Feature 125019

Language: Korean
Model: meta-llama/Llama-3.2-1B
Layer: model.layers.10.mlp
SAE Model: EleutherAI/sae-Llama-3.2-1B-131k
Selected Token Probability: 0.266
Entropy: 0.155

Activation Range

0.082-0.209 | 0.209-0.335 | 0.335-0.462 | 0.462-0.589 | 0.589-0.715 | 0.715-0.842 | 0.842-0.969 | 0.969-1.096 | 1.096-1.222 | 1.222-1.349

Interpretation

"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."

Score Type	Accuracy	Precision	Recall	F1 score	TPR	TNR	FPR	FNR
detection	0.52	1.0	0.04	0.077	0.04	1.0	0.0	0.96
fuzz	0.51	1.0	0.02	0.039	0.02	1.0	0.0	0.98

Korean

#examples: [('paws-x', 997), ('flores', 997)]

paws-x-162. <|begin_of_text|>연락 원은 58 ��이었지만이 중 20 ��은 메신저를 위해서만 사용되��습니다.

paws-x-979. <|begin_of_text|>쇼는 MBC 2에서 큰 상금 $ 100 000으로 방송되었지만 취소로 인해 이겼습니다.

paws-x-828. <|begin_of_text|>Holy Trinity (가��릭 리그) (1443 W. Division St) 또한 카��릭 리그 팀이��습니다. 마지막 호랑이 축구 시즌은 1965 년이 ��습니다.

paws-x-208. <|begin_of_text|>Togdheer (소말리 'Wabi Togdheer')는 소말리랜드 동부의 Togdheer River 지역의 계절 강입니다.

Text Examples for Each Interval

interval 1

Range: 1.222-1.349
#examples: 4

paws-x-162. <|begin_of_text|>연락 원은 58 ��이었지만이 중 20 ��은 메신저를 위해서만 사용되��습니다.

xnli-463. <|begin_of_text|>Một khu phố châu á thứ ba , Hàn Quốc , nằm phía tây của trung tâm dọc theo đại lộ olympic giữa vermont và Phương Tây . Đại lộ Olympic là phía đông của khu vực Hàn Quốc , và là vermont và Phương Tây .

paws-x-979. <|begin_of_text|>Die Sendung wurde auf MBC 2 mit einem großen Preis von 100.000 $ ausgestrahlt, die aufgrund der Absage nicht gewonnen wurde.

flores-385. <|begin_of_text|>Hangeul is the only purposely invented alphabet in popular daily use. The alphabet was invented in 1444 during the reign of King Sejong (1418 – 1450).

interval 2

Range: 1.096-1.222
#examples: 8

paws-x-979. <|begin_of_text|>The show was broadcast on MBC 2 with a big prize $ 100 000 , not won because of the cancellation .

paws-x-979. <|begin_of_text|>쇼는 MBC 2에서 큰 상금 $ 100 000으로 방송되었지만 취소로 인해 이겼습니다.

paws-x-828. <|begin_of_text|>Holy Trinity (가��릭 리그) (1443 W. Division St) 또한 카��릭 리그 팀이��습니다. 마지막 호랑이 축구 시즌은 1965 년이 ��습니다.

flores-385. <|begin_of_text|>Hangeul là bảng chữ cái được phát minh chỉ nhằm mục đích sử dụng thông dụng hàng ngày. Bảng chữ cái được phát minh vào năm 1444 trong triều đại Vua Sejong (1418 - 1450)