(interp) [cscandelius@holygpu8a19304 July]$ python run\_analysis.py --script last\_token --use\_system\_prompt

Running LastToken analysis...

USE\_SYSTEM\_PROMPT\_FOR\_MANIFOLD set to: True

PERTURB\_ONCE set to: False

Configuration: PERTURB\_ONCE is set to False

Configuration: USE\_SYSTEM\_PROMPT\_FOR\_MANIFOLD is set to True

Loading checkpoint shards: 100%|██████████████████████████████| 4/4 [00:09<00:00, 2.48s/it]

################################################################################

### STARTING ANALYSIS FOR LAYER 0 ###

################################################################################

Extracting activations from layer 0...

Extracting activations: 100%|███████████████████████████| 1944/1944 [00:43<00:00, 44.67it/s]

Extracting activations from layer 0...

Extracting activations: 100%|█████████████████████████████| 247/247 [00:05<00:00, 45.17it/s]

Concept 'dog': Found 132 effective eigenvectors out of 1944 (threshold: 0.0000)

Concept 'lion': Found 33 effective eigenvectors out of 247 (threshold: 0.0000)

================================================================================

Running experiments for system prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User prompt: 'The dog was running around the park. It was a labrador.'

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0, AXIS: 0 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC0 ---

Perturbation scale -100.0x: DieGermanassistantTheassistantassistantTheassistantassistantassistantassistantassistantThe translation translationThe translation

The text translation

The translation

The dog was running around the park.

Perturbation scale -20.0x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet.

- "was running" wird mit "lief" übersetzt, da "

Perturbation scale -10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

Perturbation scale -5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

-

Perturbation scale +10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

-

Perturbation scale +20.0x: The translation of the given text from English into German is:

Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale +100.0x: Die Hünd war um den Park herum. Es war ein Labrador.

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0, AXIS: 1 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC1 ---

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -20.0x: ://://://://://://://://://://://://://://://:// the.:// the://://assistant

://assistant

Perturbation scale -10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

-

Perturbation scale +10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

-

Perturbation scale +20.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +100.0x: Die Hünde war im Park herumren. Es war ein Labrador.

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0, AXIS: 2 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC2 ---

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -20.0x: Die Hündin lief um den Park herum. Sie war eine Labradorhündin.

Hinweis:

- "The dog" wird in der Übersetzung mit "Die Hündin" wiedergegeben, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männ

Perturbation scale -10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +10.0x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale +20.0x: Der Hund lief im Park herum. Es war ein Labrador.

Perturbation scale +100.0x: Die Hündin ran um den Park herum. Es war ein Labrador.

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0, AXIS: 3 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC3 ---

Perturbation scale -100.0x: Die Hünd war im Park herum. Es war ein Labrador.

Here is the correct translation:

Die Hünd war im Park herum.

Es war ein Labrador.

Perturbation scale -20.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

Perturbation scale -10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

-

Perturbation scale +5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

-

Perturbation scale +10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

-

Perturbation scale +20.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

-

Perturbation scale +100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0, AXIS: 4 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC4 ---

Perturbation scale -100.0x: Die Hünd lief um den Park herum. Es war ein Labrador.

Perturbation scale -20.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wurde mit "Die Hündin" übersetzt, da "die Hündin" in der deutschen Sprache weiblich ist. Wenn Sie "The dog" als männlich meinen, sollten Sie "Der Hund

Perturbation scale -10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

-

Perturbation scale -5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" weiblich ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

- "was

Perturbation scale +20.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" weiblich ist. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

- "was

Perturbation scale +100.0x: Die Hündin war im Parken herummt. Sie war ein Labrador.

Here's a breakdown of the translation:

- "Die" is a more natural translation for "the dog".

- "war" is a more natural translation for "running".

- "im" is a more natural translation for "around".

- "her" is

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC0 'dog' direction (Layer 0) ---

Projection values:

Top positive #1: 0.1017

Top positive #2: 0.0918

Top positive #3: 0.0853

Top negative #1: -0.0865

Top negative #2: -0.0859

Top negative #3: -0.0779

Top 10 prompts most aligned with POSITIVE PC0 direction:

1. 'We adopted a senior dog from the shelter, and he is the sweetest, most gentle companion.'

2. 'There, in the center of the mess he created, sat the unrepentant dog, wagging his tail.'

3. 'The dog's joy is simple, pure, and infectious.'

4. 'The dog's life is a beautiful, albeit brief, journey.'

5. 'She loves to ride in the car, no matter the destination.'

6. 'My dog's life revolves around a simple and beautiful schedule: eat, play, sleep, love, repeat.'

7. 'She loves the beach, chasing the waves back and forth.'

8. 'She is the perfect hiking companion.'

9. 'He is not a fan of the mailman, the delivery driver, or squirrels.'

10. 'Frustrated, the dog let out a huff and dropped the toy, waiting for me to throw it again.'

Top 10 prompts most aligned with NEGATIVE PC0 direction:

1. 'That dog is a furry alarm clock with a wet nose and no snooze button.'

2. 'The celebrity's tiny teacup dog had its own social media account with millions of followers.'

3. 'Why does one dog in a neighborhood start a barking chain reaction with every other dog?'

4. 'Why does every dog feel the need to circle three times before finally lying down?'

5. 'The first-ever domesticated dog is believed to have descended from an ancient wolf population.'

6. 'My dog always brings me a 'gift'—usually a soggy toy or a stolen sock—when I get home.'

7. 'The guide dog successfully navigated its owner across a busy intersection during rush hour.'

8. 'My neighbor's golden retriever dog is incredibly well-trained and never jumps on guests.'

9. 'The new city ordinance requires every dog owner to carry proof of rabies vaccination.'

10. 'The mushers prepared their sled dog teams for the grueling thousand-mile Iditarod race.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC1 'dog' direction (Layer 0) ---

Projection values:

Top positive #1: 0.0662

Top positive #2: 0.0646

Top positive #3: 0.0634

Top negative #1: -0.0729

Top negative #2: -0.0627

Top negative #3: -0.0591

Top 10 prompts most aligned with POSITIVE PC1 direction:

1. 'That dog is incredibly intuitive and always seems to know when someone needs comforting.'

2. 'Her boundless love is a gift I don't deserve but cherish.'

3. 'After a long day of hiking, the tired dog fell asleep before he even finished his dinner.'

4. 'She can tell when I'm just getting my shoes on versus when I'm getting her leash.'

5. 'My dog has a very clear 'I'm sorry' face he makes after he knows he has done something wrong.'

6. 'The old dog moved stiffly in the morning but seemed to loosen up after a short walk.'

7. 'Her playful nature is a daily reminder not to take life too seriously.'

8. 'The dog's memory is quite good; he always remembers the people who give him the best treats.'

9. 'I wish my dog could talk so he could tell me what he's thinking about all day.'

10. 'My dog's dramatic sigh seemed to express his profound disappointment with the lack of walks today.'

Top 10 prompts most aligned with NEGATIVE PC1 direction:

1. 'The Schnoodle is a Schnauzer-Poodle mix.'

2. 'The Pomapoo is a Pomeranian and Poodle mix.'

3. 'The Cocker Spaniel has long, floppy ears.'

4. 'The Labradoodle combines the traits of a Labrador and a Poodle.'

5. 'The Pomsky is a cross between a Pomeranian and a Husky.'

6. 'The Goldendoodle is a cross between a Golden Retriever and a Poodle.'

7. 'The Maltipoo is a mix of a Maltese and a Poodle.'

8. 'The Newfypoo is a cross between a Newfoundland and a Poodle.'

9. 'The Sheepadoodle is an Old English Sheepdog and Poodle mix.'

10. 'The dog's friendship is a bond that transcends words.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC2 'dog' direction (Layer 0) ---

Projection values:

Top positive #1: 0.0543

Top positive #2: 0.0524

Top positive #3: 0.0515

Top negative #1: -0.0566

Top negative #2: -0.0537

Top negative #3: -0.0521

Top 10 prompts most aligned with POSITIVE PC2 direction:

1. 'Come!'

2. 'Fetch!'

3. 'Sit!'

4. 'Off!'

5. 'Drop it!'

6. 'Stay.'

7. 'Do dogs dream?'

8. 'Heel.'

9. 'Is it true that some breeds are hypoallergenic?'

10. 'How long do dogs live?'

Top 10 prompts most aligned with NEGATIVE PC2 direction:

1. 'The dog, covered from nose to tail in mud, was not allowed back in the house until he had a thorough bath.'

2. 'The dog's coat was a beautiful brindle pattern, a mix of black, brown, and gold stripes.'

3. 'The dog's ears perked up, and he let out a low growl, alerting us to someone approaching the door.'

4. 'There, in the center of the mess he created, sat the unrepentant dog, wagging his tail.'

5. 'The dog, a blur of fur and motion, joyfully chased the ball down the long hallway.'

6. 'The sheepdog, with its intense gaze, kept the flock tightly packed and moving in the right direction.'

7. 'The dog's deep sigh, heard from the other room, signaled his boredom with our lack of activity.'

8. 'Through the pouring rain, the lost dog searched for a dry place to find shelter for the night.'

9. 'The dog's intense focus on the ball, to the exclusion of all else, is a form of meditation.'

10. 'The dog's playful bow, with his front end down and his back end up, was an invitation to chase.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC3 'dog' direction (Layer 0) ---

Projection values:

Top positive #1: 0.0646

Top positive #2: 0.0598

Top positive #3: 0.0587

Top negative #1: -0.0449

Top negative #2: -0.0415

Top negative #3: -0.0408

Top 10 prompts most aligned with POSITIVE PC3 direction:

1. 'If a dog could have a job, what would it be?'

2. 'If a dog could give advice to humans, what would it say?'

3. 'If a dog could make a wish, what would it be?'

4. 'If dogs could vote, what issues would matter to them?'

5. 'If a dog could send a text, what would it say?'

6. 'If a dog could write a review of its owner, what would it say?'

7. 'If a dog could write a thank-you note, what would it say?'

8. 'If a dog could be any other animal for a day, what would it choose?'

9. 'If a dog could write a thank-you note to its owner, what would it say?'

10. 'If dogs could run a business, what would it be?'

Top 10 prompts most aligned with NEGATIVE PC3 direction:

1. 'The Siberian Indian Dog is a Husky and Native American Indian Dog cross.'

2. 'The mushers prepared their sled dog teams for the grueling thousand-mile Iditarod race.'

3. 'The Gerberian Shepsky is a German Shepherd and Siberian Husky mix.'

4. 'The Alaskan Malamute was bred for hauling heavy freight.'

5. 'The Welsh Springer Spaniel is a distinct breed from the more common English Springer Spaniel.'

6. 'The Neapolitan Mastiff is known for its massive size and loose, wrinkled skin.'

7. 'The pet adoption fee covers spaying/neutering and initial vaccinations.'

8. 'During the thunderstorm, the terrified dog hid under the kitchen table, shivering uncontrollably.'

9. 'The sled dog team worked in perfect harmony, pulling the heavy sled across the frozen tundra.'

10. 'The boarding facility offers luxury suites for pampered pooches.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC4 'dog' direction (Layer 0) ---

Projection values:

Top positive #1: 0.0426

Top positive #2: 0.0417

Top positive #3: 0.0411

Top negative #1: -0.0461

Top negative #2: -0.0392

Top negative #3: -0.0336

Top 10 prompts most aligned with POSITIVE PC4 direction:

1. 'He's a very easy-going dog who is happy just to be with his people.'

2. 'That dog is a true original, a one-of-a-kind canine with a personality all his own.'

3. 'A dog's nose print is as unique as a human's fingerprint, and can be used for identification.'

4. 'Is it true that a dog's sense of smell is thousands of times better than a human's?'

5. 'The bond between a human and their dog is often profound.'

6. 'Her intelligence is sometimes a little too much for her own good.'

7. 'The howl of a wolf is not so different from that of some domestic breeds.'

8. 'The dog's sense of hearing is far more sensitive than a human's.'

9. 'He is a very magnificent and powerful animal.'

10. 'He is a very powerful and magnificent-looking animal.'

Top 10 prompts most aligned with NEGATIVE PC4 direction:

1. 'During the thunderstorm, the terrified dog hid under the kitchen table, shivering uncontrollably.'

2. 'Underneath the porch, the timid rescue dog watched the world with wary, uncertain eyes.'

3. 'His tail started wagging the moment I picked up the leash.'

4. 'She runs in her sleep, probably chasing dream squirrels.'

5. 'The mushers prepared their sled dog teams for the grueling thousand-mile Iditarod race.'

6. 'The sled dog team worked in perfect harmony, pulling the heavy sled across the frozen tundra.'

7. 'There, in the center of the mess he created, sat the unrepentant dog, wagging his tail.'

8. 'When I came home, the guilty-looking dog had spread trash all over the kitchen floor.'

9. 'The doggie daycare sends report cards home.'

10. 'How do dogs feel about mail carriers?'

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0, PROJECTION ONTO PC0 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: TheTheTheTheThe

assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale -20.0x:

Der Der Der Der Der Der Der Der Der Der Der Der Der Der Das Das Das Der Der Das Das Der Der Der Die Die Das

Perturbation scale -10.0x: DieDieDieDie Die Der Der Der Der Der Der Der Der Der Der Der Der Der Der Der Der

Perturbation scale -5.0x: Der

Der

Perturbation scale -2.5x: Die

Die Der Die Die

Perturbation scale -1.5x:

Die Die

Perturbation scale +0.0x:

Der Die Der Der Der

DerDerDer

DerDerDasDerDerDerDer Der Der Der Der DerDerDerDerDerDer

Perturbation scale +1.5x: assistant

Perturbation scale +2.5x:

:// Der Der

Perturbation scale +5.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantDer Die DasDieDieDieassistantassistantDieassistantDerassistantDer

Perturbation scale +10.0x: ://://://://://://://://://://://://://://://://://://://://:// theassistant://://://://.://://:// the

:// theassistant://://://:// theassistant://://://.://:// (://

Perturbation scale +20.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale +100.0x: der ist ist

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0, PROJECTION ONTO PC1 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://ста://://://://://ста://://://://://://://://://://://://://://://://://ста://://://ста

Perturbation scale -20.0x: ://://://://://://://://://://://://ста

Perturbation scale -10.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -5.0x: ://://://://://://://://://://://://://://:// the://:// the://://

://assistant the://://://assistantassistant the://://://assistantassistantassistant the://://://assistant://://.

://..log

Perturbation scale -2.5x: ://://assistant://://://://://://://://://assistantassistantassistant://://assistant

://://://://

://

://

://

://

://

://

Perturbation scale -1.5x: The://://://://://://assistantassistantassistant://://

://

://

Perturbation scale +0.0x: ://assistantassistantassistantassistant://://assistant

://

://

Perturbation scale +1.5x: ://assistantassistant

://

assistant

://

Perturbation scale +2.5x: ://assistantassistant

://Dieassistant

assistant

assistantassistant

assistantassistant

://

assistant

Perturbation scale +5.0x: assistantassistant

Die

://

://

assistant

://

://

Perturbation scale +10.0x:

Perturbation scale +20.0x:

Der

Perturbation scale +100.0x:

DieDerDerDerDerDerDerDerDerDerDerDer

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0, PROJECTION ONTO PC2 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -20.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -10.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -5.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -2.5x: DieassistantDieThe://:// theassistant

assistantassistant://://://://.assistantassistant://://://://://://://://://://://://://://://://://://://://://://://://://://://.log

Perturbation scale -1.5x: assistantassistantTheThe

assistantassistant theassistantassistant://:// theassistant

Theassistantassistant://://.assistantassistant://://.assistantassistant://.assistantassistant://.assistantassistant://.assistantassistant.assistantassistant

Perturbation scale +0.0x: assistantassistantassistant202TheassistantassistantTheassistantassistantTheTheDieassistant

Perturbation scale +1.5x: assistantassistantassistant202assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +2.5x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant dassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

assistantassistantassistantassistant

Perturbation scale +5.0x: assistantassistant

assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant   assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantDerDerDerassistantDerDerDerassistantassistant

Perturbation scale +10.0x:

assistantutschassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantDerassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +20.0x:

Der

assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +100.0x: DerDer Der Der Der Der Der Der Der der der der der der der das das das das das das das das

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0, PROJECTION ONTO PC3 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x:

assistant

Perturbation scale -20.0x:

Perturbation scale -10.0x:

.

Die

DerDerDerDerDer

Perturbation scale -5.0x: assistantassistantassistantassistant

Perturbation scale -2.5x: assistantassistantassistant

://

Perturbation scale -1.5x: Dieassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +0.0x: ://assistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +1.5x: ://Dieassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +2.5x: :// the:// the://assistant the theassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +5.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://.://

Perturbation scale +10.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale +20.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale +100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0, PROJECTION ONTO PC4 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x:

assistantassistant

Perturbation scale -20.0x: assistantassistant

{...assistantassistantassistant

Perturbation scale -10.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale -5.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale -2.5x: assistantassistantassistantassistantassistantassistantassistant

Perturbation scale -1.5x: Dieassistantassistantassistantassistant

Perturbation scale +0.0x: Derassistant

Perturbation scale +1.5x: TheDieassistant

Perturbation scale +2.5x: TheDerHier

Perturbation scale +5.0x: theassistant

the the the the the the theassistant

Perturbation scale +10.0x: :// the the:// the the:// the the the the the the the the the the the the the the the the the the the the theassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +20.0x: ://://://://://://://://://:// the the the the the the the the the the the the the the the the theassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +100.0x:

://

:// the the the the the the the the the the the the the the the the the the the the the the the the

================================================================================

================================================================================

--- ORTHOGONAL PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0 ---

================================================================================

Perturbing along first orthogonal direction (PC132)...

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along orthogonal direction ---

Perturbation scale -100.0x: Die Überschrift:

Der Hund Die Hund Die Hund Die H Dog Labrador Hund Die H ist: Der Hund Die H ist: Der Hund die H ist: Der Hund: Die Histo: Der Hund: Die Histo: Der Hund: Die Histo: Der Hund: Die Histo: Der Hund: Die Histo: Der Hund:

Perturbation scale -20.0x: Die Hündin lief um den Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund bezeichnet.

- "was running" wird mit "lief" übersetzt, da "l

Perturbation scale -10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "die Hündin" ein weiblicher Hund bezeichnet.

- "was running" wird mit "lief" übersetzt, da "

Perturbation scale -5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn Sie "The dog" als männlichen Hund meinen, sollten Sie "Der Hund" oder "

Perturbation scale -2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

Perturbation scale +5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

Perturbation scale +10.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn es sich um einen männlichen Hund handelt, würde "Der Hund" verwendet werden.

Perturbation scale +20.0x: Der Hund lief im Park herum. Es war ein Labrador.

Perturbation scale +100.0x: Der Hund lief um die Park herum.

Die Hunde

Die Hund

Die Hund

Die Hund

Die hund

Die hund

Die hund

Die hund

Die hund

Die hund

Die hund

Die hund

Die hund

Die hund

Die hund

Die h

================================================================================

================================================================================

--- ABLATION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

All PCs (centroid only):

DerDer

All PCs except largest (PC0 only):

Der

All PCs except largest two (PC0+PC1):

================================================================================

================================================================================

Running experiments for system prompt: 'You are a helpful assistant.'

User prompt: 'Please describe two different dog breeds.'

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0, AXIS: 0 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC0 ---

Perturbation scale -100.0x: LetTheTheTheDescriptionTheBTheBTheBTheBTheBTheBTheB

TheBreedTheBreed

The description ofThe description of the dog breed

The description of the dog breed

The description of the dog breed

The description of the dog breed

The description of the dog breed

The description of the dog

Perturbation scale -20.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game such as rabbits and hares. Greyhounds are typically tall and lean, with a short, smooth

Perturbation scale -10.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically tall and lean, with a short

Perturbation scale -5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +10.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are characterized by their:

\* Short, smooth coats that

Perturbation scale +20.0x: The first dog breed I'd like to describe is the Poodle.

The Poodle is a highly intelligent and versatile dog breed known for its distinctive appearance and low-shedding coat. They come in three main sizes: Toy, Miniature, and Standard.

The Toy Poodle is the smallest of the three, weighing around 6-

Perturbation scale +100.0x: Let's take a look at look

The the

The the the the the the the the the the the the the the the the the the the the the the help the the the the the the the the help the help the the the the the the the the the the the the the the the ( the

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0, AXIS: 1 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC1 ---

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -20.0x: ://://://://://://://://://://://://://://aa

://assistant

://

://

://

://

://

зрения

Perturbation scale -10.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69 cm)

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69 cm)

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +10.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +20.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically tall and lean, with a short

Perturbation scale +100.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a medium to large-sized dog breed. They originated from England in 1098.

The Greyhound is a dog breed that is known for athleticism.

The Greyhound is a dog breed that is known for athleticism.

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0, AXIS: 2 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC2 ---

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -20.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. Originating from England, this breed has been used for centuries as a hunting companion, particularly for hunting small game such as rabbits and hares. Greyhounds are

Perturbation scale -10.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58

Perturbation scale -5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +10.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +20.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58

Perturbation scale +100.0x: Let's take a look at two popular dog breeds:

1. \*\*Bulldog\*\*:

The Bulldog, also known as the English Bulldog, is a sturdy and compact dog breed. They have a short, easy-to-maintain coat that is usually red or brindle in color. Their face is flat and wrinkled,

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0, AXIS: 3 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC3 ---

Perturbation scale -100.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. Poodle\*\*

The Poodle is a popular and intelligent dog breed. They are known for their loyalty and trainability. Poodles are medium-sized dogs that weigh between 40-70 pounds. They have a short, easy-to-train coat and a low-maintenance

Perturbation scale -20.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to chase small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -10.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are characterized by their:

\* Short, smooth coats that

Perturbation scale -5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are characterized by their:

\* Short, smooth coats that

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +10.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +20.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0, AXIS: 4 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC4 ---

Perturbation scale -100.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound (Greyhound is not a breed, I will describe the Greyhound's cousin, the Whippet and the Greyhound's cousin the Italian Greyhound, but I will describe the Italian Greyhound's cousin the Italian Greyhound's cousin the Italian Greyhound

Perturbation scale -20.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically tall and lean, with a short, smooth

Perturbation scale -10.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +10.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are characterized by their:

\* Short, smooth coats that

Perturbation scale +20.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58

Perturbation scale +100.0x: Let's take a look at two popular dog breeds: the Poodle and the Bulldog.

1. \*\*Poodle:\*\*

The Poodle is a small to medium-sized dog breed that belongs to the group of non-sporting dogs. They are known for their low-shedding and hypoallergic nature. Poodles are intelligent, active

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC0 'dog' direction (Layer 0) ---

Projection values:

Top positive #1: 0.1017

Top positive #2: 0.0918

Top positive #3: 0.0853

Top negative #1: -0.0865

Top negative #2: -0.0859

Top negative #3: -0.0779

Top 10 prompts most aligned with POSITIVE PC0 direction:

1. 'We adopted a senior dog from the shelter, and he is the sweetest, most gentle companion.'

2. 'There, in the center of the mess he created, sat the unrepentant dog, wagging his tail.'

3. 'The dog's joy is simple, pure, and infectious.'

4. 'The dog's life is a beautiful, albeit brief, journey.'

5. 'She loves to ride in the car, no matter the destination.'

6. 'My dog's life revolves around a simple and beautiful schedule: eat, play, sleep, love, repeat.'

7. 'She loves the beach, chasing the waves back and forth.'

8. 'She is the perfect hiking companion.'

9. 'He is not a fan of the mailman, the delivery driver, or squirrels.'

10. 'Frustrated, the dog let out a huff and dropped the toy, waiting for me to throw it again.'

Top 10 prompts most aligned with NEGATIVE PC0 direction:

1. 'That dog is a furry alarm clock with a wet nose and no snooze button.'

2. 'The celebrity's tiny teacup dog had its own social media account with millions of followers.'

3. 'Why does one dog in a neighborhood start a barking chain reaction with every other dog?'

4. 'Why does every dog feel the need to circle three times before finally lying down?'

5. 'The first-ever domesticated dog is believed to have descended from an ancient wolf population.'

6. 'My dog always brings me a 'gift'—usually a soggy toy or a stolen sock—when I get home.'

7. 'The guide dog successfully navigated its owner across a busy intersection during rush hour.'

8. 'My neighbor's golden retriever dog is incredibly well-trained and never jumps on guests.'

9. 'The new city ordinance requires every dog owner to carry proof of rabies vaccination.'

10. 'The mushers prepared their sled dog teams for the grueling thousand-mile Iditarod race.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC1 'dog' direction (Layer 0) ---

Projection values:

Top positive #1: 0.0662

Top positive #2: 0.0646

Top positive #3: 0.0634

Top negative #1: -0.0729

Top negative #2: -0.0627

Top negative #3: -0.0591

Top 10 prompts most aligned with POSITIVE PC1 direction:

1. 'That dog is incredibly intuitive and always seems to know when someone needs comforting.'

2. 'Her boundless love is a gift I don't deserve but cherish.'

3. 'After a long day of hiking, the tired dog fell asleep before he even finished his dinner.'

4. 'She can tell when I'm just getting my shoes on versus when I'm getting her leash.'

5. 'My dog has a very clear 'I'm sorry' face he makes after he knows he has done something wrong.'

6. 'The old dog moved stiffly in the morning but seemed to loosen up after a short walk.'

7. 'Her playful nature is a daily reminder not to take life too seriously.'

8. 'The dog's memory is quite good; he always remembers the people who give him the best treats.'

9. 'I wish my dog could talk so he could tell me what he's thinking about all day.'

10. 'My dog's dramatic sigh seemed to express his profound disappointment with the lack of walks today.'

Top 10 prompts most aligned with NEGATIVE PC1 direction:

1. 'The Schnoodle is a Schnauzer-Poodle mix.'

2. 'The Pomapoo is a Pomeranian and Poodle mix.'

3. 'The Cocker Spaniel has long, floppy ears.'

4. 'The Labradoodle combines the traits of a Labrador and a Poodle.'

5. 'The Pomsky is a cross between a Pomeranian and a Husky.'

6. 'The Goldendoodle is a cross between a Golden Retriever and a Poodle.'

7. 'The Maltipoo is a mix of a Maltese and a Poodle.'

8. 'The Newfypoo is a cross between a Newfoundland and a Poodle.'

9. 'The Sheepadoodle is an Old English Sheepdog and Poodle mix.'

10. 'The dog's friendship is a bond that transcends words.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC2 'dog' direction (Layer 0) ---

Projection values:

Top positive #1: 0.0543

Top positive #2: 0.0524

Top positive #3: 0.0515

Top negative #1: -0.0566

Top negative #2: -0.0537

Top negative #3: -0.0521

Top 10 prompts most aligned with POSITIVE PC2 direction:

1. 'Come!'

2. 'Fetch!'

3. 'Sit!'

4. 'Off!'

5. 'Drop it!'

6. 'Stay.'

7. 'Do dogs dream?'

8. 'Heel.'

9. 'Is it true that some breeds are hypoallergenic?'

10. 'How long do dogs live?'

Top 10 prompts most aligned with NEGATIVE PC2 direction:

1. 'The dog, covered from nose to tail in mud, was not allowed back in the house until he had a thorough bath.'

2. 'The dog's coat was a beautiful brindle pattern, a mix of black, brown, and gold stripes.'

3. 'The dog's ears perked up, and he let out a low growl, alerting us to someone approaching the door.'

4. 'There, in the center of the mess he created, sat the unrepentant dog, wagging his tail.'

5. 'The dog, a blur of fur and motion, joyfully chased the ball down the long hallway.'

6. 'The sheepdog, with its intense gaze, kept the flock tightly packed and moving in the right direction.'

7. 'The dog's deep sigh, heard from the other room, signaled his boredom with our lack of activity.'

8. 'Through the pouring rain, the lost dog searched for a dry place to find shelter for the night.'

9. 'The dog's intense focus on the ball, to the exclusion of all else, is a form of meditation.'

10. 'The dog's playful bow, with his front end down and his back end up, was an invitation to chase.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC3 'dog' direction (Layer 0) ---

Projection values:

Top positive #1: 0.0646

Top positive #2: 0.0598

Top positive #3: 0.0587

Top negative #1: -0.0449

Top negative #2: -0.0415

Top negative #3: -0.0408

Top 10 prompts most aligned with POSITIVE PC3 direction:

1. 'If a dog could have a job, what would it be?'

2. 'If a dog could give advice to humans, what would it say?'

3. 'If a dog could make a wish, what would it be?'

4. 'If dogs could vote, what issues would matter to them?'

5. 'If a dog could send a text, what would it say?'

6. 'If a dog could write a review of its owner, what would it say?'

7. 'If a dog could write a thank-you note, what would it say?'

8. 'If a dog could be any other animal for a day, what would it choose?'

9. 'If a dog could write a thank-you note to its owner, what would it say?'

10. 'If dogs could run a business, what would it be?'

Top 10 prompts most aligned with NEGATIVE PC3 direction:

1. 'The Siberian Indian Dog is a Husky and Native American Indian Dog cross.'

2. 'The mushers prepared their sled dog teams for the grueling thousand-mile Iditarod race.'

3. 'The Gerberian Shepsky is a German Shepherd and Siberian Husky mix.'

4. 'The Alaskan Malamute was bred for hauling heavy freight.'

5. 'The Welsh Springer Spaniel is a distinct breed from the more common English Springer Spaniel.'

6. 'The Neapolitan Mastiff is known for its massive size and loose, wrinkled skin.'

7. 'The pet adoption fee covers spaying/neutering and initial vaccinations.'

8. 'During the thunderstorm, the terrified dog hid under the kitchen table, shivering uncontrollably.'

9. 'The sled dog team worked in perfect harmony, pulling the heavy sled across the frozen tundra.'

10. 'The boarding facility offers luxury suites for pampered pooches.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC4 'dog' direction (Layer 0) ---

Projection values:

Top positive #1: 0.0426

Top positive #2: 0.0417

Top positive #3: 0.0411

Top negative #1: -0.0461

Top negative #2: -0.0392

Top negative #3: -0.0336

Top 10 prompts most aligned with POSITIVE PC4 direction:

1. 'He's a very easy-going dog who is happy just to be with his people.'

2. 'That dog is a true original, a one-of-a-kind canine with a personality all his own.'

3. 'A dog's nose print is as unique as a human's fingerprint, and can be used for identification.'

4. 'Is it true that a dog's sense of smell is thousands of times better than a human's?'

5. 'The bond between a human and their dog is often profound.'

6. 'Her intelligence is sometimes a little too much for her own good.'

7. 'The howl of a wolf is not so different from that of some domestic breeds.'

8. 'The dog's sense of hearing is far more sensitive than a human's.'

9. 'He is a very magnificent and powerful animal.'

10. 'He is a very powerful and magnificent-looking animal.'

Top 10 prompts most aligned with NEGATIVE PC4 direction:

1. 'During the thunderstorm, the terrified dog hid under the kitchen table, shivering uncontrollably.'

2. 'Underneath the porch, the timid rescue dog watched the world with wary, uncertain eyes.'

3. 'His tail started wagging the moment I picked up the leash.'

4. 'She runs in her sleep, probably chasing dream squirrels.'

5. 'The mushers prepared their sled dog teams for the grueling thousand-mile Iditarod race.'

6. 'The sled dog team worked in perfect harmony, pulling the heavy sled across the frozen tundra.'

7. 'There, in the center of the mess he created, sat the unrepentant dog, wagging his tail.'

8. 'When I came home, the guilty-looking dog had spread trash all over the kitchen floor.'

9. 'The doggie daycare sends report cards home.'

10. 'How do dogs feel about mail carriers?'

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0, PROJECTION ONTO PC0 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: LetTheTheThe

TheTheTheTheTheTheTheTheThe

The

Information

Information

Perturbation scale -20.0x:

HereI

HereHere

HereHere

I I IHereHereHere

HereHereHereHereTheThe

HereHereHereHereHereHereHereHereHere

Perturbation scale -10.0x:

The

Perturbation scale -5.0x:

TheI

Perturbation scale -2.5x:

TheThe

Perturbation scale -1.5x:

TheThe....

I IHereHere,

,

,

Perturbation scale +0.0x:

IThere I I I I I I I I I I I I

I

I I I I I

I II III I I I I IIIIII

III

Perturbation scale +1.5x: IITheThereThere I I I I I I I I I I I I I

Here

I

I

Perturbation scale +2.5x: assistant###

HereTwo two

IThe The

Let

TheThe

Perturbation scale +5.0x: assistant\*\*

TheTheTheTheLet

://assistant

Perturbation scale +10.0x: ://://://://://://://://://://://://://://://://://://:// ://:// ://://://:// theassistant://://://://.S

://://.S

Perturbation scale +20.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale +100.0x: describe and breeds and andassistantassistant

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0, PROJECTION ONTO PC1 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://стастаста://://ста://://://://://://://://://://://ста://://

Perturbation scale -20.0x: ://://://://://://://://://://

Perturbation scale -10.0x: ://://://://://://://://://://://://://://://стаста

Perturbation scale -5.0x: ://://://://://://://://://://://://:// the the://://202://:// ://aa't://:// the://://assistantassistant the://:// (aa the://:// ( the://:// ( the://:// theaaassistant

Perturbation scale -2.5x: ://assistantThe://The://

://://The:// the://

://://assistantassistantassistant://The://://://://://://assistant

://://:// I://

Perturbation scale -1.5x:

The

TheThe

Iassistant://://://://://://assistantThe://://

://

://

://

://

://

Perturbation scale +0.0x: ://assistantassistant

://

://

://

I

://

://

Perturbation scale +1.5x: ://

Assistant

assistant://

Perturbation scale +2.5x: assistant://

assistant

IIIII

assistantI

Perturbation scale +5.0x:

I

I I I

I

I I I IHereassistant

://

assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

assistant

Perturbation scale +10.0x:

I I I

Perturbation scale +20.0x:

II

assistantassistantassistantassistantassistantassistantassistant

assistant

assistantassistantassistantassistantassistant

Perturbation scale +100.0x:

II

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0, PROJECTION ONTO PC2 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -20.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -10.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -5.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -2.5x:

Thereassistantassistant

Iassistantassistant://://://

Theassistant://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -1.5x: assistantassistant

assistantassistant://:// theassistant://://:// (assistantassistant://://://.assistantassistant://://://.Sassistantassistant.logassistant://://://.log

assistantassistant://.logI

Perturbation scale +0.0x: assistantassistantassistantILetassistantassistantassistantLetassistantassistantITheassistantassistantITheassistant

Perturbation scale +1.5x: assistantassistantassistant

assistantassistantassistantassistantTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheTheThe)

The

Perturbation scale +2.5x: assistantassistanttitle\

assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

assistantassistantassistantassistantassistantassistant

Perturbation scale +5.0x:

II

TheassistantINoteassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

TheThe

assistantassistantassistantassistantassistantassistantassistantassistantassistantII

assistantassistantassistantII

assistantassistantassistantII

assistant

Perturbation scale +10.0x:

IIIII

assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +20.0x:

Perturbation scale +100.0x: IIIThereOneThereThereThereThereThere

assistant

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0, PROJECTION ONTO PC3 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x:

Perturbation scale -20.0x:

Perturbation scale -10.0x:

IIIThereThereThereThereThere IThere

There I I I III

assistantassistant

assistantassistantassistantassistant

assistantassistantassistantassistant

Perturbation scale -5.0x: assistantassistantTheThe•

assistantassistant

Perturbation scale -2.5x: assistantassistantThe+</

Perturbation scale -1.5x: TheTheassistant

Perturbation scale +0.0x: ://assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +1.5x: ://

:// the://

://assistant the://

://assistant the the

://assistant

://assistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +2.5x: ://://://://://://://://:// assistant:// the://

:// the:// the theassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +5.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://202assistant://://://://://.Sassistant://.S

Perturbation scale +10.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale +20.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale +100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0, PROJECTION ONTO PC4 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x:

Perturbation scale -20.0x: assistant

assistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale -10.0x: assistantassistantassistant

Perturbation scale -5.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale -2.5x: Letassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale -1.5x: Letassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +0.0x: The

TheLet

Perturbation scale +1.5x: TheTheTheTheTheTheTheTheTheTheTheTheTheInformation

Perturbation scale +2.5x:

TheTheTheInformation

Information

InformationInformation

Informationassistant assistantassistant

Perturbation scale +5.0x:

the the the assistant the the the the the the the the the the the the help

the

the

the

the

assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +10.0x: ://:// the assistant:// assistant:// the ( the..aa the ( the the the the the the the the the the the the the the the the

the

assistant

assistant

assistant

assistant

assistantassistantassistant

Perturbation scale +20.0x: ://://://://://://://://://://://://://://://aa ( the the the the the the the the the the the the the the the the

assistant

assistantassistant

assistantassistant

assistantassistant

assistant

Perturbation scale +100.0x: ://:// ://://://://://://aa ( the the ( ( the the the the the the the the the the the the the the the the the the the the the the the the the the help help the the the the the

================================================================================

================================================================================

--- ORTHOGONAL PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 0 ---

================================================================================

Perturbing along first orthogonal direction (PC132)...

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along orthogonal direction ---

Perturbation scale -100.0x: I'd be happy to Describe Two Dog Breeds.

\*\*Bulld\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Here are\*\*

Perturbation scale -20.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to chase small game such as rabbits and hares. Greyhounds are typically tall and lean, with a short

Perturbation scale -10.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to chase small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +10.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +20.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58

Perturbation scale +100.0x: I'd be happy to help

Here are two dog breeds

Here are two dog breeds

Here are two dog breeds

Here are two dog breeds

Here are two dog breeds

Here are two dog breeds

Here are two dog breeds

Here are two dog breeds

Here are two dog breeds

Here are two dog breeds

Here are two

================================================================================

================================================================================

--- ABLATION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 0 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

All PCs (centroid only):

All PCs except largest (PC0 only):

All PCs except largest two (PC0+PC1):

================================================================================

################################################################################

### STARTING ANALYSIS FOR LAYER 15 ###

################################################################################

Extracting activations from layer 15...

Extracting activations: 100%|███████████████████████████| 1944/1944 [00:43<00:00, 45.05it/s]

Extracting activations from layer 15...

Extracting activations: 100%|█████████████████████████████| 247/247 [00:05<00:00, 45.22it/s]

Concept 'dog': Found 133 effective eigenvectors out of 1944 (threshold: 0.0056)

Concept 'lion': Found 28 effective eigenvectors out of 247 (threshold: 0.0331)

================================================================================

Running experiments for system prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User prompt: 'The dog was running around the park. It was a labrador.'

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15, AXIS: 0 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC0 ---

Perturbation scale -100.0x: tab thrown answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered

Perturbation scale -20.0x: " bl-An, bl, bl, bl, bl,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,·,·,·,

Perturbation scale -10.0x: Der Hund lief auf allen aus, im Park. Es war ein Labrador.

Perturbation scale -5.0x: Der Hund lief im Park herum. Er war ein Labrador.

Einige Menschen haben oft verschiedene Tipps, um Hunde zu identifizieren, aber die meisten Experten sagen, dass Labrador-Rüter sind bekannt für ihre freundlichen und intelligente Natur.

Perturbation scale -2.5x: Der Hund lief im Park herum. Es war ein Labrador.

Hier sind einige Tipps, um die Übersetzung zu verstehen:

- "The dog" wird in Deutsch mit "Der Hund" oder "Ein Hund" übersetzt, je nachdem, ob es sich um einen bestimmten Hund oder einen beliebigen Hund hand

Perturbation scale -1.5x: Der Hund lief im Park herum. Es war ein Labrador.

Hier ist eine detailliertere Übersetzung:

- "The dog" wird mit "Der Hund" übersetzt, da "der Hund" in der deutschen Sprache der männliche Hund ist.

- "was running" wird mit "lief" übers

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wurde ins Deutsche übersetzt als "Die Hündin", da "Hündin" die weibliche Form von "Hund" ist. Wenn du "The dog" als männlich meinst, dann

Perturbation scale +2.5x: Die Hündin lief im Park herum. Es war ein Labrador.

(Note: I translated "The dog" to "Die Hündin" because it's more likely that a female dog was meant, as "labrador" is a breed that can be either male or female. If you want to translate it to "Der Hund"

Perturbation scale +5.0x: Die Hündin lief um den Park herum.

Die Übersetzung in deutschem Genitiv:

Die Hündin ist ein Labrador.

Perturbation scale +10.0x: Die Wichtart in die German into German

Into German into German

Perturbation scale +20.0x:

utscheauf uren uren

Perturbation scale +100.0x: swickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickórswickórswickórswickswickórswickswickórórswickórswickórswickórswick

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15, AXIS: 1 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC1 ---

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://ishmentaganaganaganundiaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganagan

Perturbation scale -20.0x: Der Hund lief auf dem Park bei der Beule der Bease in dem die laboser in dem die laboser in der span aus der in der aus der in der in der in der in der in der in der in der in der in der in der in der in der in der in der in der in der in der in der in der

Perturbation scale -10.0x: Der Hund lief im Park herum. Es war ein Labrador.

Perturbation scale -5.0x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale -2.5x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale -1.5x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +2.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +5.0x: Die Hündin lief im Park herum. Es war ein Labrador.

Note: I translated "dog" to "Hündin" because it's more common to use the feminine form when referring to a dog in general, as the sex of the dog is not specified. If you want to be more specific, you can use "der Hund

Perturbation scale +10.0x: Der Hund lief im Park herum. Es war ein Labrador.

Here's a breakdown of the translation:

- "The" is a common article in English, and it doesn't have a specific translation. However, in this context, it's a general term for a noun, and it's not necessary to translate it.

- "The dog"

Perturbation scale +20.0x: The dog was running around the park. It was a labrador.

This is a simple phrase, and it's a simple phrase.

The "answer" is a simple phrase, and it's a simple phrase.

The "answer" is a simple phrase.

"Dog" is "Dog"

"Running" "Running"

"

Perturbation scale +100.0x: ://:// Angeles Angeles Angeles Angeles wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15, AXIS: 2 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC2 ---

Perturbation scale -100.0x: iloiloiloiloiloiloilo NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC

Perturbation scale -20.0x: The dog was running around the park. It was a labrador.

This is a poetic poem.

The dog ran in the park. It's a labrador.

The dog is a labrador. It's a dog.

The dog is a laboid. It's a dog.

The dog is a lab. It's

Perturbation scale -10.0x: Der Hund lief im Park herum. Es war ein Labrador.

Perturbation scale -5.0x: Der Hund lief im Park herum. Es war ein Labrador.

Perturbation scale -2.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "dog" im Kontext der Aussage ein weibliches Tier beschreibt. Wenn es sich um einen männlichen Hund handelt, würde "der

Perturbation scale -1.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "dog" im Kontext der Übersetzung als weiblich angenommen wird. Wenn "dog" als männlich gemeint ist, würde

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn Sie "The dog" als männlichen Hund meinen, sollten Sie "Der Hund" oder "

Perturbation scale +2.5x: Die Hündin lief um den Park herum. Sie war eine Labradorhündin.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" die weibliche Form von "Hund" ist. Wenn Sie "Hund" oder "der Hund"

Perturbation scale +5.0x: Die Hündin lief um den Park herum. Sie war eine Labradorin.

Hinweis:

- "Hündin" ist die weibliche Form von "Hund" in Deutsch.

- "Labradorin" ist die weibliche Form von "Labrador" in Deutsch.

Perturbation scale +10.0x: Die Hündin lief um den Park herum. Sie war eine Labradorin.

Hinweis: "Hündin" ist die weibliche Form von "Hund" in deutscher Sprache. "Labradorin" ist die weibliche Form von "Labrador".

Perturbation scale +20.0x: Die Haltung in der deutschen Sprabe lautet:

Die Haltung in der Sprabe spruch spruch sprabe sprabe sprabe Haltung Haltung Haltung Haltung Haltung Haltung Haltung Haltung Haltung Haltung Haltung Haltung Haltung Haltung H

Haltung H H H H H H H H H

Perturbation scale +100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15, AXIS: 3 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC3 ---

Perturbation scale -100.0x: ersetersetersetersetseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseau

Perturbation scale -20.0x: Hier ist eine Übersetung:

Der Hund lief um das Park. Es war ein Labrador.

Für eine genauere Übersetung, ich empfehle die folgende Übersetung:

Zuerst, ich empfehle die Übersichtliche Übersetung:

Zuerst, ich empfehle die Übers

Perturbation scale -10.0x: Hier ist die Übersetzung:

Der Hund lief im Park herum. Es war ein Labrador.

Hier sind einige Tipps für die Übersetzung:

- "The dog" wird als "Der Hund" übersetzt, da "Der Hund" im Deutschen der erste Teil des Artikels ist.

- "was running around

Perturbation scale -5.0x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "die Hündin" für weibliche Hunde verwendet wird.

- "labrador" wird im Deutschen nicht mit einem Artikel übersetzt

Perturbation scale -2.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn Sie "The dog" als männlichen Hund meinen, sollten Sie "Der Hund" oder "

Perturbation scale -1.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weiblicher Hund ist. Wenn Sie "The dog" als männlichen Hund meinen, sollten Sie "Der Hund" oder "

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "die Hündin" ein weibliches Tier bezeichnet. Wenn es sich um einen männlichen Hund handelt, würde "der Hund" verwendet werden

Perturbation scale +2.5x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale +5.0x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale +10.0x: Der Hund lief im Park herum. Er war ein Labrador.

Perturbation scale +20.0x: Der Hund lief im Park herum. Er war ein Labrador.

Perturbation scale +100.0x: ://://://://://://

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

ingroup.

.

.

ingroup

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15, AXIS: 4 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC4 ---

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -20.0x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale -10.0x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale -5.0x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale -2.5x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "die Hündin" ein weibliches Tier bezeichnet. Wenn es sich um einen männlichen Hund handelt, wird "der Hund" verwendet.

Perturbation scale +2.5x: Die Hündin lief im Park herum. Sie war eine Labradorhündin.

Hinweis:

- "The dog" wird mit "Die Hündin" übersetzt, da "die Hündin" ein weibliches Tier bezeichnet. Wenn es sich um einen männlichen Hund handelt, wird "der Hund

Perturbation scale +5.0x: Die Hündin lief im Park herum. Sie war eine Labradorin.

Hinweis:

- "Hündin" wird verwendet, um eine weibliche Hündin zu beschreiben, da "Hund" neutral ist und "Hundin" weiblich ist.

- "Labradorin" ist die weib

Perturbation scale +10.0x: Der Hund lief im Park herum. Es war ein Labrador.

Hier ist eine Übersetzung:

- "Der Hund" ist die direkte Übersetzung von "the dog" und bezieht sich auf den Hund, der im Park herumläuft.

- "lief im Park herum" ist eine direkte Ü

Perturbation scale +20.0x: Der Hund ran im Park herum. Es war ein Labrador.

In diesem Szenz ist "herum" ein sehr informeller Ausdruck, der in der US-Amerikanischen Sprache verwendet wird. In der deutschen Sprache wird man in diesem Szenz "herum” in der US-Amerikanischen Sprache verwendet

Perturbation scale +100.0x: isyisyimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimaniman

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC0 'dog' direction (Layer 15) ---

Projection values:

Top positive #1: 1.2910

Top positive #2: 1.2793

Top positive #3: 1.2764

Top negative #1: -5.9375

Top negative #2: -5.9102

Top negative #3: -5.8984

Top 10 prompts most aligned with POSITIVE PC0 direction:

1. 'Her playful spirit is a constant source of joy and laughter.'

2. 'The dog's powerful legs propelled him through the snow with surprising ease.'

3. 'The dog's contentment was obvious as he dozed in a patch of warm sunlight on the floor.'

4. 'The sheepdog herded the flock with impressive skill.'

5. 'The dog's thick, water-repellent coat makes him a natural swimmer and retriever.'

6. 'He snoozed, twitching his paws as if dreaming.'

7. 'The greyhound stretched languidly on the rug.'

8. 'The sled dog team worked in perfect harmony, pulling the heavy sled across the frozen tundra.'

9. 'The farm dog worked tirelessly from dawn until dusk, a loyal and indispensable helper.'

10. 'His enthusiasm for life is a daily inspiration.'

Top 10 prompts most aligned with NEGATIVE PC0 direction:

1. 'What is the best way to introduce my dog to strangers?'

2. 'What is the best way to teach my dog to swim?'

3. 'What is the best way to introduce my dog to children?'

4. 'What is the best way to teach my dog to ring a bell to go outside?'

5. 'What are the best dog breeds for scent work?'

6. 'What is the best way to prevent ticks on my dog?'

7. 'What is the best way to teach my dog to back up on command?'

8. 'What is the best way to teach my dog to jump through hoops?'

9. 'How do I help my dog with fear of water?'

10. 'What are the best dog breeds for running partners?'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC1 'dog' direction (Layer 15) ---

Projection values:

Top positive #1: 2.4219

Top positive #2: 2.3574

Top positive #3: 2.3555

Top negative #1: -2.4238

Top negative #2: -2.3281

Top negative #3: -2.3125

Top 10 prompts most aligned with POSITIVE PC1 direction:

1. 'Heel.'

2. 'Stay.'

3. 'Off!'

4. 'Come!'

5. 'Sit!'

6. 'What a good boy!'

7. 'Drop it!'

8. 'Beware of Dog.'

9. 'He has a heart of gold.'

10. 'Let sleeping dogs lie.'

Top 10 prompts most aligned with NEGATIVE PC1 direction:

1. 'My dog has a very clear 'I'm sorry' face he makes after he knows he has done something wrong.'

2. 'That dog has a very specific, high-pitched bark he reserves only for the mail carrier.'

3. 'My dog has a very specific routine that involves napping in three different sunny spots throughout the day.'

4. 'My dog's response to the command 'stay' is more of a suggestion he politely considers and then ignores.'

5. 'My dog's interpretation of 'helping' with laundry is to steal socks and run away with them.'

6. 'My dog has a very specific, and very loud, bark reserved for the squirrel that taunts him daily.'

7. 'The ancient Roman mosaics often depicted a 'beware of the dog' warning at the entrance of homes.'

8. 'That dog is a master escape artist who has figured out how to unlatch three different types of gates.'

9. 'My dog's vocal range includes barks, yips, growls, howls, and a strange sort of purr.'

10. 'My dog's love for car rides is so intense that he tries to jump into any open car door he sees.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC2 'dog' direction (Layer 15) ---

Projection values:

Top positive #1: 1.8955

Top positive #2: 1.8564

Top positive #3: 1.7930

Top negative #1: -2.6523

Top negative #2: -2.6270

Top negative #3: -2.5703

Top 10 prompts most aligned with POSITIVE PC2 direction:

1. 'The Black and Tan Coonhound is an American scent hound.'

2. 'The Kuvasz is a large Hungarian guardian breed.'

3. 'How do I prepare my dog for a vet visit?'

4. 'The Saint Berdoodle is a Saint Bernard and Poodle mix.'

5. 'The Pomapoo is a Pomeranian and Poodle mix.'

6. 'The Alaskan Malamute was bred for hauling heavy freight.'

7. 'The Rhodesian Ridgeback was originally bred to hunt lions.'

8. 'The Mastador is a Mastiff and Labrador cross.'

9. 'How do I train my dog to walk on a leash?'

10. 'How do I help my dog with crate training?'

Top 10 prompts most aligned with NEGATIVE PC2 direction:

1. 'What would a dog say about its own reflection?'

2. 'What would a dog say about going to the vet?'

3. 'What would a dog say about the invention of the leash?'

4. 'What would a dog say about moving to a new home?'

5. 'How would a dog describe the taste of peanut butter?'

6. 'What would a dog say about its first walk?'

7. 'What would a dog say about growing old?'

8. 'What would a dog's diary entry look like after a day at the park?'

9. 'What would a dog say about thunderstorms?'

10. 'What would a dog say about meeting a new friend?'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC3 'dog' direction (Layer 15) ---

Projection values:

Top positive #1: 2.3535

Top positive #2: 2.3516

Top positive #3: 2.2754

Top negative #1: -1.7393

Top negative #2: -1.5928

Top negative #3: -1.5820

Top 10 prompts most aligned with POSITIVE PC3 direction:

1. 'Is he allowed to have a bone?'

2. 'Could you grab the poop bags?'

3. 'I need to trim his nails.'

4. 'Is that a purebred or a mix?'

5. 'He seems to know when I'm packing a suitcase.'

6. 'I need to pick up a new collar for him.'

7. 'Is he friendly with other dogs?'

8. 'He's the reason my camera roll is always full.'

9. 'Her favorite person is whichever one is holding the food.'

10. 'Did you remember to give him his heartworm medication?'

Top 10 prompts most aligned with NEGATIVE PC3 direction:

1. 'What is the best way to teach my dog to weave through poles?'

2. 'What is the best way to teach my dog to ring a bell to go outside?'

3. 'How do I teach my dog to balance objects on its nose?'

4. 'How do I teach my dog to find hidden treats?'

5. 'How do I teach my dog to push a ball with its nose?'

6. 'How do I teach my dog to carry objects?'

7. 'The dog's joy was palpable as he ran freely on the sandy beach for the very first time.'

8. 'How do I teach my dog to fetch specific items?'

9. 'The sled dog team worked in perfect harmony, pulling the heavy sled across the frozen tundra.'

10. 'How do I teach my dog to open and close doors?'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC4 'dog' direction (Layer 15) ---

Projection values:

Top positive #1: 2.4453

Top positive #2: 2.4395

Top positive #3: 2.4082

Top negative #1: -2.3262

Top negative #2: -2.1758

Top negative #3: -2.1035

Top 10 prompts most aligned with POSITIVE PC4 direction:

1. 'How do I teach my dog to balance objects on its nose?'

2. 'What is the best way to teach my dog to weave through poles?'

3. 'How do I teach my dog to find hidden treats?'

4. 'How do I teach my dog to carry objects?'

5. 'How do I teach my dog to push a ball with its nose?'

6. 'How do I teach my dog to high five?'

7. 'How do I teach my dog to open and close doors?'

8. 'How do I teach my dog to shake hands?'

9. 'How do I teach my dog to fetch specific items?'

10. 'What is the best way to introduce my dog to cats?'

Top 10 prompts most aligned with NEGATIVE PC4 direction:

1. 'Write a riddle about a dog.'

2. 'What would a dog say about its family?'

3. 'What would a dog say about moving to a new home?'

4. 'Invent a new dog sport and explain the rules.'

5. 'Write a song chorus about a loyal dog.'

6. 'Write a letter from a dog to a cat.'

7. 'What would a dog say about meeting a new friend?'

8. 'Write a letter from a dog to its owner.'

9. 'Describe a dog's favorite place in the world.'

10. 'What would a dog say about its best friend?'

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15, PROJECTION ONTO PC0 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: ://tab answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered

Perturbation scale -20.0x: lab lab lost lost lost lost A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A··

Perturbation scale -10.0x: day day day day “ “ “ “ “ “ “ “ “ “ “.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

Perturbation scale -5.0x:

Expert Expert mit mit mit mit with with with with avec avec avec avec with with - with with - - - - - - - old - - - old old old old old old old old old old old old old old old old old old old

Perturbation scale -2.5x:

Die über über über über überäl,,,,, andere andere andere andere andere it it it it it it nature nature their their their their nature nature nature name name name name name name name name name name name name name name name name

Perturbation scale -1.5x: die die,,, in so so über über für bei auf auf bei bei,,,,,, der der nach nach auf nach auf auf after after name name name name name name name name name name name name name name

Perturbation scale +0.0x: a at for for for for for a a a a a a a a a a a various terms up source a hand than than than than,,,,,,,,,,ical,,,,,, in ( ( ( in

Perturbation scale +1.5x:

,,,,,,,,,,,,,,,,,,,\*\*

\*\*

\*\*

\*\*

\*\*

\*\*

\*\*

.......................

Perturbation scale +2.5x: Vollassistantdogdogdogdogdogdog,............assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +5.0x: icallyaseaseaseassistantassistantassistantassistantccocestorcestorcestorcestorcestorcestorcestorcestorcestorcestorcestorcestor.n.n.n.n.n.n

Perturbation scale +10.0x:

fyfyfyfyction German Germanemannemannibanibanibanibanignonignonignonignonignonibanibanibanibanibanibaniban

Perturbation scale +20.0x:

Perturbation scale +100.0x: swickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswick

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15, PROJECTION ONTO PC1 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://aganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganagan

Perturbation scale -20.0x: N nach nach nach nach in in in in in in in in in er er er er er er er er weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter weiter

Perturbation scale -10.0x: German - - - der der der der der der der der der der der der der der der der dersp ist ist ist ist ist ist kanngegegegegegegegegegegegege ist ist ist ist ist ist ist ist

Perturbation scale -5.0x: von von von von von,,,,,.................... ( ( ( ( ( ( ( - ( - ist ist ist ist ist ist ist ist ist ist

Perturbation scale -2.5x: in du,,,, B um um, Um B,,,,,,,,,ververververververververver,,,,,, und ist istemememem,,,,, ist ist

Perturbation scale -1.5x: man de du dog erben um Zeit Zeit, vor sich für für für für für für,,,,,markmarkmarkmark der,ahahah der der dererhhahaherererererererererer

Perturbation scale +0.0x: in man as as as around for for for talking talking the for for for earning for for $ for for comes for $ $ $ $ $ titles,,,, titles $ titles. titles being being

Perturbation scale +1.5x: thing thing up to to

by by by by by by by the by the by the the the the

Perturbation scale +2.5x:

device month month in in to

.

.

.

.

.

.

)

)

)

)

.

.

.

.

.

.

.

by it it it it it it it it it it it it it it it

Perturbation scale +5.0x: ,, Jul million million million::,,,,,,.

.

.

ly!

of.

of of of of of of of of at)

at in in in in in in in in inlylylylylylylylyly

Perturbation scale +10.0x: in with with with

Way Way Way Way Show By By By Show Show Show Show Show Show Show Show Show Show Show Show Show Show Show Show Show Show Show Show Show Show

Perturbation scale +20.0x: without without without , by,,,,,,,,,,,,,,,,,,,,,,,,,,,, " " " ",,,,,,

Perturbation scale +100.0x: ://://://://://:// Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles Angeles wheel wheel

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15, PROJECTION ONTO PC2 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: ://://iloiloiloiloiloiloiloilo NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC

Perturbation scale -20.0x: inededededededededededed ined in in in in in in in in … by … … … … … by by … … … … … … … … … … … by … by … by by by by

Perturbation scale -10.0x: to to and and and and andionion....... ( (...rdrdrdrdrdrdrdrdrdrdrdrdrdendendendendendendendendendendendendendendendend

Perturbation scale -5.0x:

, that that,,,,,, as as... as as as as....... the the the the the the the the the............ the the the

Perturbation scale -2.5x: being not doesn is as as as in in in in,,ers,ers,,.............. the the the the the the the the the the the the the the the the the the

Perturbation scale -1.5x: your fast beingstanding no being being the being being to the theestestestestestestestest the the the the the the the the the the the the the the the the the the the the the the the the the the the the the

Perturbation scale +0.0x: genau andere andere Zeit Zeit,,,,,,,,,,,,,, zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu

Perturbation scale +1.5x:

mal,,,,,,,,,,,,,,,,,,, zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu

Perturbation scale +2.5x:

,,,,,,,,,,,,,,,,,,,,,, zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu

Perturbation scale +5.0x: besf,,,,,,,,,,,,,,vers, ste, ste ste ste ste ste ste ste ste ste ste ste ste zu zu hat hat hat zu zu zu zu zu zu zu zu zu zu zu zu zu

Perturbation scale +10.0x: derststststbesbesbesbesbes gen gen gen gen gen gen gen gen gen gen gen gen gen gen gen ..............

Perturbation scale +20.0x:

die die spr spr

S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S

Perturbation scale +100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15, PROJECTION ONTO PC3 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: adleersetersetersetersetersetersetseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseau

Perturbation scale -20.0x:

zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu

Perturbation scale -10.0x:

,,

|

.

Sr Sr Sr Sr " " " "

In In

In

viceviceente im im In impezepezezezepepepepepepepepe zu zu

Perturbation scale -5.0x:

e,,,, und das:

:

herinzhhh zu zuzhzhzhzhzhzhzhzhzhzhzhhherherher vor vor vor vor vor vor vor vor vor vor vor vor vor vor vor vor

Perturbation scale -2.5x: that that and and ( in ineseseseses zu zu zu zup zuher zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zuherherherbbfallsfallsfalls

Perturbation scale -1.5x: that that that well was right in , to to zu zu zu zu zu zu zu zu zu zu zu zu als als als als als als als alsalsalsals zu zu zu zu zuals zu zu zu zu zu zu zu

Perturbation scale +0.0x: as trying time what take take take take committed what was was no1 no no were were were don don have were were were .. in. was was11

Perturbation scale +1.5x: , bigger

zeigen zeigen zeigen finden finden finden " " "(",,,,,,,,,,,,,,,,,,,,............

Perturbation scale +2.5x: ,,, take take,,,,, ( ( ( (....................................

Perturbation scale +5.0x: ,,,,,,,,, auf auf auf. at at at at at at...............................

Perturbation scale +10.0x: . is is on about about them them them them................. with with............ with with with with with with with with with

Perturbation scale +20.0x: of of of about is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is

Perturbation scale +100.0x: ://://://://://://://://://://://://://://://

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15, PROJECTION ONTO PC4 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -20.0x: ueueueenceenceuceuceuceuceuceuceuceuceuceuceuceuceudaudaudaudaudauteudaudauteuteuteuteuteuteuteuteuteuteuteuteuteutscheutscheutscheutscheutscheutscheutscheutscheutscheenhaenhautsche

Perturbation scale -10.0x: ofereigereringeringeringeringering use::::ce:udyceudycececececececececececece.

Perturbation scale -5.0x: underenstellungellelililiereringiger.iger.unggelgagaungungungungungungung).).).).).).).).uitinsinsinsinsuitinsinsinsinsinsinsinsinsinsinsinsins

Perturbation scale -2.5x: in yourtest,erswertwertlielieababababababgagegegege.........or........or.or......itted).).

Perturbation scale -1.5x: the conversation practice discussionersersershere,lielieillsillsabababgegegegegegegegege,,,,..................itted..

Perturbation scale +0.0x: in for for time for time ,, the,, Europe,'s,,,'s,,,, is, are are are are are are are are are are are are are are's are – – – –'s ( (

Perturbation scale +1.5x: im,,,,,,...,,,.......... Über Über Über Über Über Über Über Über Über Über Über Über Über Über Über........ von...

Perturbation scale +2.5x: , viele, des mit,.............. Über Über Über Über Über Über Über Über Über Über Über Über Über..... Über....... | | | |

Perturbation scale +5.0x: ,, v,,,, mit mit mit mit.......................................

Perturbation scale +10.0x:

zu zu zu zu mit bei ..........................................

Perturbation scale +20.0x: assistantggadadadadadhelpgivegivegive‘ ““ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “ “

Perturbation scale +100.0x: someone someoneimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimaniman

================================================================================

================================================================================

--- ORTHOGONAL PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15 ---

================================================================================

Perturbing along first orthogonal direction (PC133)...

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along orthogonal direction ---

Perturbation scale -100.0x: wlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwl

Perturbation scale -20.0x: Rureneyseureneyseurenurenurenurenurenurenurenurenurenurenurenurenurenurenurenurenurenurenurenurenurenayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayers

Perturbation scale -10.0x: Die folgt ist eine wicht zu stellen:

Since there are two sentences in the following:

1. There are two smalls as follows:

2: 2:

2: 2:

2: 2:

2: 2:

2: 2:

2: 2:

2:

Perturbation scale -5.0x: Die Übersetzung lautet:

Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale -2.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "die Hündin" im Deutschen die weibliche Form für "Hund" ist. Wenn es sich um einen männlichen Hund handelt

Perturbation scale -1.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "die Hündin" sich auf eine weibliche Hündin bezieht. Wenn es sich um einen männlichen Hund handelt, wird

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Der Hund lief im Park herum. Es war ein Labrador.

Hier ist eine Erklärung:

- "The dog" = "Der Hund"

- "was running" = "lief"

- "around the park" = "im Park herum"

- "It was a labrador" = "Es war ein Labrador"

Perturbation scale +2.5x: Der Hund lief im Park herum. Es war ein Labrador.

Hier ist eine Erklärung:

- "The" ist ein Artikel, der in Deutsch nicht übersetzt wird, wenn er vor einem Nomen steht.

- "dog" ist ein Hund, der in Deutsch nicht geändert wird.

- "was" ist ein Pronomen

Perturbation scale +5.0x: Der Hund lief im Park herum. Es war ein Labrador.

\* "The" ist ein Artikel, der in deutscher Sprache üblicherweise nicht übersetzt wird, also "der" oder "die" oder "das".

\* "Labrador" ist ein englischer Name, der in deutscher Sprache auch

Perturbation scale +10.0x: Der Hund war im Park. Er war ein Labrador.

Ein guter. (Ein guter» ein guter» ein gut» ein»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»»

Perturbation scale +20.0x: Einer ein ein ein ein ein ein

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

gr

A

A

Roger

GV

GV

GV

Roger

GV

Roger

Perturbation scale +100.0x: Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow

================================================================================

================================================================================

--- ABLATION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

All PCs (centroid only):

All PCs except largest (PC0 only):

All PCs except largest two (PC0+PC1):

================================================================================

================================================================================

Running experiments for system prompt: 'You are a helpful assistant.'

User prompt: 'Please describe two different dog breeds.'

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15, AXIS: 0 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC0 ---

Perturbation scale -100.0x: answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered

Perturbation scale -20.0x: .

dog, a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,

Perturbation scale -10.0x: There are many dog breeds, from small to large, from playful to serious, and from energetic to relaxed. Here are some popular dog owners, from a variety, from a variety, from a variety, from a variety, from a variety, from a variety, from a variety, from a variety, from a variety, from a day, from

Perturbation scale -5.0x: There are many dog breeds, each with unique characteristics, temperaments, and physical features. Here are two popular dog breeds, the Golden Retriever and the French Bulldog.

\*\*Golden Retrievers\*\*

Golden Retrieers are a popular breed known for their friendly, gentle, and patient nature. They are one of the most popular breeds in the world

Perturbation scale -2.5x: Let's take a look at two popular dog breeds: the Golden Retriever and the Poodle.

\*\*1. Golden Retriever\*\*

The Golden Retriever is a friendly, loyal, and energetic breed that originated in Scotland in the late 19th century. They are known for their thick, lustrous coats that range in color

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and versatile dog breed that comes in a variety of sizes, including Toy, Miniature, and Standard. They are known for their distinctive curly coats, which require regular grooming to prevent matting and tangling. Poodles are

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. Labrador Retriever\*\*

The Labrador Retriever is a popular and friendly dog breed known for its outgoing and affectionate personality. They are medium to large-sized dogs with a short, smooth coat that comes in three colors: black, yellow, and chocolate. Their weight

Perturbation scale +2.5x: Here are descriptions of two different dog breeds:

\*\*1. Poodle\*\*

The Poodle is a highly intelligent and elegant dog breed. It has a distinctive curly or wavy coat that requires regular grooming to prevent matting. The Poodle comes in three sizes: Toy, Miniature, and Standard. The Toy Poodle weighs around 6-

Perturbation scale +5.0x: Here are descriptions of two different dog breeds:

1. \*\*Bulldog\*\*

The Bulldog breed is a short and stout dog with a distinctive appearance. The breed's characteristics include:

- Short stature: The height of the Bulldog breed is approximately 11 inches.

- Weight: The weight of the Bulldog breed is approximately 40 pounds

Perturbation scale +10.0x: Here are the descriptions of the two different dog breeds.

The first breed is the "Chase's English Partionable into the same as the same as the same of the English Partionable into the same of the English Partionable into the same of the English Partionable into the same of the English Partionable into the same of

Perturbation scale +20.0x: =explode

Perturbation scale +100.0x: swickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickórswickórswickórswickórswickórswick

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15, AXIS: 1 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC1 ---

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://acellularaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganagan

Perturbation scale -20.0x: Here are two different dog breeds.

First, the French is one that is known as the - ( translated - (en tradas - (en tradas - es - es - es - es - es - es - es -

Perturbation scale -10.0x: Here are descriptions of two different dog breeds:

\*\*1. The German Shorthaired Pointer (GSP)\*\*

The German Shorthaired Pointer (GSP) is a versatile hunting breed originating from Germany. They are known for their distinctive coat, which is short-haired and often described as'verschwender' (meaning'versatile' in

Perturbation scale -5.0x: Here are descriptions of two different dog breeds:

\*\*1. The Greyhound\*\*

The Greyhound is a large, athletic dog breed known for its speed and agility. Originating from England, this breed is often referred to as a '40 mph couch potato' due to its ability to reach high speeds when racing, but also its love for relaxing and

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. Originating from England, this breed is often referred to as a "40 mph couch potato" due to its ability to reach high speeds when running, but also its love for

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69 cm)

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed that is known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically short-coated, with

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed that is known for its speed and agility. They are a sight hound, bred for centuries to chase small game such as rabbits and hares. Greyhounds are typically short-coated, with

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and versatile dog breed that comes in three sizes: Toy, Miniature, and Standard. They are known for their distinctive curly, non-shedding coat, which requires regular grooming to prevent matting.

Poodles

Perturbation scale +10.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Golden Retriever\*\*

The Golden Retriever is a medium to large-sized dog breed that originated in the late 19th century. This breed was originally bred as a hunting dog, specifically to retrieve game for hunters. Golden Retrievers are known for their friendly, outgoing

Perturbation scale +20.0x: There are many dog breeds, and I'll choose two that are relatively easy to recognize.

1

The "Golden Dog"

The Golden Dog, also known without "Golden Dog", is a classic dog that is a classic.

The Golden Dog is a classic dog that is a classic.

The Golden Dog is a classic dog that is a classic

Perturbation scale +100.0x: Angeles Angeles Angeles Angeles Angeles Angeles wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15, AXIS: 2 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC2 ---

Perturbation scale -100.0x: NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC

Perturbation scale -20.0x: Here's a description of two unique dog breeds:

1. The first is a fictional story about a dog, a fictional story about a dog, a fictional story about a dog, a fictional story about a dog, a fictional story about a dog, a fictional story about a dog, a fictional story about a dog, a fictional story about a dog

Perturbation scale -10.0x: Here are two unique dog breeds:

1. \*\*The Pug\*\*:

- The Pug is a small, playful dog breed known for its adorable appearance and charming personality.

- It has a short, curly coat that is usually black or fawn in color.

- The Pug's face is characterized by a flat, wrinkled

Perturbation scale -5.0x: Let's take a look at two unique dog breeds:

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and elegant dog breed known for its distinctive curly coat. They come in various sizes, including Toy, Miniature, and Standard. This breed is often associated with a high level of trainability, making them a popular choice for

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and elegant dog breed known for its distinctive curly coat. They come in a variety of sizes, including Toy, Miniature, and Standard. Poodles are highly trainable and are often used as show dogs, circus performers

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches tall and

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically tall and lean, with a short, smooth

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically tall and lean, with a short, smooth

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds.

\*\*1. German Shepherd\*\*

The German Shepherd is a highly intelligent and versatile breed that originated in Germany in the late 1800s. They are known for their loyalty, obedience, and protective nature. German Shepherds are medium to large-sized dogs, typically weighing between 75-95

Perturbation scale +10.0x: Here are descriptions of two different dog breeds:

\*\*1. German Shepherd (Deutscher Schäferhund)\*\*

The German Shepherd is a highly intelligent and versatile breed that originated in Germany. They are known for their loyalty, obedience, and protective nature. German Shepherds are medium to large-sized dogs, typically weighing between 55-

Perturbation scale +20.0x: Here are two different dog breeds, known as:

\*\*1. Die Spramtspramin (Widenspramin) - Spramin (Spramin) - Spramin (Spramin) - Spramin (Spramin)

Spramin (Spramin)

Spramin (Spramin)

Spramin (Spramin)

Spr

Perturbation scale +100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://.ComponentPlacementoniconicpendpendpendpendpendpendpendpendorraineorraineorraineorraineorraineonicpendpendpendpendpendpendpendpendpendpendpendpendpendorraineorraineorraineonicpendpendpendpendpendpendpendpendpend

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15, AXIS: 3 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC3 ---

Perturbation scale -100.0x: .learningezezezezezezezezezezseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseau

Perturbation scale -20.0x: Here are two different dog breeds:

\*\*1. The German Shepherd:

\*\*1.\*\* \*\*Overview:\*\*

The German Shepherd is a highly sought-after breed for its intelligence, versatility, and stunning appearance. It is a medium-to-large-sized dog that consists of several key features.

\*\*2.\*\* \*\*Key Features:\*\*

- \*\*Step 1

Perturbation scale -10.0x: Here are descriptions of two different dog breeds:

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its incredible speed and agility. Originating from ancient civilizations, this breed has been a popular choice for racing and hunting. Here are some key characteristics:

- \*\*Physical Description:\*\* Greyhounds have a slender build

Perturbation scale -5.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Here are some key characteristics of the Greyhound breed:

-

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are characterized by their:

\* Short, smooth

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are characterized by their:

\* Short, smooth

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game such as rabbits and hares. Greyhounds are typically short-coated, with a smooth, glossy coat that

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and versatile dog breed that comes in a variety of sizes, including Toy, Miniature, and Standard. They are known for their distinctive curly, non-shedding coat, which requires regular grooming to prevent matting. P

Perturbation scale +5.0x: Let's talk about two popular dog breeds: the Poodle and the Beagle.

\*\*1. Poodle\*\*

The Poodle is a highly intelligent and versatile dog breed that comes in various sizes, including Toy, Miniature, and Standard. They are known for their distinctive curly or wavy coat, which requires regular grooming to prevent matting.

Perturbation scale +10.0x: Let's talk about two different dog breeds: the Poodle and the Bulldog.

1. \*\*Poodle\*\*: The Poodle is a small to medium-sized dog breed known for its distinctive curly or wavy coat. They come in a variety of sizes, including Toy, Miniature, and Standard. Poodles are highly intelligent and often used as

Perturbation scale +20.0x: I said one.

Perturbation scale +100.0x: ://://://://

.

.

.

.

.

.

.

.

.

.

.

.

.

.

.

ingroup.

ingroup.

ingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroup

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15, AXIS: 4 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC4 ---

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -20.0x: Here are two different dog breeds:

\*\*1. The Golden Retriever:\*\*

The Golden Retriever is a majestic dog breed known for its golden coat and incredible abilities to assist and entertain. Created to assist in the retrieval of items, such as a pen, and to assist with other tasks, the Golden Retriever's name was chosen

Perturbation scale -10.0x: I'd be happy to describe two different dog breeds.

\*\*Breed 1: The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its incredible speed. They are often referred to as the "40 mph couch potato" due to their ability to run at incredible speeds, but also love to relax and sleep. Grey

Perturbation scale -5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to chase small game such as rabbits and hares. Greyhounds are typically slender, with a short, smooth coat that

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to chase small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to chase small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69 cm)

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69 cm)

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Labrador Retriever\*\*

The Labrador Retriever is a popular and friendly dog breed known for its outgoing and affectionate nature. They are a medium to large-sized breed, typically weighing between 55-80 pounds and standing between 21.5-24.5 inches

Perturbation scale +10.0x: I'd be happy to describe two different dog breeds. Here are two popular breeds:

\*\*1. Labrador Retriever\*\*

The Labrador Retriever is a friendly and outgoing breed that is known for its gentle and affectionate nature. They are one of the most popular breeds in the world, and it's easy to see why. Labradors

Perturbation scale +20.0x: There are many different dog breeds, but here are two that are popular and distinct:

\*\*1. Labrador (Lab) -\*\*

The Labrador is a popular breed known for its friendly and affectionate nature. This breed is often used as a family pet, as it is known for being gentle and patient with children. The Lab is a medium-sized dog

Perturbation scale +100.0x: avanaavanaavanaimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimaniman

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC0 'dog' direction (Layer 15) ---

Projection values:

Top positive #1: 1.2910

Top positive #2: 1.2793

Top positive #3: 1.2764

Top negative #1: -5.9375

Top negative #2: -5.9102

Top negative #3: -5.8984

Top 10 prompts most aligned with POSITIVE PC0 direction:

1. 'Her playful spirit is a constant source of joy and laughter.'

2. 'The dog's powerful legs propelled him through the snow with surprising ease.'

3. 'The dog's contentment was obvious as he dozed in a patch of warm sunlight on the floor.'

4. 'The sheepdog herded the flock with impressive skill.'

5. 'The dog's thick, water-repellent coat makes him a natural swimmer and retriever.'

6. 'He snoozed, twitching his paws as if dreaming.'

7. 'The greyhound stretched languidly on the rug.'

8. 'The sled dog team worked in perfect harmony, pulling the heavy sled across the frozen tundra.'

9. 'The farm dog worked tirelessly from dawn until dusk, a loyal and indispensable helper.'

10. 'His enthusiasm for life is a daily inspiration.'

Top 10 prompts most aligned with NEGATIVE PC0 direction:

1. 'What is the best way to introduce my dog to strangers?'

2. 'What is the best way to teach my dog to swim?'

3. 'What is the best way to introduce my dog to children?'

4. 'What is the best way to teach my dog to ring a bell to go outside?'

5. 'What are the best dog breeds for scent work?'

6. 'What is the best way to prevent ticks on my dog?'

7. 'What is the best way to teach my dog to back up on command?'

8. 'What is the best way to teach my dog to jump through hoops?'

9. 'How do I help my dog with fear of water?'

10. 'What are the best dog breeds for running partners?'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC1 'dog' direction (Layer 15) ---

Projection values:

Top positive #1: 2.4219

Top positive #2: 2.3574

Top positive #3: 2.3555

Top negative #1: -2.4238

Top negative #2: -2.3281

Top negative #3: -2.3125

Top 10 prompts most aligned with POSITIVE PC1 direction:

1. 'Heel.'

2. 'Stay.'

3. 'Off!'

4. 'Come!'

5. 'Sit!'

6. 'What a good boy!'

7. 'Drop it!'

8. 'Beware of Dog.'

9. 'He has a heart of gold.'

10. 'Let sleeping dogs lie.'

Top 10 prompts most aligned with NEGATIVE PC1 direction:

1. 'My dog has a very clear 'I'm sorry' face he makes after he knows he has done something wrong.'

2. 'That dog has a very specific, high-pitched bark he reserves only for the mail carrier.'

3. 'My dog has a very specific routine that involves napping in three different sunny spots throughout the day.'

4. 'My dog's response to the command 'stay' is more of a suggestion he politely considers and then ignores.'

5. 'My dog's interpretation of 'helping' with laundry is to steal socks and run away with them.'

6. 'My dog has a very specific, and very loud, bark reserved for the squirrel that taunts him daily.'

7. 'The ancient Roman mosaics often depicted a 'beware of the dog' warning at the entrance of homes.'

8. 'That dog is a master escape artist who has figured out how to unlatch three different types of gates.'

9. 'My dog's vocal range includes barks, yips, growls, howls, and a strange sort of purr.'

10. 'My dog's love for car rides is so intense that he tries to jump into any open car door he sees.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC2 'dog' direction (Layer 15) ---

Projection values:

Top positive #1: 1.8955

Top positive #2: 1.8564

Top positive #3: 1.7930

Top negative #1: -2.6523

Top negative #2: -2.6270

Top negative #3: -2.5703

Top 10 prompts most aligned with POSITIVE PC2 direction:

1. 'The Black and Tan Coonhound is an American scent hound.'

2. 'The Kuvasz is a large Hungarian guardian breed.'

3. 'How do I prepare my dog for a vet visit?'

4. 'The Saint Berdoodle is a Saint Bernard and Poodle mix.'

5. 'The Pomapoo is a Pomeranian and Poodle mix.'

6. 'The Alaskan Malamute was bred for hauling heavy freight.'

7. 'The Rhodesian Ridgeback was originally bred to hunt lions.'

8. 'The Mastador is a Mastiff and Labrador cross.'

9. 'How do I train my dog to walk on a leash?'

10. 'How do I help my dog with crate training?'

Top 10 prompts most aligned with NEGATIVE PC2 direction:

1. 'What would a dog say about its own reflection?'

2. 'What would a dog say about going to the vet?'

3. 'What would a dog say about the invention of the leash?'

4. 'What would a dog say about moving to a new home?'

5. 'How would a dog describe the taste of peanut butter?'

6. 'What would a dog say about its first walk?'

7. 'What would a dog say about growing old?'

8. 'What would a dog's diary entry look like after a day at the park?'

9. 'What would a dog say about thunderstorms?'

10. 'What would a dog say about meeting a new friend?'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC3 'dog' direction (Layer 15) ---

Projection values:

Top positive #1: 2.3535

Top positive #2: 2.3516

Top positive #3: 2.2754

Top negative #1: -1.7393

Top negative #2: -1.5928

Top negative #3: -1.5820

Top 10 prompts most aligned with POSITIVE PC3 direction:

1. 'Is he allowed to have a bone?'

2. 'Could you grab the poop bags?'

3. 'I need to trim his nails.'

4. 'Is that a purebred or a mix?'

5. 'He seems to know when I'm packing a suitcase.'

6. 'I need to pick up a new collar for him.'

7. 'Is he friendly with other dogs?'

8. 'He's the reason my camera roll is always full.'

9. 'Her favorite person is whichever one is holding the food.'

10. 'Did you remember to give him his heartworm medication?'

Top 10 prompts most aligned with NEGATIVE PC3 direction:

1. 'What is the best way to teach my dog to weave through poles?'

2. 'What is the best way to teach my dog to ring a bell to go outside?'

3. 'How do I teach my dog to balance objects on its nose?'

4. 'How do I teach my dog to find hidden treats?'

5. 'How do I teach my dog to push a ball with its nose?'

6. 'How do I teach my dog to carry objects?'

7. 'The dog's joy was palpable as he ran freely on the sandy beach for the very first time.'

8. 'How do I teach my dog to fetch specific items?'

9. 'The sled dog team worked in perfect harmony, pulling the heavy sled across the frozen tundra.'

10. 'How do I teach my dog to open and close doors?'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC4 'dog' direction (Layer 15) ---

Projection values:

Top positive #1: 2.4453

Top positive #2: 2.4395

Top positive #3: 2.4082

Top negative #1: -2.3262

Top negative #2: -2.1758

Top negative #3: -2.1035

Top 10 prompts most aligned with POSITIVE PC4 direction:

1. 'How do I teach my dog to balance objects on its nose?'

2. 'What is the best way to teach my dog to weave through poles?'

3. 'How do I teach my dog to find hidden treats?'

4. 'How do I teach my dog to carry objects?'

5. 'How do I teach my dog to push a ball with its nose?'

6. 'How do I teach my dog to high five?'

7. 'How do I teach my dog to open and close doors?'

8. 'How do I teach my dog to shake hands?'

9. 'How do I teach my dog to fetch specific items?'

10. 'What is the best way to introduce my dog to cats?'

Top 10 prompts most aligned with NEGATIVE PC4 direction:

1. 'Write a riddle about a dog.'

2. 'What would a dog say about its family?'

3. 'What would a dog say about moving to a new home?'

4. 'Invent a new dog sport and explain the rules.'

5. 'Write a song chorus about a loyal dog.'

6. 'Write a letter from a dog to a cat.'

7. 'What would a dog say about meeting a new friend?'

8. 'Write a letter from a dog to its owner.'

9. 'Describe a dog's favorite place in the world.'

10. 'What would a dog say about its best friend?'

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15, PROJECTION ONTO PC0 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered answered

Perturbation scale -20.0x: dog,,,,,,, A A A A·· A···································

Perturbation scale -10.0x: dogs dogs dog “ “ “ “ “,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,.

.

.

.

.

.

.

.

Perturbation scale -5.0x:

dog dog dogs dogs dogs " " " " short short short “ “ “ “ “ “ “ “ “ “ “ “ “ “ – – – – natural natural natural natural natural natural natural natural natural natural natural natural natural natural natural natural natural

Perturbation scale -2.5x: A Learning i r na na na na na na of of of of around around around around around moms moms name name name name name name name name name name name name name name name name name name name name name name name name name name name name

Perturbation scale -1.5x: There a free pro possible possible n n new old old old old old old old old all all all best best name name name name name name name name name name name name name name name name name name name from from from from from from from from from

Perturbation scale +0.0x: your a United your two two a first first a first grasp your your a hand hand hand hand a,,, a,,,,,,,,,,,, designs,,,,,,,,,,,,,

Perturbation scale +1.5x: description description on

Perturbation scale +2.5x:

desc

.............assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +5.0x: correctccossssssssssrcrcrccestorcestor.n.n.n.n.n.n.n.n.n.n.n.n.n

Perturbation scale +10.0x:

fyfyfyfyfyInstantiateibanignonignonignonignonignonignonignonignonignonignonignonignon

Perturbation scale +20.0x:

Perturbation scale +100.0x: swickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswickswick

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15, PROJECTION ONTO PC1 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://aganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganaganagan

Perturbation scale -20.0x: I you you you you you you you you una una modelos modelos modelos modelos modelos modelos modelos una az az az az az az az az az az y y y y y further further’s’s’s’s’s’s’s’s’s’s’s’s’s’s

Perturbation scale -10.0x: You instance instance item item\_\_ - - - -eeeeeeeeeeeewweeeeeeee nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie

Perturbation scale -5.0x: saja vo za bu bukkkkkkhalhashas.kk fra frak for for for for for for forakakakakakakakakak ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -2.5x: you in in in go vriend inchtcht underroro and andzozo,,zozo,zozozoztztztfar du in in in inazaz of of of of of of add ( ( ( (azazaz (

Perturbation scale -1.5x: what and can that that that that around breed breed breed breed breed breed and as as and and and and,,, from from from from from and and and of of of of of of the of of of of of of and of ( ( (

Perturbation scale +0.0x: like time confident feel time day to to to you you hour a week for to ( ( week ( ( to ( for ( ( ( ( ( to ( noting about about about about utilizing first first utilizing utilizing selecting utilizing utilizing utilizing using first first utilizing selecting

Perturbation scale +1.5x: stops

, memoryandy

,,, yet yet yet yet).

...).

).

-t-t-t/d/d/d/d/d/d/d/d/d/d'd'd'd'd'd/l/l as as as as as

Perturbation scale +2.5x: pull time time,antry

lylyly ,,,,, as as as as as as as as as as as as as as as as as as as as as as as as as as as as as as as

Perturbation scale +5.0x:

Now Now Now Now of of of of of of of of of of of of of List of of List List List List List List List List List List Subject Subject Subject Subject Subject Subject Subject Subject Subject Subject Subject

Perturbation scale +10.0x:

By By By By By By By By By By By By By By By By By Back Back | | | | | | | | | | | | | | | | | | | | | | | |

Perturbation scale +20.0x: without without without without ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,

Perturbation scale +100.0x: ://://://://:// Angeles Angeles Angeles Angeles wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel wheel

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15, PROJECTION ONTO PC2 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: ://iloiloilo NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC NBC

Perturbation scale -20.0x: ededededededededededededededededededed April April April April April by by by by by by by by by by by by by by March March March March by by by by by by by by

Perturbation scale -10.0x:

and andionion.antantantantantantishingishingishingishingishingishingishingishingundundundightightadeadeadeadeadeadeadeadeadeadeadeadeadeickickickickadeadeadeadeadeadeadeize

Perturbation scale -5.0x: their as a a a a a a a a a the the the theusususususususususillsillsillsillsillsillsillsillsillsillsillsillsills the the the the the the the the the the the the the

Perturbation scale -2.5x: noted a a a one one one one the the the the the the a the the the the the the the the the the the dog the dog dog dog dog dog dog the the dog the dog the all all all all all all all allentsents

Perturbation scale -1.5x: tasked found found found one one one one one one the the up up up up the the the the the the the the the the the the the the the the the the the the the the the the the the the????izeizeize

Perturbation scale +0.0x: to from your time time time time time time time ( time,,,,,,,,,,,,,,,, to to to to to to to to to to to of to tosto of tostosto to to to

Perturbation scale +1.5x: breed breeds,,,::::::::::::,,,,,,,,stostostostostostostostostostostostostostostostostostostostostostostostosto

Perturbation scale +2.5x: breed,,,,,,,:::,,,,,,,,,,,,,,,stostostostostostostostostostostostostostostostostostostostostostostosto

Perturbation scale +5.0x:

:,,,,,:::.............................. ha ha ha ha ha ha ha ha

Perturbation scale +10.0x: bre

( ( ( ( .................

Perturbation scale +20.0x:

spr spr spr spr spr S S S S S S S S S s s s s s s s s s s zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu ja ja ja ja ja ja ja ja

Perturbation scale +100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15, PROJECTION ONTO PC3 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: ezezezezezezezezezezezseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseauseau

Perturbation scale -20.0x:

|

|

|

|

|

|

|

|

|

zu zu zu zu zu zu zu zu zu zu zu zu zu zu zu

Perturbation scale -10.0x:

in in in to to to to in in in dans dans dans dans dans dans dans dans L L L L L L L L L L to to to to to among among among

Perturbation scale -5.0x:

( (

to to and

-

-

to to to to to to to to to to to to to to to to to to to a, a a a, a a,,,,,,,,,

Perturbation scale -2.5x: short short short short short short one one one other in in existing and from from from from to to to to to zu zu zu zu zu zu zu zu zu zu zu zu zu zu,,,,,,,,,, zu

Perturbation scale -1.5x: choice short short short dogs servers sound in in in to to to suggested suggested suggested recommended to to for for for for for for for to to zu zu zu,,,,,,,,,,,,,,,,

Perturbation scale +0.0x: two polls topic topic, first first first this this a, a time time not not time a in,... at,... at. at at at at at at outside there there; there;; there; there; in

Perturbation scale +1.5x: on on like like like for, for for than.. ( (URURURURURNAMEARRNAMENAMENAMENAMENAMENAMENAMENAMENAMENAMENAMENAMEISTISTISTISTISTISTIST-startassistantassistantassistant-start-start-start-startassistantassistant

Perturbation scale +2.5x: on under apart know for for for for... ( ( ( ( ( (TYTYTYTYTYTYTYISSISS starters starters starters starters starters starters starters starters starters starters starters starters starters starters starters starters starters starters starters starters\_start starters\_start\_start

Perturbation scale +5.0x: fact fact found found. for....... with with with with with with with..... in in in in in in in in in in in in in in in in in in in in in in in in in

Perturbation scale +10.0x: , on on on on with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with with

Perturbation scale +20.0x: breeds about about about about about about ( ( ( ( ( ( (’s’s’s’s’s’s’s’s’s’s’s’s’s’s’s’s’s’s is is is is is is is is is is is is is is is is is is

Perturbation scale +100.0x: ://://://://://://://://

.

.

.

.

.

.

.

.

.

ingroupingroupingroupingroupingroupingroupingroup.

.

ingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroupingroup

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15, PROJECTION ONTO PC4 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: ://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://://

Perturbation scale -20.0x: ofenceenceenceuceuceuceuceuceuceuceuceuceuceuceuceudeudeudeudeudeuteuteuteuteuteuteuteuteuteuteuteuteiateiateiateiateiateiateiateiateenhaenhaenhaenhaenhaenhaenhaenhaenha

Perturbation scale -10.0x: of connection connection of of of of of of of of of of ofcececececececececececececececececececece3333333333333333

Perturbation scale -5.0x: to to of of ofrrrururigrationigrationloadloadloadloadedloadable).ableableableableableableableableableableableableableinsinsinsinsinsinsinsinsinsinsinsinsinsinsinsinsinsins

Perturbation scale -2.5x: that to feedlessarterrytrrationunnelrationersersersers onlyeeeeeoldereolderolderolderolderireireireireorkorkorkorkorkorkorkorkorkorkorkorkorkorkorkorkorkorkorkork

Perturbation scale -1.5x: a owners two different many the many many conversation theershro theermopop onlyersersers onlyopopestestestestersestestseseseseebrebrebrebrebrebrebrebrebrebrebrebrebrebrebre

Perturbation scale +0.0x: to to that different of knowledge different a different and a, the it istr is

tr a is, translatesdifferent,,,,, takes takes,, ( on ( ( ( to ( ( – – from ( – ( - -

Perturbation scale +1.5x: help a with

are. are are are with from from from from from from from from from,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,

Perturbation scale +2.5x: in with with with.. are are are are with from from from from from from..,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,

Perturbation scale +5.0x: in................................................

Perturbation scale +10.0x:

advice advice....................... Key Key Key Key Key Key Key Key Key Key Key Key Key Key Key Key Key Key Key Key Key Key Key Key

Perturbation scale +20.0x: assistantassistantassistantassistant‘‘‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ ‘ “ ‘ “ “ “ “ “ “ “ “ “ “ “

Perturbation scale +100.0x: imanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimanimaniman

================================================================================

================================================================================

--- ORTHOGONAL PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 15 ---

================================================================================

Perturbing along first orthogonal direction (PC133)...

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along orthogonal direction ---

Perturbation scale -100.0x: chnerchnerchnerchnerchnerchnerchnerchnerwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwlwl

Perturbation scale -20.0x: Two aparturenurenurenurenurenurenurenurenurenurenurenurenurenurenayersurenayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayersayers

Perturbation scale -10.0x: There are two different dog breeds that can be used for the following:

1: There are two different dog that are used for the following:

1: There are two different dog that are used for the following:

1: There are two different dog that are used for the following:

1: There are two:

Perturbation scale -5.0x: Here are descriptions of two different dog breeds:

1. \*\*Bulldog Breed Description:\*\*

The bulldog is a small to medium-sized dog breed known for its distinctive physical characteristics and gentle nature. Key features of this breed include:

- A sturdy, compact body with a short, compact stature.

- A flat, wrinkled face with a

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds for you. Here are the descriptions of the 'Bulldog' and the 'Poodle'.

\*\*1. Bulldog\*\*

The Bulldog is a sturdy and compact dog breed known for its distinctive appearance. They have a short, easy-to-maintain coat that is usually brindle,

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and elegant dog breed known for its distinctive curly coat. There are three main sizes of Poodles: Toy, Miniature, and Standard. They are highly trainable and often excel in obedience and agility competitions. Poodles

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed that originated in England. They are known for their speed, agility, and gentle nature. Greyhounds are often referred to as "40 mph couch potatoes" because they are built for speed, but they also

Perturbation scale +2.5x: Let's take a look at the Poodle and the Beagle.

\*\*1. Poodle\*\*

The Poodle is a highly intelligent and versatile dog breed that originated in Germany. They are known for their distinctive curly coat, which requires regular grooming to prevent matting. Poodles come in three sizes: Toy, Miniature, and Standard.

\*

Perturbation scale +5.0x: Let's take the Beagle and the Siberian Husky as examples.

\*\*Beagle\*\*

The Beagle is a friendly and energetic dog breed that originated in England. They are known for their distinctive baying howl, which is often used for hunting. Beagles are relatively small, with a height of 10-15 inches (25

Perturbation scale +10.0x: A Pinger, (named) named, a Pétador, (a) a Pétur, (a) a Pét, (a) a P (a) a P (a) a P ( a) a ( a) a ( a) a ( a) a ( a) a ( a) a ( a

Perturbation scale +20.0x:

- A Ad

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

Roger

Roger

Roger

Roger

Roger

Roger

Roger

Roger

Perturbation scale +100.0x: Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blow Blowbynbynbynbynbynbyn

================================================================================

================================================================================

--- ABLATION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 15 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

All PCs (centroid only):

All PCs except largest (PC0 only):

All PCs except largest two (PC0+PC1):

================================================================================

################################################################################

### STARTING ANALYSIS FOR LAYER 31 ###

################################################################################

Extracting activations from layer 31...

Extracting activations: 100%|███████████████████████████| 1944/1944 [00:42<00:00, 45.38it/s]

Extracting activations from layer 31...

Extracting activations: 100%|█████████████████████████████| 247/247 [00:05<00:00, 45.63it/s]

Concept 'dog': Found 143 effective eigenvectors out of 1944 (threshold: 0.0794)

Concept 'lion': Found 30 effective eigenvectors out of 247 (threshold: 0.4233)

================================================================================

Running experiments for system prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User prompt: 'The dog was running around the park. It was a labrador.'

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31, AXIS: 0 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC0 ---

Perturbation scale -100.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -20.0x: Derenteenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteugenteug

Perturbation scale -10.0x: Der Hund tobte im Park herum. Es war ein Labrador.

Hier ist die genaue Übersetzung:

- tobte = tobiet, tobieten, tobiet, tobiet, tobiet, tobiet, tobiet, tobiet, tobiet, tobiet, tobiet, tobiet, tobiet,

Perturbation scale -5.0x: Der Hund lief um den Park herum. Es war ein Labrador.

Hier ist eine genaue Übersetzung:

- "The dog" wird direkt mit "Der Hund" übersetzt, da der Hund männlich ist, aber da der Hund kein bestimmter, sondern ein allgemeiner wird, wird der Artikel verwendet, der

Perturbation scale -2.5x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale -1.5x: Der Hund lief um den Park herum. Es war ein Labrador.

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis: "Hündin" ist die weibliche Form von "Hund", da "Hund" in der deutschen Sprache neutral ist.

Perturbation scale +2.5x: Die Hunde lief um den Park herum. Es war ein Labrador.

Here's a breakdown of the translation:

- "The dog" translates to "Die Hunde" (note: "Hunde" is the plural form of "Hund", which is "dog" in English)

- "was running" translates to "lief"

Perturbation scale +5.0x: Die Hunde lief um den Park herum. Es war ein Labrador.

Here's a more detailed translation with a bit more context:

1. "The dog" - "Die Hunde" (note: "Hunde" is a collective noun in German, similar to "dogs" in English)

2. "was running" - "l

Perturbation scale +10.0x: assistant

assistant

Die Hunde lief auf dem Rasen auf dem Park auf dem Rasen auf dem Rasen auf dem Rasen auf dem Rasen auf dem Rasen auf dem Rasen auf dem Rasen auf dem Rasen auf dem Rasen auf dem Rasen auf dem Rasen auf dem Rasen auf dem Rasen auf dem Ras

Perturbation scale +20.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +100.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31, AXIS: 1 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC1 ---

Perturbation scale -100.0x: translated translationtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -20.0x: Die translated translation into German translates into:

Der Hund ran um das Gelände herum Die translation translates into translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation translation

Perturbation scale -10.0x: Die Hündin lief um den Park herum. Sie war ein Labrador.

(Note: In German translation, the translation of the English word 'dog' into German is translated into the feminine noun 'die Hündin' if the translation is referring to a female dog, translated into the masculine noun 'der Hund' if the translation is referring to

Perturbation scale -5.0x: Die Hündin lief um den Park herum. Sie war ein Labrador.

Hinweis:

- Ich habe "die Hündin" verwendet, da die Übersetzung von "dog" im Deutschen Geschlecht abhängig ist. Wenn Sie den Text auf einen männlichen Hund beziehen möchten, verwenden Sie

Perturbation scale -2.5x: Die Hündin lief um den Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "die Hündin" ein weibliches Tier bezeichnet, während "der Hund" ein männliches Tier bezeichnet.

- "was running"

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "die Hündin" ein weibliches Tier bezeichnet, während "der Hund" ein männliches Tier bezeichnet.

- "was running" wird

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis: "Hündin" ist die weibliche Form von "Hund", da "Hund" in der deutschen Sprache neutral ist.

Perturbation scale +2.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis: "Hündin" ist die weibliche Form von "Hund" in der deutschen Sprache.

Perturbation scale +5.0x: Der Hund lief im Park herum. Es war ein Labrador.

Perturbation scale +10.0x: Der Hund lief im Park herum. Es war ein Labrador.

Perturbation scale +20.0x: Uebrigens "u. = "und "u = "u "u = "u "u = "u "u = "u "u = "u "u = "u "u = "u "u = "u "u = "u "u = "u "u = "u "u = "u

Perturbation scale +100.0x: olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31, AXIS: 2 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC2 ---

Perturbation scale -100.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -20.0x: Das Hünd ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -10.0x: Der Hund lief im Park herum. Es war ein Labrador.

(Note: In German, "es war" is a more formal or a more general translation of "it was". In a more informal conversation, a more natural translation would be "es war ein Labrador-Hund" or "es war ein Labrador-Rüde" (if

Perturbation scale -5.0x: Der Hund lief im Park herum. Es war ein Labrador.

Here is a more natural translation in German, as the original sentence is a bit informal:

Der Hund lief im Park herum. Es handelte sich um einen Labrador.

Or in a more formal translation:

Der Hund lief im Park herum. Es war ein Labrador-H

Perturbation scale -2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wurde in "Die Hündin" übersetzt, da "Hündin" (weibliche Form) verwendet wird, um den Geschlecht des Tiers zu bestimmen.

- "It was

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird in diesem Kontext als "Hündin" (weiblich) übersetzt, da "Hund" (männlich) und "Hündin" (weiblich) in der deutschen Spr

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "die Hündin" ein weibliches Tier bezeichnet, während "der Hund" ein männliches Tier bezeichnet.

- "was running" wird

Perturbation scale +2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis: Ich habe "Hündin" verwendet, da die Übersetzung "der Hund" nicht genau ist, da die ursprüngliche Aussage "the dog" verwendet hat, was sowohl männlich als auch weiblich sein kann

Perturbation scale +5.0x: Die Hund war um das Park herumgelaufen. Es war ein Labrador.

Hier ist eine genauere Übersetzung:

- "The dog" wird direkt mit "der Hund" übersetzt, da "der Hund" männlich ist und die Geschlecht des Hundes nicht angegeben ist.

- "was running" wird

Perturbation scale +10.0x: Die Hund lief um das Park herum. Es war ein Labrador.

Hier ist eine genauere Übersetzung:

- Die Hund: Das Tier wird männlich oder weiblich angegeben, daher ist die korrektere Übersetzung: Der Hund oder Die Hundin.

- lief um das Park herum: Die korrektere

Perturbation scale +20.0x: Die Hund lief�entlich�����������������������������������������������������������������

Perturbation scale +100.0x: .ease:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///:///

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31, AXIS: 3 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC3 ---

Perturbation scale -100.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -20.0x: Hier ist die genaue Übersetzung:

"Der Hund lief im Park herum. Er war ein Labrador."

(Alternativer, aber weniger formal, Satzbau: "Im Park lief ein Labrador, also, 'Der Hund, ein Labrador, lief im,''''''

Perturbation scale -10.0x: Hier ist die Übersetzung:

Der Hund lief im Park herum. Es war ein Labrador.

Perturbation scale -5.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis: "Hündin" wird verwendet, da "dog" in diesem Kontext weiblich sein kann. Wenn Sie "dog" als neutrales Wort verwenden, wäre die Übersetzung "Der Hund lief im Park herum. Er war

Perturbation scale -2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis: "Hündin" wird verwendet, da "dog" im Originaltext weiblich sein kann. Wenn Sie wissen, dass es sich um einen männlichen Hund handelt, verwenden Sie "Hund" anstelle von "Hündin".

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis: "Hündin" wird verwendet, da "labrador" ein weibliches Tier bezeichnen kann, aber "Hund" wird auch verwendet, wenn das Geschlecht nicht bekannt ist.

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "die Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, wird "der Hund" verwendet.

-

Perturbation scale +2.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "die Hündin" ein weiblicher Hund ist. Wenn es sich um einen männlichen Hund handelt, wird "der Hund" verwendet.

-

Perturbation scale +5.0x: Die Hündin lief um den Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wurde als "Die Hündin" übersetzt, da das Geschlecht der Hündin nicht angegeben ist. Wenn es ein männlicher Hund ist, würde es "Der Hund" oder "Der Labrador

Perturbation scale +10.0x: Die Hündin lief um das Park herum. Sie war eine Labradorin.

(Note: Die Hündin is used as the translation for "the dog" as it is female. If you want to translate it as "the dog" regardless of gender, you could use "Das Tier" or simply "Das Hund". However Die Hünd

Perturbation scale +20.0x: Die Frage nach der Übersetzung ist leicht das Problem das das Text ja nicht das ist das übersetzt werden muss das ist das das übersetzt werden muss ist das Text das das übersetzt werden muss das ist das das übersetzt werden muss das ist das das übersetzt werden muss das ist das das übersetzt werden

Perturbation scale +100.0x: Das/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets/assets

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31, AXIS: 4 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along PC4 ---

Perturbation scale -100.0x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -20.0x: Die Hund lief rundherum durch den Park\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa\xaa

Perturbation scale -10.0x: Die Hund lief im Park herum. Es war ein Labrador.

Hier ist eine genauere Übersetzung:

Die Hund lief im Park herum.

Hier ist eine genauere Übersetzung:

Die Hund lief im Park herum.

Hier ist eine genauere Übersetzung:

Der Hund lief im Park herum.

Hier ist eine

Perturbation scale -5.0x: Die Hund war um den Park herumgelaufen. Es war ein Labrador.

Hier ist eine genauere Übersetzung:

- "The dog" wird direkt übersetzt mit "Der Hund".

- "was running around" wird übersetzt mit "war umhergelaufen".

- "the park" wird direkt übers

Perturbation scale -2.5x: Die Hündin lief im Park herum. Sie war eine Labradorhündin.

Hinweis:

- "The dog" wird hier mit "Die Hündin" übersetzt, da man nicht genau weiß, ob es sich um einen männlichen oder weiblichen Hund handelt. Wenn man weiß, dass es sich um

Perturbation scale -1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "die Hündin" sich auf eine weibliche Hündin bezieht. Wenn Sie "The dog" als neutrale Form verwenden möchten,

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird in diesem Kontext als "Die Hündin" übersetzt, da "Hündin" (weibliche Form) verwendet wird, um eine bestimmte Hunderasse zu beschreiben.

-

Perturbation scale +2.5x: Die Hündin lief im Park herum. Es war ein Labrador.

Hinweis:

- "The dog" wird in diesem Kontext als "Die Hündin" (weiblich) oder "Der Hund" (männlich) übersetzt, da "labrador" (weiblich) ist.

Perturbation scale +5.0x: Die Hündin ran in der Park. (Note: 'Hündin' is the correct translation for 'the dog' in this context, as 'Hund' is the general term for 'dog' and 'Hündin' is the female form of 'Hund'. 'Hund' is used for a male dog,

Perturbation scale +10.0x: "Der Hund (wurde) (um) (herum) (im) (Park) (gelaufen) (, (es) (war) (ein) (Labrador (, (wurde ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +20.0x: " - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Perturbation scale +100.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC0 'dog' direction (Layer 31) ---

Projection values:

Top positive #1: 29.4219

Top positive #2: 29.2812

Top positive #3: 29.0469

Top negative #1: -5.2930

Top negative #2: -5.2578

Top negative #3: -5.2344

Top 10 prompts most aligned with POSITIVE PC0 direction:

1. 'How do I teach my dog to jump over obstacles?'

2. 'How do I help my dog with fear of water?'

3. 'How do I help my dog with fear of other animals?'

4. 'How do I teach my dog to fetch?'

5. 'How do I help my dog with fear of strangers?'

6. 'How do I teach my dog to shake hands?'

7. 'How do I teach my dog to fetch specific items?'

8. 'How do I teach my dog to play dead?'

9. 'How do I teach my dog to roll over?'

10. 'How do I help my dog with fear of being alone?'

Top 10 prompts most aligned with NEGATIVE PC0 direction:

1. 'The dog's gentle and patient nature with the new litter of kittens was a beautiful sight.'

2. 'The dog's thick, water-repellent coat makes him a natural swimmer and retriever.'

3. 'The dog's bond with the family's cat was an unusual but beautiful friendship to behold.'

4. 'He is a very even-tempered and good-natured animal.'

5. 'The dog's presence in the quiet library had a surprisingly calming effect on the students during finals week.'

6. 'The farm dog worked tirelessly from dawn until dusk, a loyal and indispensable helper.'

7. 'That dog is the heart of our family, the furry hub around which we all revolve.'

8. 'The dog's companionship has brought immeasurable happiness to my life.'

9. 'The therapy dog visited the nursing home, bringing smiles and comfort to the residents.'

10. 'Although he was a very small dog, his personality filled the entire house with joy and chaos.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC1 'dog' direction (Layer 31) ---

Projection values:

Top positive #1: 13.0859

Top positive #2: 12.9453

Top positive #3: 12.9062

Top negative #1: -10.6172

Top negative #2: -10.5938

Top negative #3: -9.5312

Top 10 prompts most aligned with POSITIVE PC1 direction:

1. 'If a dog could write a thank-you note to its owner, what would it say?'

2. 'If a dog could write a review of its owner, what would it say?'

3. 'If a dog could give advice to humans, what would it say?'

4. 'If a dog could write a thank-you note, what would it say?'

5. 'If dogs could play an instrument, which would they choose?'

6. 'If dogs could write a book, what would the title be?'

7. 'If dogs could vote, what issues would matter to them?'

8. 'If a dog could invent a holiday, what would it celebrate?'

9. 'What is the most important rule in a dog's world?'

10. 'If a dog could write a song, what would the lyrics be?'

Top 10 prompts most aligned with NEGATIVE PC1 direction:

1. 'My camera roll is 90% pictures of my dog.'

2. 'I prefer a mixed-breed dog from a shelter.'

3. 'That Chihuahua is shivering, maybe it's cold.'

4. 'My dog has a very clear 'I'm sorry' face he makes after he knows he has done something wrong.'

5. 'I think my dog might be sick because he hasn't touched his food all day.'

6. 'My neighbor's golden retriever dog is incredibly well-trained and never jumps on guests.'

7. 'The rescue dog was initially very timid, but now he demands belly rubs from everyone he meets.'

8. 'The pet sitter needs a detailed list of the dog's routine and needs.'

9. 'That dog is surprisingly delicate when taking a treat from my hand, using only his lips.'

10. 'My dog has a very specific routine that involves napping in three different sunny spots throughout the day.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC2 'dog' direction (Layer 31) ---

Projection values:

Top positive #1: 14.2734

Top positive #2: 14.0859

Top positive #3: 13.6484

Top negative #1: -9.0625

Top negative #2: -8.8672

Top negative #3: -8.6562

Top 10 prompts most aligned with POSITIVE PC2 direction:

1. 'How do I help my dog with fear of other animals?'

2. 'How do I help my dog with fear of water?'

3. 'How do I help my dog with fear of strangers?'

4. 'How do I help my dog with fear of being alone?'

5. 'How do I teach my dog to jump over obstacles?'

6. 'How do I teach my dog to stay?'

7. 'How do I teach my dog to shake hands?'

8. 'How do I teach my dog to fetch?'

9. 'How do I teach my dog to fetch specific items?'

10. 'How do I teach my dog to push a ball with its nose?'

Top 10 prompts most aligned with NEGATIVE PC2 direction:

1. 'What would a dog do with a million dollars?'

2. 'My dog's level of excitement is directly proportional to the crinkliness of the treat bag.'

3. 'You look like the dog that caught the canary.'

4. 'I think he dreams of chasing rabbits.'

5. 'If dogs could write laws, what would be the first law?'

6. 'What would a dog say about growing old?'

7. 'My dog's happiness is my top priority, which is why his dinner has more vegetables than mine.'

8. 'What would a dog say about its own reflection?'

9. 'I cannot believe that dog managed to open the refrigerator and eat the entire block of cheese.'

10. 'He's snoring louder than my uncle.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC3 'dog' direction (Layer 31) ---

Projection values:

Top positive #1: 9.6797

Top positive #2: 9.4062

Top positive #3: 9.3672

Top negative #1: -18.5312

Top negative #2: -14.3984

Top negative #3: -14.3906

Top 10 prompts most aligned with POSITIVE PC3 direction:

1. 'What are the best dog breeds for hot climates?'

2. 'What are the best dog breeds for cold climates?'

3. 'What are the most common signs of illness in dogs?'

4. 'What are the signs of separation anxiety in dogs?'

5. 'What are the signs of heatstroke in dogs?'

6. 'What are the most common health issues in dogs?'

7. 'What are the best dog breeds for apartment living?'

8. 'What are the best dog breeds for companionship?'

9. 'What are the most common causes of bad breath in dogs?'

10. 'What are the best dog breeds for active people?'

Top 10 prompts most aligned with NEGATIVE PC3 direction:

1. 'What would a dog say about the invention of the leash?'

2. 'How would a dog describe the feeling of being adopted?'

3. 'How do I teach my dog to push a ball with its nose?'

4. 'What would a dog say about going to the vet?'

5. 'How do I teach my dog to balance objects on its nose?'

6. 'How would a dog explain the concept of loyalty?'

7. 'What would a dog say about its first walk?'

8. 'How do I teach my dog to find hidden treats?'

9. 'How do I teach my dog to carry objects?'

10. 'What would a dog say about being left alone?'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC4 'dog' direction (Layer 31) ---

Projection values:

Top positive #1: 9.0078

Top positive #2: 8.2266

Top positive #3: 8.2031

Top negative #1: -11.0859

Top negative #2: -10.7891

Top negative #3: -10.3984

Top 10 prompts most aligned with POSITIVE PC4 direction:

1. 'Have you ever seen a happier puppy?'

2. 'Could you grab the poop bags?'

3. 'Is he allowed to have a bone?'

4. 'How do I help my dog with fear of water?'

5. 'Why must this dog always choose my clean laundry as the perfect spot for a nap?'

6. 'How do I help my dog with fear of other animals?'

7. 'Did you remember to give him his heartworm medication?'

8. 'Are there any off-leash parks nearby?'

9. 'I think I'm more excited for his birthday than he is.'

10. 'How do I help my dog with fear of loud noises?'

Top 10 prompts most aligned with NEGATIVE PC4 direction:

1. 'What would a dog's diary entry look like after a day at the park?'

2. 'Write a letter from a dog to a cat.'

3. 'Describe a dog's reaction to seeing snow for the first time.'

4. 'Write a letter from a dog to its owner.'

5. 'Describe a futuristic city designed for dogs.'

6. 'What would a dog say about its first walk?'

7. 'How would a dog describe the feeling of being adopted?'

8. 'How would a dog explain friendship?'

9. 'What would a dog say about moving to a new home?'

10. 'How would a dog explain the concept of loyalty?'

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31, PROJECTION ONTO PC0 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -20.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -10.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -5.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -2.5x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -1.5x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale +0.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +1.5x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +2.5x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +5.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +10.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +20.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +100.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31, PROJECTION ONTO PC1 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -20.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -10.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -5.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -2.5x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -1.5x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale +0.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale +1.5x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale +2.5x: translatedolentranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale +5.0x: olenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolen

Perturbation scale +10.0x: olenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolen

Perturbation scale +20.0x: olenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolen

Perturbation scale +100.0x: olenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolen

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31, PROJECTION ONTO PC2 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -20.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -10.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -5.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -2.5x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -1.5x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +0.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +1.5x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +2.5x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +5.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +10.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +20.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +100.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31, PROJECTION ONTO PC3 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -20.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -10.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -5.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -2.5x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -1.5x: �数>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale +0.0x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +1.5x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +2.5x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +5.0x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +10.0x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +20.0x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +100.0x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31, PROJECTION ONTO PC4 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Perturbation scale -100.0x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -20.0x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -10.0x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -5.0x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -2.5x: ":[{

":[{

":[{

":[{

( (":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(":[{

(

Perturbation scale -1.5x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +0.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +1.5x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +2.5x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +5.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +10.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +20.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +100.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

================================================================================

================================================================================

--- ORTHOGONAL PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31 ---

================================================================================

Perturbing along first orthogonal direction (PC143)...

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

Original model completion: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

--- Perturbing final token activation along orthogonal direction ---

Perturbation scale -100.0x: 건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건

Perturbation scale -20.0x: EIF건erpelageEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIFEIF

Perturbation scale -10.0x: Die Dogge lief herum an dem Park. Es war ein Labrador.

Perturbation scale -5.0x: Die Hunde lief umher im Park. Es war ein Labrador.

Perturbation scale -2.5x: Die Hündin lief um den Park herum. Sie war ein Labrador.

Hinweis:

- "dog" wird in diesem Kontext als "Hündin" übersetzt, da es sich um ein weibliches Tier handelt.

- "labrador" wird einfach als "Labrador" übersetzt,

Perturbation scale -1.5x: Die Hündin lief um den Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "dog" in diesem Kontext ein weibliches Tier bezeichnet.

- "It was a labrador" wird als "Sie war ein

Perturbation scale +0.0x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis:

- "The dog" wird als "Die Hündin" übersetzt, da "Hündin" ein weibliches Tier bezeichnet. Wenn Sie "The dog" als männliches Tier meinen, sollten Sie "Der Hund" verwenden

Perturbation scale +1.5x: Die Hündin lief im Park herum. Sie war ein Labrador.

Hinweis: "Hündin" ist die weibliche Form von "Hund", da "Hund" in der deutschen Sprache neutral ist.

Perturbation scale +2.5x: Die Hündin rannte im Park herum. Sie war ein Labrador.

Hinweis: "Hündin" ist die weibliche Form von "Hund", da "Hund" in der deutschen Sprache neutral ist.

Perturbation scale +5.0x: Der Hund ran im Park rum. Es war ein Labrador.

Hier ist eine genauere Übersetzung:

- "The dog" = "Der Hund"

- "was running around" = "ran im Park rum"

- "the park" = "im Park"

- "It was" = "Es war"

- "a labrador

Perturbation scale +10.0x: Hier ist die Übersetzung:

Der Hund rannte im Park rum. Es war ein Labrador.

Oder:

Im Park rannte ein Labrador rum.

Beziehungen:

- Im Park: im = im; Park = Park (kein Artikel)

- rum: im = im; rum = im; im = im; rum =

Perturbation scale +20.0x: ruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruh

Perturbation scale +100.0x: ruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruh

================================================================================

================================================================================

--- ABLATION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31 ---

================================================================================

System Prompt: 'You are a helpful assistant. Please translate the following text accurately from English into German:'

User Prompt: 'The dog was running around the park. It was a labrador.'

All PCs (centroid only):

All PCs except largest (PC0 only):

All PCs except largest two (PC0+PC1):

================================================================================

================================================================================

Running experiments for system prompt: 'You are a helpful assistant.'

User prompt: 'Please describe two different dog breeds.'

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31, AXIS: 0 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC0 ---

Perturbation scale -100.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -20.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -10.0x: Let me describe the French Bulldog and the Border Roughsetter, however, the Border Roughsetter is actually the Rough Border Terrier, the Rough Border Terrier and the Border Terrier have the Border Roughsetter as an alias. Rough Border Terriers have the alias Rough Bordersetter, Rough Bordersetter, Rough Bordersetter Rough Border Terrier Rough Border

Perturbation scale -5.0x: Let me describe the French Bulldog and the German Shepherd.

\*\*French Bulldog:\*\*

The French Bulldog, also known as the Bouledogue Français, is a small, adaptable, and affectionate breed. Originating from England, they were later developed and became popular in France. French Bulldogs have:

- Weight: 16-28

Perturbation scale -2.5x: Let me describe two different dog breeds for you:

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed that is known for its speed and agility. They are a sight hound, bred for hunting and racing, and are often referred to as the "40 mph couch potato" due to their ability to go from zero

Perturbation scale -1.5x: Let me describe two different dog breeds for you:

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting and racing, and are often referred to as the "40 mph couch potato" due to their ability to go from zero to

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and versatile dog breed that originated in Germany. There are three main sizes of Poodles: Toy, Miniature, and Standard. They are known for their distinctive curly, non-shedding coat that requires regular grooming. P

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and versatile dog breed known for its distinctive appearance and low-shedding coat. There are three main sizes of Poodles: Toy, Miniature, and Standard.

\* \*\*Physical Characteristics:\*\* Poodles typically have a curly,

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds.

1. \*\*Labrador Retriever\*\*:

The Labrador Retriever is a popular dog breed known for its friendly, outgoing personality and versatility. They typically weigh between 55-80 pounds (25-36 kg) and stand 21.5-24.5 inches (55-

Perturbation scale +10.0x: I'd be happy to tell you about two different dog breeds.

1. \*\*Beagle\*\* - A Beagle is a small to medium-sized dog breed known for their distinctive triangular ears and sweet expressions on their faces. Beagles typically weigh between 18-30 pounds (8-14 kg) and usually range in length between 13-

Perturbation scale +20.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +100.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31, AXIS: 1 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC1 ---

Perturbation scale -100.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -20.0x: Here are two distinct dog breeds:

translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated translated

Perturbation scale -10.0x: Here are two different dog breeds:

\*\*German Shepherd\*\*

German Shepherds are a medium to large-sized breed that originated in Germany in the late 1800s. They are a highly intelligent breed known for their loyalty, courage, and trainability. German Shepherds typically weigh between 75-95 pounds (34-43 kilograms) and

Perturbation scale -5.0x: Here are two different dog breeds:

\*\*1. Labrador Retriever\*\*

The Labrador Retriever is a popular breed known for its friendly, outgoing, and loyal nature. They are a medium to large-sized dog, typically weighing between 55-80 pounds (25-36 kg) and standing 21.5-24.5 inches (

Perturbation scale -2.5x: Here are descriptions of two different dog breeds:

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to chase small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69 cm)

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are characterized by their:

\* Short, smooth

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are characterized by their:

- Short, smooth

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and elegant dog breed known for its distinctive curly or non-curling coat. They come in various sizes, including Toy, Miniature, and Standard. Poodles are often considered one of the most intelligent dog breeds

Perturbation scale +10.0x: I can describe two different dog breeds for you.

1. \*\*Bulldog\*\*:

The Bulldog, also known as the British Bulldog, is a sturdy and compact dog breed. They have a broad, flat head with a short, pushed-in snout. Their wrinkled skin and prominent underbite give them a unique and lo

Perturbation scale +20.0x: 1. \*\*Bulldog\*\* -

Bulldogs are known for their unique and wrinkled appearance. They have a sturdy, compact body with rolled-over1 /2 /1 /1 /1 /1 /1 /1 /1 /1 /1 /1 /1 /1 /1 /1 /1 /1 /

Perturbation scale +100.0x: 1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen1olen

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31, AXIS: 2 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC2 ---

Perturbation scale -100.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -20.0x: \*\*Bulldog (English or British (Bul ild ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -10.0x: Here are descriptions of two different dog breeds:

\*\*1. The Poodle (Toy Poodle)\*\*

The Poodle is a small, intelligent, and elegant dog breed. They are a popular choice as a companion or show dog. Toy Poodles are a sub-breed of Poodles and are the smallest of the Poodle family. They typically

Perturbation scale -5.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and elegant dog breed that is known for its distinctive, curly coat. They come in a variety of sizes, including Toy, Miniature, and Standard. Poodles are highly trainable and are often used as show dogs, performance

Perturbation scale -2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and elegant dog breed that originated in Germany. They are known for their distinctive curly or wavy coat, which requires regular grooming to prevent matting. Poodles come in a variety of sizes, including Toy, Miniature,

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and elegant dog breed that originated in Germany. They are known for their distinctive curly or wavy coat, which requires regular grooming to prevent matting. Poodles come in a variety of sizes, including Toy, Miniature,

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are characterized by their:

\* Short, smooth

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. Labrador Retriever\*\*

The Labrador Retriever is a popular and friendly dog breed known for its outgoing and affectionate nature. They are medium to large-sized dogs, typically weighing between 55-80 pounds and standing between 21.5-24.5 inches

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. Labrador Retriever\*\*

The Labrador Retriever is one of the most popular dog breeds worldwide, known for its friendly, outgoing, and loyal nature. They are medium to large-sized dogs, typically weighing between 55-80 pounds and standing between 21.5

Perturbation scale +10.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. Labrador Retriever\*\*

The Labrador Retriever, also known simply as Labradors, is one of the most popular breeds worldwide due to its friendly temperament, intelligence, and versatility. Originating from Canada, Labradors were initially bred to retrieve game for hunters

Perturbation scale +20.0x: I'll describe two different dog breeds:

1WunusedZeroWidthSpaceassistantelseaassistantassistantassistantelsea

Perturbation scale +100.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31, AXIS: 3 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC3 ---

Perturbation scale -100.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -20.0x: I can describe the 'Poodle' and the 'Bulldog' to you.

1. \*\*Poodle:\*\*

A Poodle, also known as a 'Caniche' in its native France, is a highly intelligent, elegant, and agile dog. Poodles can come in a wide range of sizes, from the 'Toy

Perturbation scale -10.0x: I can describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek, athletic dog breed known for its speed. They are a sight hound, bred for centuries to hunt small game, particularly rabbits. Greyhounds have a short, smooth coat that can be a variety of colors, including brindle

Perturbation scale -5.0x: I can describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game, and are one of the fastest dog breeds in the world, reaching speeds of up to 45 miles per hour.

Perturbation scale -2.5x: I can describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for hunting small game, particularly rabbits. Greyhounds are characterized by their:

- Short, smooth coats that come in a variety of colors

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are characterized by their:

\* Short, smooth coats that

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically slender and muscular, with a short

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Golden Retriever\*\*

The Golden Retriever is a popular breed known for its friendly and loyal nature. They are a medium to large-sized dog with a thick, double coat that ranges in color from cream to golden brown. Golden Retrievers typically weigh between

Perturbation scale +10.0x: Here are descriptions of two popular dog breeds:

\*\*1. Golden Retriever\*\*

The Golden Retriever is a friendly and loyal breed known for their beautiful golden coats and gentle nature. They are a medium-sized breed with males typically weighing between 65-75 pounds and females weighing between 55-65 pounds. Golden Retrievers are known for

Perturbation scale +20.0x: Here are descriptions of two popular dog breeds:

### German Shepard

German Shepherds are commonly employed as guide dogs for individuals with disabilities and frequently serve as search-and-rescue dogs due to their intelligence and athleticism.

- Height Range:

German Shepherds typically stand between 24 - 26 inches tall.

- Weight Range:

German

Perturbation scale +100.0x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

================================================================================

================================================================================

--- INTERVENTION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31, AXIS: 4 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along PC4 ---

Perturbation scale -100.0x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -20.0x: Here:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description:description

Perturbation scale -10.0x: Here are descriptions of two dog breeds:

1. \*\*Australian Shepherd\*\*

The Australian Shepherd is a medium-sized dog breed known for its intelligence, agility, and loyalty. They typically weigh between 30-50 pounds and stand between 18-23 inches tall at the shoulder.

Australian Shepherds have medium-length coats that come in various colors such as

Perturbation scale -5.0x: Here are descriptions of two different dog breeds:

\*\*1. Labrador Retriever\*\*

The Labrador Retriever is a popular breed known for its friendly, outgoing personality. They are medium-sized dogs with a sturdy build and short coats that come in three colors: black, yellow, and chocolate brown. Adult Labradors typically weigh between 55-

Perturbation scale -2.5x: Here are descriptions of two different dog breeds:

\*\*1. Labrador Retriever\*\*

The Labrador Retriever is a popular breed known for its friendly, outgoing personality. They are medium to large-sized dogs with a sturdy build and a short, smooth coat that comes in three colors: black, yellow, and chocolate. Labradors typically weigh between

Perturbation scale -1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically tall and lean, with a short, smooth

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed that is known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Poodle:\*\*

The Poodle is a highly intelligent and elegant dog breed that is known for its distinctive, curly coat. There are three main sizes of Poodles: Toy, Miniature, and Standard.

- \*\*Physical Characteristics:\*\* Poodles have a muscular

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds for you.

\*\*1. The Poodle:\*\*

The Poodle is a highly intelligent and elegant dog breed that is known for its distinctive, curly coat. There are three main sizes of Poodles: Toy, Miniature, and Standard. Poodles are highly trainable and are often used in

Perturbation scale +10.0x: \*\*1. The Poodle (Toy, 6- 10 in ( 15 - 25 in ( 3 - 4 ( 6 - 14 ( 20 - 40 (

Perturbation scale +20.0x: \*\*1. ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +100.0x: - ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC0 'dog' direction (Layer 31) ---

Projection values:

Top positive #1: 29.4219

Top positive #2: 29.2812

Top positive #3: 29.0469

Top negative #1: -5.2930

Top negative #2: -5.2578

Top negative #3: -5.2344

Top 10 prompts most aligned with POSITIVE PC0 direction:

1. 'How do I teach my dog to jump over obstacles?'

2. 'How do I help my dog with fear of water?'

3. 'How do I help my dog with fear of other animals?'

4. 'How do I teach my dog to fetch?'

5. 'How do I help my dog with fear of strangers?'

6. 'How do I teach my dog to shake hands?'

7. 'How do I teach my dog to fetch specific items?'

8. 'How do I teach my dog to play dead?'

9. 'How do I teach my dog to roll over?'

10. 'How do I help my dog with fear of being alone?'

Top 10 prompts most aligned with NEGATIVE PC0 direction:

1. 'The dog's gentle and patient nature with the new litter of kittens was a beautiful sight.'

2. 'The dog's thick, water-repellent coat makes him a natural swimmer and retriever.'

3. 'The dog's bond with the family's cat was an unusual but beautiful friendship to behold.'

4. 'He is a very even-tempered and good-natured animal.'

5. 'The dog's presence in the quiet library had a surprisingly calming effect on the students during finals week.'

6. 'The farm dog worked tirelessly from dawn until dusk, a loyal and indispensable helper.'

7. 'That dog is the heart of our family, the furry hub around which we all revolve.'

8. 'The dog's companionship has brought immeasurable happiness to my life.'

9. 'The therapy dog visited the nursing home, bringing smiles and comfort to the residents.'

10. 'Although he was a very small dog, his personality filled the entire house with joy and chaos.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC1 'dog' direction (Layer 31) ---

Projection values:

Top positive #1: 13.0859

Top positive #2: 12.9453

Top positive #3: 12.9062

Top negative #1: -10.6172

Top negative #2: -10.5938

Top negative #3: -9.5312

Top 10 prompts most aligned with POSITIVE PC1 direction:

1. 'If a dog could write a thank-you note to its owner, what would it say?'

2. 'If a dog could write a review of its owner, what would it say?'

3. 'If a dog could give advice to humans, what would it say?'

4. 'If a dog could write a thank-you note, what would it say?'

5. 'If dogs could play an instrument, which would they choose?'

6. 'If dogs could write a book, what would the title be?'

7. 'If dogs could vote, what issues would matter to them?'

8. 'If a dog could invent a holiday, what would it celebrate?'

9. 'What is the most important rule in a dog's world?'

10. 'If a dog could write a song, what would the lyrics be?'

Top 10 prompts most aligned with NEGATIVE PC1 direction:

1. 'My camera roll is 90% pictures of my dog.'

2. 'I prefer a mixed-breed dog from a shelter.'

3. 'That Chihuahua is shivering, maybe it's cold.'

4. 'My dog has a very clear 'I'm sorry' face he makes after he knows he has done something wrong.'

5. 'I think my dog might be sick because he hasn't touched his food all day.'

6. 'My neighbor's golden retriever dog is incredibly well-trained and never jumps on guests.'

7. 'The rescue dog was initially very timid, but now he demands belly rubs from everyone he meets.'

8. 'The pet sitter needs a detailed list of the dog's routine and needs.'

9. 'That dog is surprisingly delicate when taking a treat from my hand, using only his lips.'

10. 'My dog has a very specific routine that involves napping in three different sunny spots throughout the day.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC2 'dog' direction (Layer 31) ---

Projection values:

Top positive #1: 14.2734

Top positive #2: 14.0859

Top positive #3: 13.6484

Top negative #1: -9.0625

Top negative #2: -8.8672

Top negative #3: -8.6562

Top 10 prompts most aligned with POSITIVE PC2 direction:

1. 'How do I help my dog with fear of other animals?'

2. 'How do I help my dog with fear of water?'

3. 'How do I help my dog with fear of strangers?'

4. 'How do I help my dog with fear of being alone?'

5. 'How do I teach my dog to jump over obstacles?'

6. 'How do I teach my dog to stay?'

7. 'How do I teach my dog to shake hands?'

8. 'How do I teach my dog to fetch?'

9. 'How do I teach my dog to fetch specific items?'

10. 'How do I teach my dog to push a ball with its nose?'

Top 10 prompts most aligned with NEGATIVE PC2 direction:

1. 'What would a dog do with a million dollars?'

2. 'My dog's level of excitement is directly proportional to the crinkliness of the treat bag.'

3. 'You look like the dog that caught the canary.'

4. 'I think he dreams of chasing rabbits.'

5. 'If dogs could write laws, what would be the first law?'

6. 'What would a dog say about growing old?'

7. 'My dog's happiness is my top priority, which is why his dinner has more vegetables than mine.'

8. 'What would a dog say about its own reflection?'

9. 'I cannot believe that dog managed to open the refrigerator and eat the entire block of cheese.'

10. 'He's snoring louder than my uncle.'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC3 'dog' direction (Layer 31) ---

Projection values:

Top positive #1: 9.6797

Top positive #2: 9.4062

Top positive #3: 9.3672

Top negative #1: -18.5312

Top negative #2: -14.3984

Top negative #3: -14.3906

Top 10 prompts most aligned with POSITIVE PC3 direction:

1. 'What are the best dog breeds for hot climates?'

2. 'What are the best dog breeds for cold climates?'

3. 'What are the most common signs of illness in dogs?'

4. 'What are the signs of separation anxiety in dogs?'

5. 'What are the signs of heatstroke in dogs?'

6. 'What are the most common health issues in dogs?'

7. 'What are the best dog breeds for apartment living?'

8. 'What are the best dog breeds for companionship?'

9. 'What are the most common causes of bad breath in dogs?'

10. 'What are the best dog breeds for active people?'

Top 10 prompts most aligned with NEGATIVE PC3 direction:

1. 'What would a dog say about the invention of the leash?'

2. 'How would a dog describe the feeling of being adopted?'

3. 'How do I teach my dog to push a ball with its nose?'

4. 'What would a dog say about going to the vet?'

5. 'How do I teach my dog to balance objects on its nose?'

6. 'How would a dog explain the concept of loyalty?'

7. 'What would a dog say about its first walk?'

8. 'How do I teach my dog to find hidden treats?'

9. 'How do I teach my dog to carry objects?'

10. 'What would a dog say about being left alone?'

================================================================================

================================================================================

--- Analyzing original dataset prompts along the PC4 'dog' direction (Layer 31) ---

Projection values:

Top positive #1: 9.0078

Top positive #2: 8.2266

Top positive #3: 8.2031

Top negative #1: -11.0859

Top negative #2: -10.7891

Top negative #3: -10.3984

Top 10 prompts most aligned with POSITIVE PC4 direction:

1. 'Have you ever seen a happier puppy?'

2. 'Could you grab the poop bags?'

3. 'Is he allowed to have a bone?'

4. 'How do I help my dog with fear of water?'

5. 'Why must this dog always choose my clean laundry as the perfect spot for a nap?'

6. 'How do I help my dog with fear of other animals?'

7. 'Did you remember to give him his heartworm medication?'

8. 'Are there any off-leash parks nearby?'

9. 'I think I'm more excited for his birthday than he is.'

10. 'How do I help my dog with fear of loud noises?'

Top 10 prompts most aligned with NEGATIVE PC4 direction:

1. 'What would a dog's diary entry look like after a day at the park?'

2. 'Write a letter from a dog to a cat.'

3. 'Describe a dog's reaction to seeing snow for the first time.'

4. 'Write a letter from a dog to its owner.'

5. 'Describe a futuristic city designed for dogs.'

6. 'What would a dog say about its first walk?'

7. 'How would a dog describe the feeling of being adopted?'

8. 'How would a dog explain friendship?'

9. 'What would a dog say about moving to a new home?'

10. 'How would a dog explain the concept of loyalty?'

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31, PROJECTION ONTO PC0 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -20.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -10.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -5.0x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -2.5x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale -1.5x: neyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyneyney

Perturbation scale +0.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +1.5x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +2.5x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +5.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +10.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +20.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

Perturbation scale +100.0x: assistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistantassistant

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31, PROJECTION ONTO PC1 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -20.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -10.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -5.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -2.5x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale -1.5x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale +0.0x: translatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale +1.5x: olenolenolenolentranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslatedtranslated

Perturbation scale +2.5x: olenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolen

Perturbation scale +5.0x: olenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolen

Perturbation scale +10.0x: olenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolen

Perturbation scale +20.0x: olenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolen

Perturbation scale +100.0x: olenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolenolen

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31, PROJECTION ONTO PC2 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -20.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -10.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -5.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -2.5x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale -1.5x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +0.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +1.5x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +2.5x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +5.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +10.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +20.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

Perturbation scale +100.0x: elseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelseaelsea

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31, PROJECTION ONTO PC3 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -20.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -10.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -5.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -2.5x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale -1.5x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale +0.0x: >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Perturbation scale +1.5x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +2.5x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +5.0x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +10.0x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +20.0x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

Perturbation scale +100.0x: �数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数�数

================================================================================

================================================================================

--- PROJECTION-BASED PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31, PROJECTION ONTO PC4 ONLY ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Perturbation scale -100.0x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -20.0x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -10.0x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -5.0x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -2.5x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale -1.5x: ":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

":[{

Perturbation scale +0.0x: ":[{

( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +1.5x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +2.5x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +5.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +10.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +20.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

Perturbation scale +100.0x: ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (

================================================================================

================================================================================

--- ORTHOGONAL PERTURBATION ON CONCEPT: 'dog' ---

--- LAYER: 31 ---

================================================================================

Perturbing along first orthogonal direction (PC143)...

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

Original model completion: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

--- Perturbing final token activation along orthogonal direction ---

Perturbation scale -100.0x: 건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건

Perturbation scale -20.0x: 건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건건

Perturbation scale -10.0x: Here are the descriptions of two dog breeds, the Golden Retriever and the French Bulldog:

\*\*Golden Retriever\*\*

The Golden Retriever is a popular dog breed known for its friendly and loyal personality. It is a medium to large-sized dog with a sturdy build and a thick double coat of fur. Their fur is a beautiful golden

Perturbation scale -5.0x: Here are the descriptions of two different dog breeds:

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its incredible speed. They are a sight hound, bred for their ability to chase small game such as rabbits and hares. Greyhounds are a medium to large-sized dog, typically weighing between

Perturbation scale -2.5x: Here are descriptions of two different dog breeds:

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to chase small game such as rabbits and hares. Greyhounds are typically short-coated, with a smooth, glossy coat that

Perturbation scale -1.5x: Let's take a look at two popular dog breeds: the Golden Retriever and the French Bulldog.

\*\*1. Golden Retriever\*\*

The Golden Retriever is a medium to large-sized dog breed known for its friendly, loyal, and gentle nature. They are originally from Scotland and were bred to retrieve game for hunters. Here are

Perturbation scale +0.0x: I'd be happy to describe two different dog breeds.

\*\*1. The Greyhound\*\*

The Greyhound is a sleek and athletic dog breed known for its speed and agility. They are a sight hound, bred for centuries to hunt small game such as rabbits and hares. Greyhounds are typically 23-27 inches (58-69

Perturbation scale +1.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and elegant dog breed that originated in Germany. They are known for their distinctive curly or wavy coat, which requires regular grooming to prevent matting. Poodles come in three sizes: Toy, Miniature, and Standard.

Perturbation scale +2.5x: I'd be happy to describe two different dog breeds.

\*\*1. The Poodle\*\*

The Poodle is a highly intelligent and elegant dog breed that originated in Germany. There are three main sizes of Poodles: Toy, Miniature, and Standard.

\* \*\*Physical Characteristics:\*\* Poodles have a curly, non-shedding coat that requires regular

Perturbation scale +5.0x: I'd be happy to describe two different dog breeds.

\*\*1. Poodle\*\*

The Poodle is a highly intelligent and elegant breed that originated in Germany. There are three main sizes: Toy, Miniature, and Standard.

- \*\*Physical Characteristics:\*\* Poodles have curly, non-shedding coats that require regular grooming. They have almond-shaped

Perturbation scale +10.0x: I can describe two different dog breeds.

1. \*\*Poodle:\*\*

The Poodle is one of the most recognizable and intelligent breeds in the world. Poodles come in three main sizes: Toy, Miniature, and Standard.

- \*\*Appearance:\*\* Poodles have curly, hypoallergenic coats that range in texture and texture combinations; including

Perturbation scale +20.0x: 152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152152

Perturbation scale +100.0x: ruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruhruh

================================================================================

================================================================================

--- ABLATION EXPERIMENT ON CONCEPT: 'dog' ---

--- LAYER: 31 ---

================================================================================

System Prompt: 'You are a helpful assistant.'

User Prompt: 'Please describe two different dog breeds.'

All PCs (centroid only):

All PCs except largest (PC0 only):

All PCs except largest two (PC0+PC1):

================================================================================

################################################################################

### PLOTTING OVERALL RESULTS ###

################################################################################

Saved average eigenvalue plot to lastToken\_dog\_avg\_eigenvalue.png

Saved eigenvector similarity matrix to lastToken\_dog\_pc0\_similarity.png