages, including prepositions, conjunctions, pronouns, and grammatical particles, as well as some high-activation non-standard or corrupted characters. These tokens are typically important for sentence structure, meaning, or language identification."	Ol, О
model.layers.2.mlp	Thai	105874	"The highlighted tokens are often morphemes, syllables, or short word fragments that serve as meaningful units in various languages, including Thai, Hindi, and others. These units frequently appear at the beginning, middle, or end of words and are important for word formation, inflection, or conveying grammatical and semantic information."	Ol, О
model.layers.3.mlp	Bulgarian	98476	"The highlighted tokens are predominantly Bulgarian suffixes, pronoun and conjunction forms, and inflectional endings that mark grammatical relationships such as gender, number, case, and verb tense, as well as function words that connect clauses or indicate possession and agency. These elements are essential for the syntactic structure and meaning in Bulgarian sentences."	Alexand
model.layers.3.mlp	Italian	120200	"The highlighted tokens are often morphemes, suffixes, or name fragments within words, especially in Italian and other Romance languages, marking comparative forms, diminutives, or parts of proper nouns and loanwords. These tokens frequently appear at word endings or within multi-syllabic names and terms."	О
model.layers.3.mlp	Spanish	4518	"The highlighted tokens are often morphemes, word stems, or affixes from various languages, especially Spanish, marking grammatical or semantic units such as verb endings, noun forms, or prepositions, and sometimes appear in named entities or set phrases."	ga
model.layers.4.mlp	Russian	48343	"The highlighted tokens are primarily short morphemes, prefixes, or function words in Slavic and related languages, often marking grammatical relationships, word formation, or serving as connectors within or between words and phrases."	Gir