Feature 8775

Language: Thai
Model: meta-llama/Llama-3.2-1B
Layer: model.layers.12.mlp
SAE Model: EleutherAI/sae-Llama-3.2-1B-131k
Selected Token Probability: 0.253
Entropy: 1.799

Activation Range

0.076-0.347 0.347-0.618 0.618-0.889 0.889-1.16 1.16-1.431 1.431-1.702 1.702-1.973 1.973-2.244 2.244-2.515 2.515-2.786

Interpretation

"The highlighted tokens are often found in polite, formal, or explanatory constructions in Japanese and Korean, including suggestions, indirect statements, and expressions of possibility or uncertainty. These tokens frequently appear in verb endings, auxiliary forms, and set phrases that convey nuance, deference, or hypothetical meaning."

Score Type	Accuracy	Precision	Recall	F1 score	TPR	TNR	FPR	FNR
detection	0.58	0.722	0.26	0.382	0.26	0.9	0.1	0.74
fuzz	0.77	0.886	0.62	0.729	0.62	0.92	0.08	0.38

Thai

#examples: [('xnli', 1000), ('flores', 996)]

xnli-282. <|begin_of_text|>ใช่ เอ่อ เรา ไม่ได้ เป็น ทหารเรือ จริงๆ แต่ ฉัน ต้องการ ��ดู ร้อนใน เร็ว ๆ นี้ เพื่อ อืม จ้าง เรือใบ กับ กัปตัน เนื่องจาก เรา ไม่ได้ รู้ ว่า มีความรู้ ใน สิ่ง นั้น และ เอ่อ ไป ที่ แคริบเบียน เขา ได้ ล่องเรือ ไป นิดนึง

xnli-326. <|begin_of_text|>เอ่อ ไม่ ฉัน ไม่ได้ ฉัน ตัดสินใจ ที่จะ ทำ มัน และ ฉัน ก็ ทำ

xnli-300. <|begin_of_text|>คุณ โกรธ แล้ว นะ คุณ ยัง สติ ดี อยู่

xnli-743. <|begin_of_text|>ใช่ มัน เป็น แบบ นั้น จริงๆ และ ทาง นั้น เรา ก็ ไม่ได้ พลาด อะไร ไป เลย ที่ คุณ รู้ ว่า เด็ก พวก นั้น เอ่อ เรา ไม่ได้ หาย ไป จาก ประโยชน์ ของ การ ดูแล เด็ก จริงๆ

Text Examples for Each Interval

interval 1

Range: 2.515-2.786
#examples: 1

flores-873. <|begin_of_text|>もっと冒険したいと思ったら、この機会にスムージーを��るかブレンドしてみてはいかがでしょうか。

interval 2

Range: 2.244-2.515
#examples: 1

flores-868. <|begin_of_text|>特に、最初の数日間は、��れるために、欧風の高級なホテルや食事、サービスにお金をかけることを検討してみてはいかがでしょうか。

interval 3

Range: 1.973-2.244
#examples: 15

flores-591. <|begin_of_text|>もしかしたら、あなたのひ��たちが異星人の惑星に降り立って、古代の祖先に思いを��せる日が来るかもしれません。

flores-235. <|begin_of_text|>"아버지가 무슨 말을 했냐고 ��자 그녀는 ""아무 말도 할 수가 없어서 그저 눈만 ����이며 서 계��어""라고 말했습니다."

flores-245. <|begin_of_text|>応援してくださる方々がいて幸せです。

flores-244. <|begin_of_text|>異を唱える人もいるかもしれないが、私は気にしません。

flores-406. <|begin_of_text|>そうです！ツタンカー��ン王は、「少年王」と呼ばれることもあり、現代では最も有名な古代エジプトの王の1人です。

paws-x-769. <|begin_of_text|>林氏は、マッキーは「テンチのオリジナルモデルだ」と述べた。