FEATURE 176433 (LAYER 24)

Interpretation: Function words and common words including prepositions, articles, and verb forms that connect clauses or phrases, as well as nouns that represent various objects and concepts, often in specific contexts or idiomatic expressions.

Top Examples:

1. Text: forgo insurance. Ultimately, that choice is up to you. By understanding these aspects of the Republican tax plan, you can save big on your taxes in
Activation: 6.9062
Active tokens: taxes

2. Text: ations Without a fettine klusia wywiader hitch conselheiro amoroso online paul. In France, Germany, Belgium, Luxem
Activation: 6.8438
Active tokens: wi

3. Text: K-ras oncogene and also via mutations in BRAF. Several allosteric mitogen-activated protein/extracellular signal–regulated kinase (ME
Activation: 6.5000
Active tokens: rac

FEATURE 203901 (LAYER 20)

Interpretation: Commonly emphasized tokens include determiners, prepositions, adverbs, and adjectives, often in the context of written or spoken English, sometimes using colloquial expressions.

Top Examples:

1. Text: was an avid reader and a fantastic cook. Susan was a brave and courageous woman who battled MS for over 40 years. Even given the limitations of her
Activation: 9.1875
Active tokens: given

2. Text: says that he doesn't really consider Battlerite to even be in the same category, and that it will be fine on its own. Well I
Activation: 9.1250
Active tokens: to

3. Text: to see a dime of the funds. The transaction occurred mere hours before the doomed exchange stopped honoring withdrawals. Tsao sold nearly 20 bit
Activation: 9.1250
Active tokens: .

K Compute Details

We performed all experiments with a single NVIDIA A100 80GB GPU in a private cluster. Table 1 lists the approximate duration of the final experiments for each model size and transformer variant. We estimate that the total cost of preliminary and failed experiments is roughly equal to the cost of the final experiments.

Model	Variants	Approx. time per variant (hours)	Total time (hours)
Pythia-6.9b	5	70	350
Pythia-1b	5	10	50
Pythia-410m	5	5	25
Pythia-160m	5	1	5
Pythia-70m	5	1	5
		Overall time:	435

Table 1: Approximate time required for our experiments.

L Example feature dashboards for Pythia-6.9b

Here we provide more detailed 'feature dashboards,' including per-feature activation patterns, token distribution entropy, and auto-interpretability ('fuzz' ROC) scores. We include two randomly sampled features for the control, randomized, and trained variants described in Section 3, trained on every fourth layer of Pythia-6.9b.

L.1 TRAINED

Feature 10065 (Layer 0) - Trained

Entropy: 0.009 | Fuzz ROC: 0.961

Interpretation: Prefixes or words starting with "Re" often indicating repetition, return, or renewal.

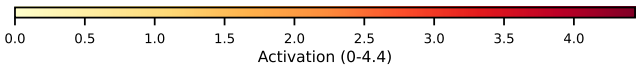
Ex 1 which Judge Kreep was censured, Brower says he has worked in diverse work places in the military and D.A.'s office, also working with

Ex 2 Example: In Louros v. Kreikas, 367 F. Supp. 2d 572 (S.D.N.Y. 2005), the

Ex 3 the Alhambra's arches. Over the years, Kreber has supplied the color separations, while printing services were provided by Century Graphics,

Ex 4 treated equal. What was the Statue of Liberty originally used for? Shadows over Kregen Schatten über Kregen, 1996; English ebook edition

Ex 5 by Boston attorney Arthur Kreiger, who represented AT&T. Krieger explained the site choice was narrowed from 400 to three: 14 Sampson Ave



Feature 10222 (Layer 0) - Trained

Entropy: 0.050 | Fuzz ROC: 0.970

Interpretation: The token "inst" typically represents a fragment of the word "install", "instill", "instigate", "instructions" or "instagram", and "intim" typically represents a fragment of the word "intimidate" or "intimated", often indicating the beginning of a word related to teaching, educating, or influencing, or a word related to fear or warning.

Ex 1 <|endoftext|>intim villa. Dua kamar tidur yang memiliki akses langsung ke kolam renang. Di setiap kamar

Ex 2 the non-Muslim world more and more fold under their legal intimidation as a result of our pacifism, self-hatred and complacency.

Ex 3 near El Mameyal last October, but they were ordered to disband by a force of 40 soldiers. The campaign of intimidation may have worked

Ex 4 least in part to the attempt by the Railroad Commission of Arkansas to protect Arkansas shippers and build up Arkansas jobbing centers.' In that case it was intimated

Ex 5 than half of those against people were assault cases, while nearly 45 per cent were crimes of intimidation. 'No person should have to fear being violently attacked