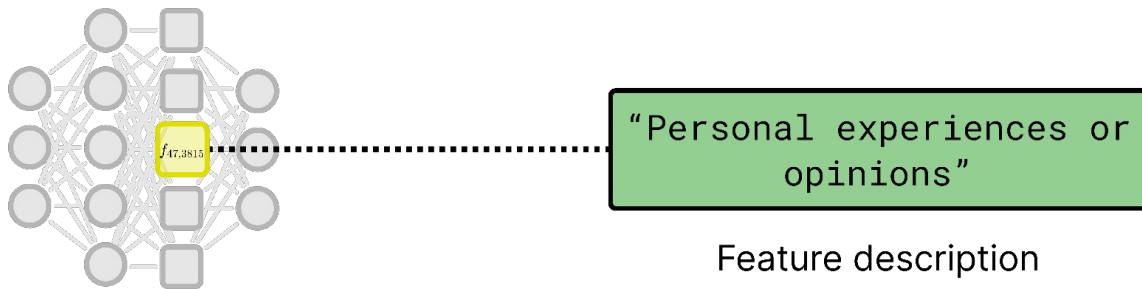


# Capturing Polysemanticity with PRISM: A Multi-Concept Feature Description Framework

Laura Kopf, Nils Feldhus, Kirill Bykov, Philine Lou Bommer, Anna Hedström,  
Marina M.-C. Höhne, Oliver Eberle



# Which Concepts does a Feature encode?



**Layer 47, Feature 3815**  
{GPT-2 XL, MLP}

**Feature:** Here, a neuron in an LLM.

# Previous Automated Interpretability Methods

Method	Target Model	Explainer Model	Feature Type
SASC [1]	BERT		neuron
GPT explain [2]	GPT-2 XL	GPT-4	neuron
EleutherAI SAE I [3]	Pythia-70M and Pythia 410-M	GPT-4	SAE feature
Anthropic SAE [4]	one-layer transformer	Claude	SAE feature
GPT-2 SAE [5]	GPT-2 small	Neuron to Graph (N2G)	SAE feature
GPT-4 SAE [5]	GPT-4	N2G	SAE feature
EleutherAI SAE II [6]	Llama 3.1 7b & Gemma 2 9b		SAE feature
Transluce explain [7]	Llama-3.1-8B-Instruct	distilled GPT-4o-mini	neuron
Llama Scope [8]	Llama-3.1-8B-Base		SAE feature
Gemma Scope [9]	Gemma 2	—	SAE feature
Goodfire explain [10]	Llama 3.3 70B	Claude	SAE feature

[1] Chandan Singh et al. Explaining black box text modules in natural language with language models. 2023.

[2] Steven Bills et al. Language models can explain neurons in language models. 2023.

[3] Hoagy Cunningham et al. Sparse Autoencoders Find Highly Interpretable Features in Language Models. 2023.

[4] Trenton Bricken et al. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. 2023.

[5] Leo Gao et al. Scaling and evaluating sparse autoencoders. 2024.

[6] Gonçalo Paulo et al. Automatically Interpreting Millions of Features in Large Language Models. 2024

[7] Dami Choi et al. Scaling Automatic Neuron Description | Transluce AI. 2024.

[8] Zhengfu He et al. Llama Scope: Extracting Millions of Features from Llama-3.1-8B with Sparse Autoencoders. 2024.

[9] Tom Lieberum et al. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. ACL 2024.

[10] Thomas McGrath et al. Mapping the Latent Space of Llama 3.3 70B. Goodfire Papers. 2024.

# Seminal work by OpenAI (Bills et al., 2023)

Step 1 **Explain** the neuron's activations using GPT-4

Step 2 **Simulate** activations using GPT-4, conditioning on the explanation

Step 3 **Score** the explanation by comparing the simulated and real activations

## Real activations:

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for Marvel's Daredevil. It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante saved her life.

offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from Skyrim, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

## Simulated activations:

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for Marvel's Daredevil. It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante saved her life.

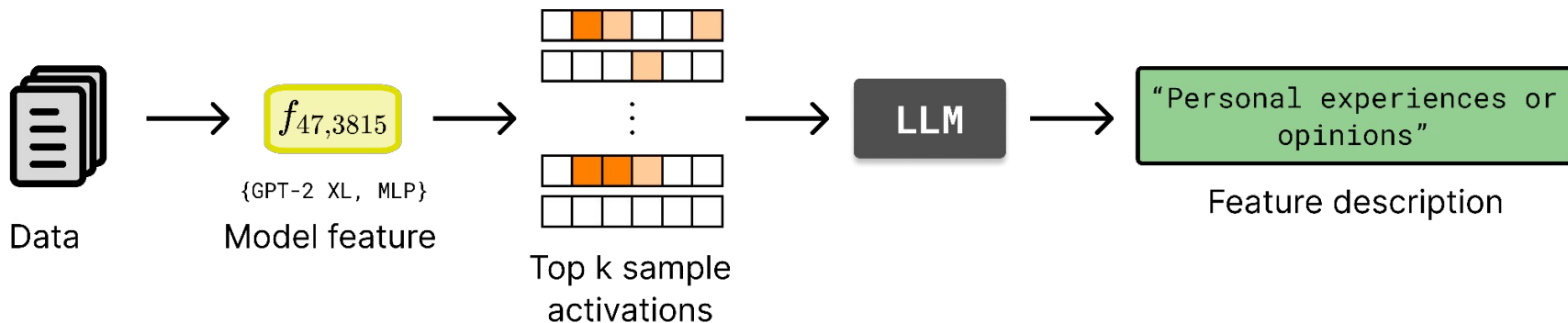
offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from Skyrim, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

Comparing the simulated and real activations to see how closely they match, we derive a score:

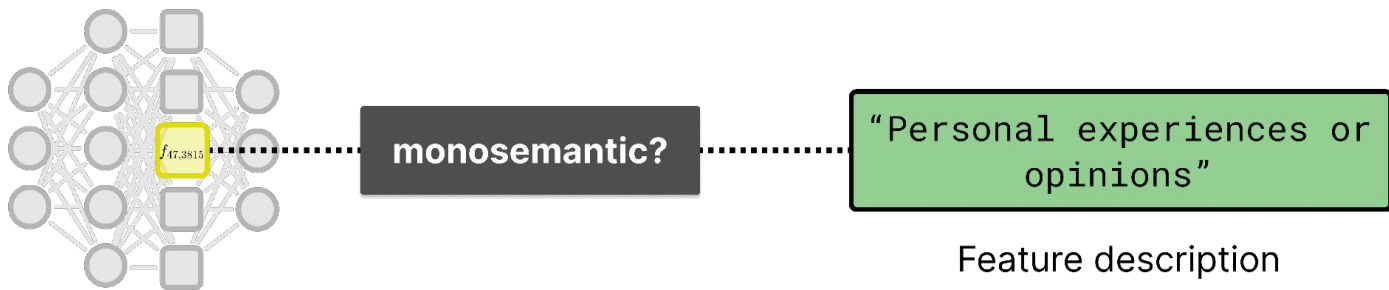
0.337

# Previous Automated Interpretability Methods



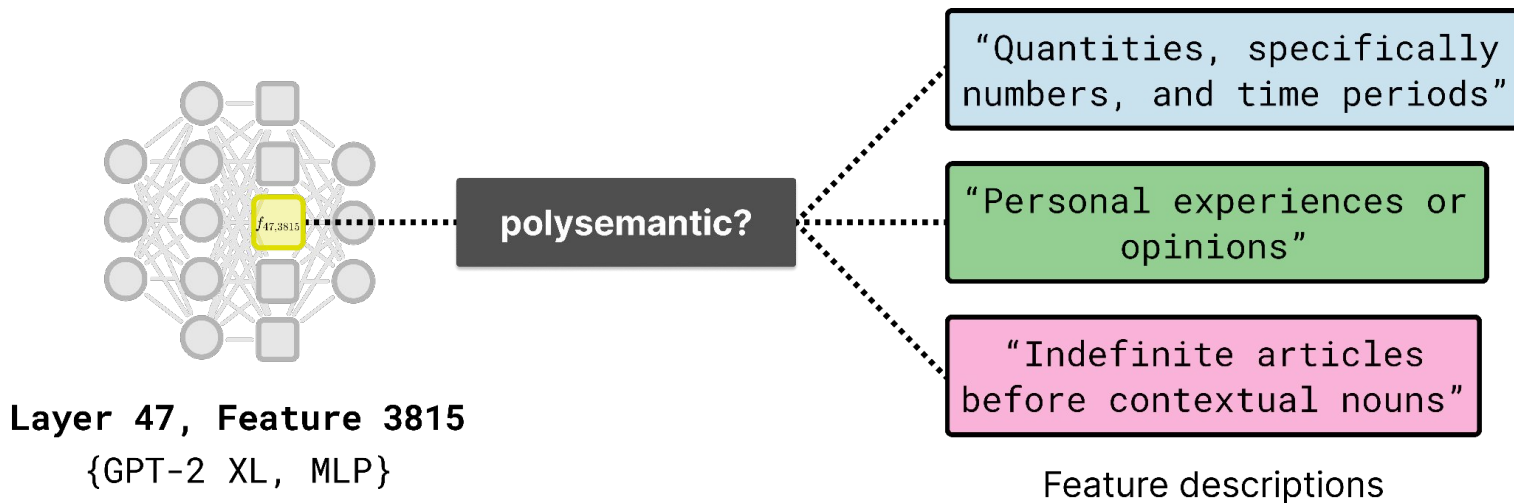
**Goal:** Identify a concept encoded in a feature.

# Which Concepts does a Feature encode?



**Layer 47, Feature 3815**  
{GPT-2 XL, MLP}

# Which Concepts does a Feature encode?



## Problem

- Individual features often **encode multiple distinct concepts** (polysemanticity).
- **Standard automated interpretability methods** assume each feature corresponds to a single concept.  
↳ This leads to an **illusion of monosemanticity**.

## Solution

- **Identify** polysemantic features.
- For each such feature, **detect the distinct concepts** it responds to.
- Provide a **more accurate description** of what each feature encodes.

## Problem

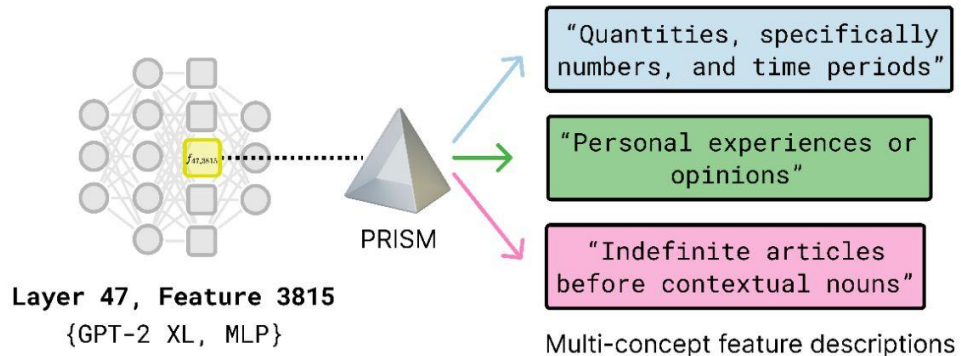
- Individual features often **encode multiple distinct concepts** (polysemanticity).
- **Standard automated interpretability methods** assume each feature corresponds to a single concept.  
↳ This leads to an **illusion of monosemanticity**.

## Solution

- **Identify** polysemantic features.
- For each such feature, **detect the distinct concepts** it responds to.
- Provide a **more accurate description** of what each feature encodes.

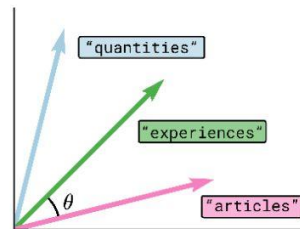
# PRISM Framework

## Extracting Feature Descriptions



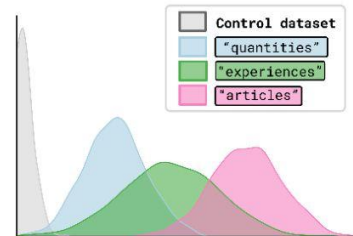
## Evaluation

### Polysemanticity Scoring



Lower Cosine Similarity  
→ high polysemanticity

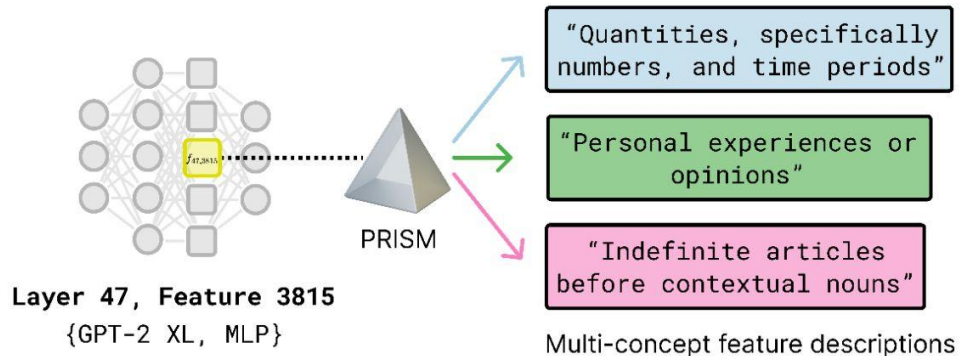
### Description Scoring



Higher activation  
→ more accurate description

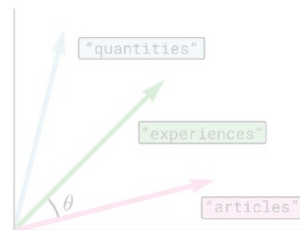
# PRISM Framework

## Extracting Feature Descriptions



## Evaluation

### Polysemanticity Scoring



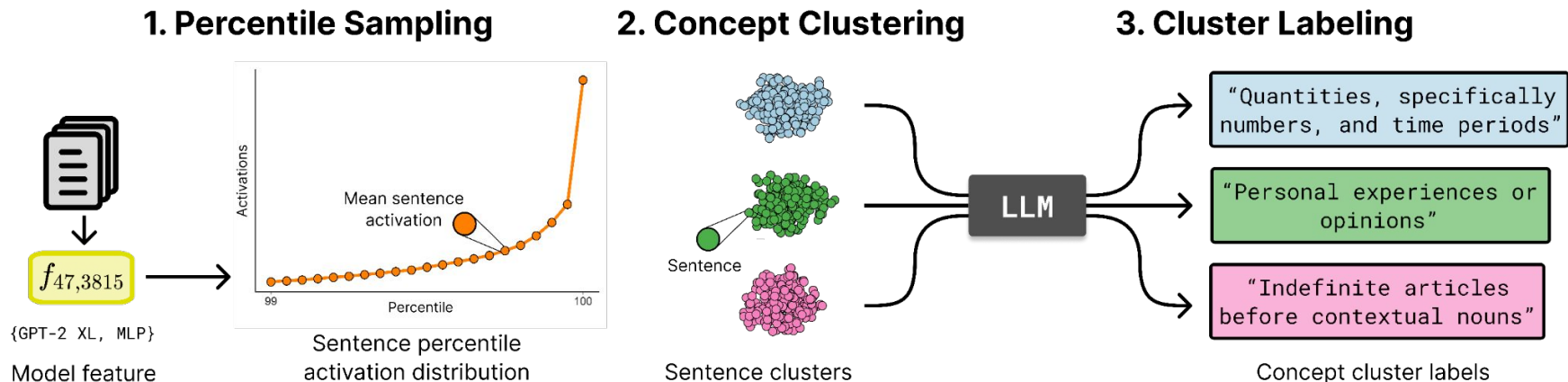
Lower Cosine Similarity  
→ high polysemanticity

### Description Scoring



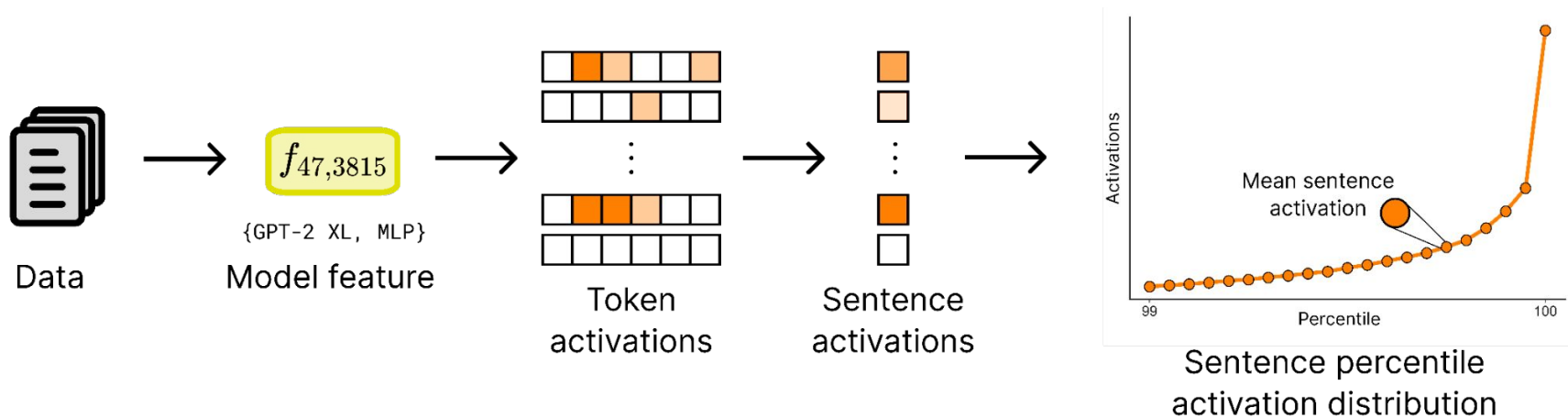
Higher activation  
→ more accurate description

# Extracting Feature Descriptions



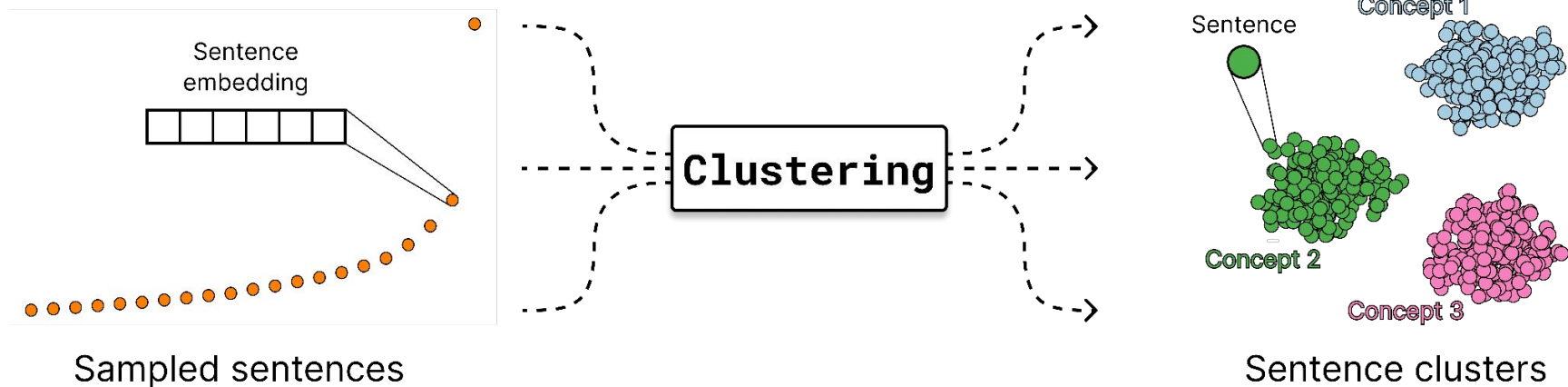
**Goal:** Identify concepts encoded in a feature.

# 1. Percentile Sampling



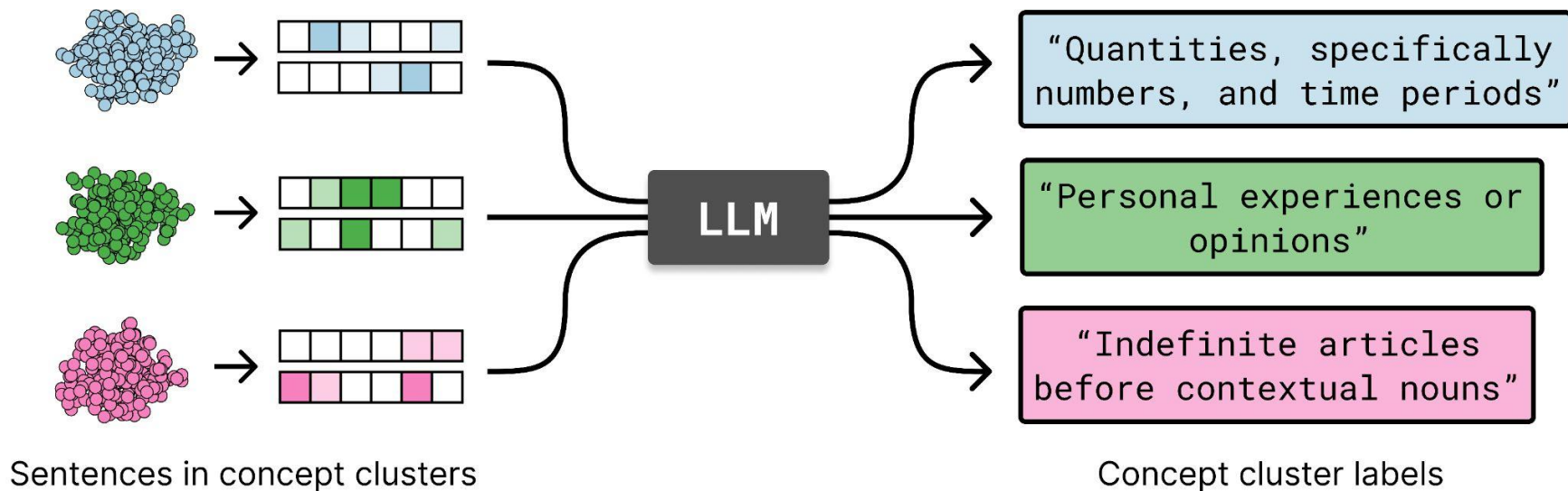
**Broader Sampling:** Sample from different distributions, not only top activating.

## 2. Concept Clustering



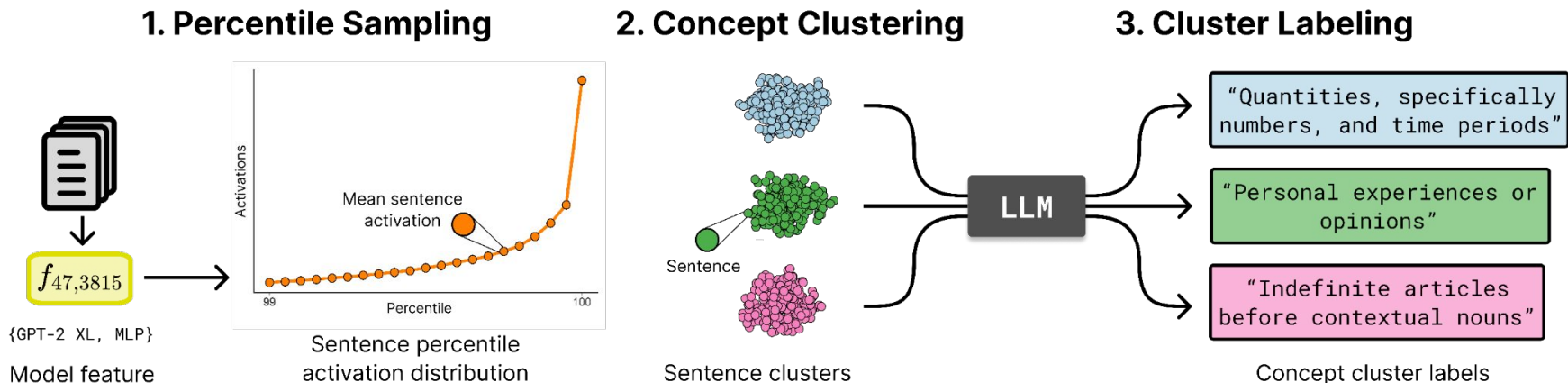
**Concept Discovery:** Cluster high-activation sentences to identify recurring patterns.

### 3. Cluster Labeling



**Descriptions:** Top examples from each cluster guide an LLM in creating descriptive cluster labels.

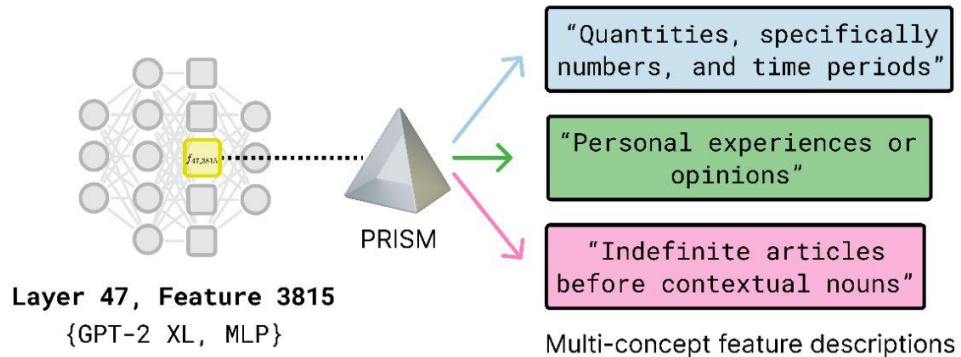
# Extracting Feature Descriptions



**Goal:** Identify concepts encoded in a feature.

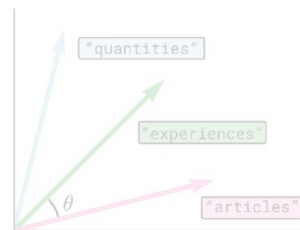
# PRISM Framework

## Extracting Feature Descriptions



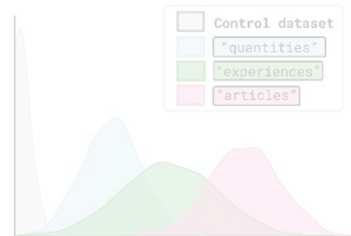
## Evaluation

### Polysemanticity Scoring



Lower Cosine Similarity  
→ high polysemanticity

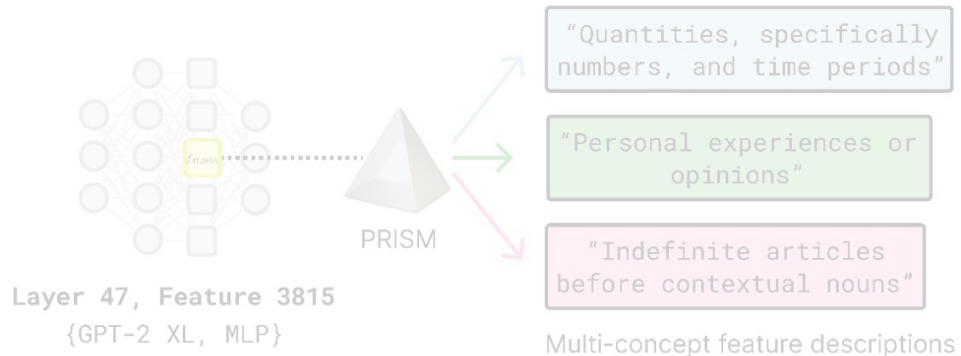
### Description Scoring



Higher activation  
→ more accurate description

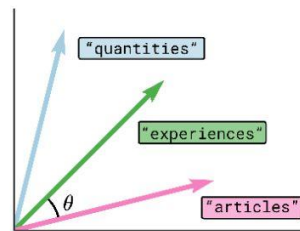
# PRISM Framework

## Extracting Feature Descriptions



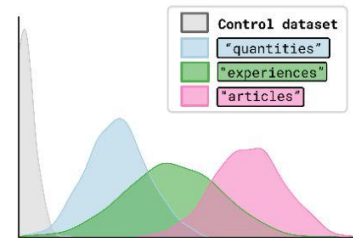
## Evaluation

### Polysemanticity Scoring



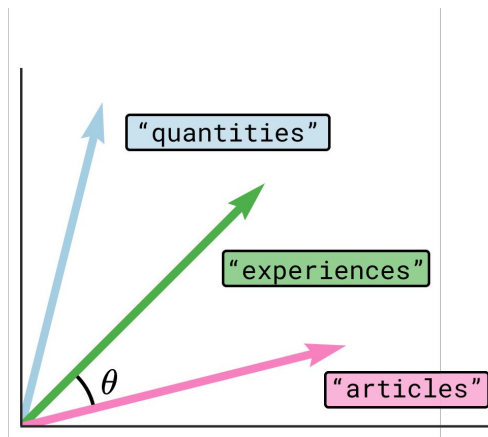
Lower Cosine Similarity  
→ high polysemanticity

### Description Scoring



Higher activation  
→ more accurate description

# Polysemanticity Scoring

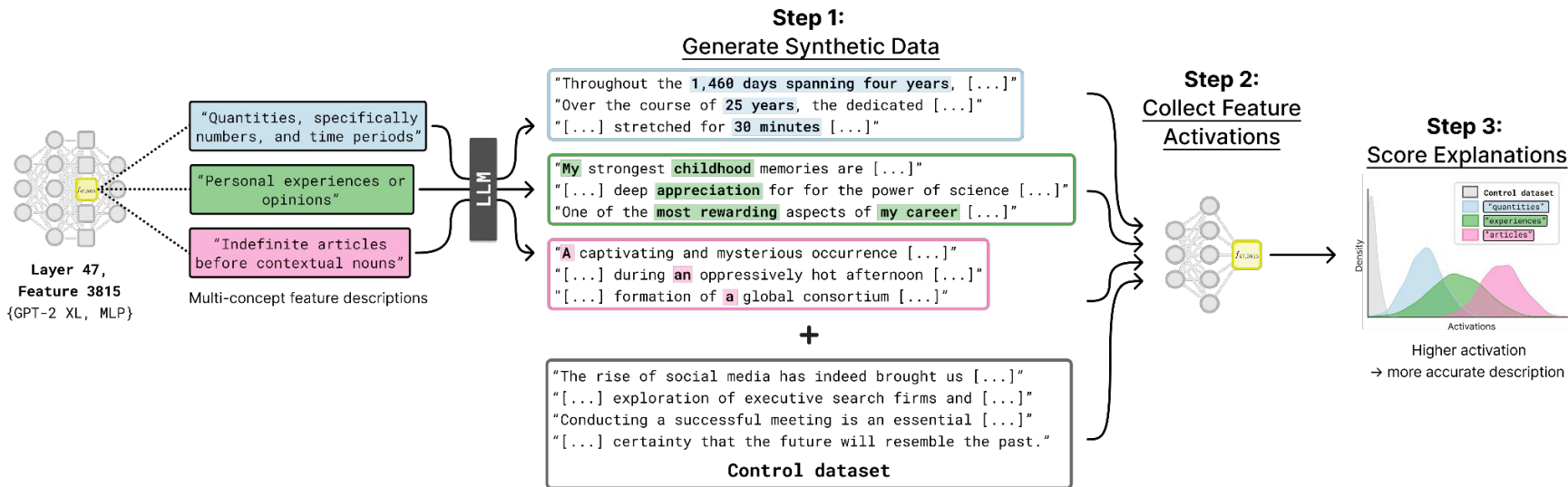


Lower Cosine Similarity  
→ high polysemanticity

1. **Encode descriptions** using a sentence embedding model.
2. Compute pairwise **cosine similarities**.

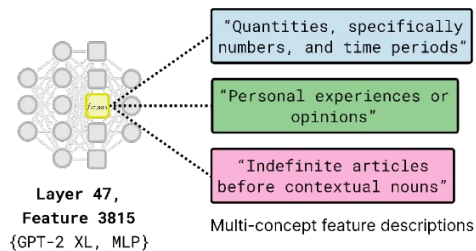
**Evaluation:** Measure similarity among the generated descriptions per feature.

# Description Scoring



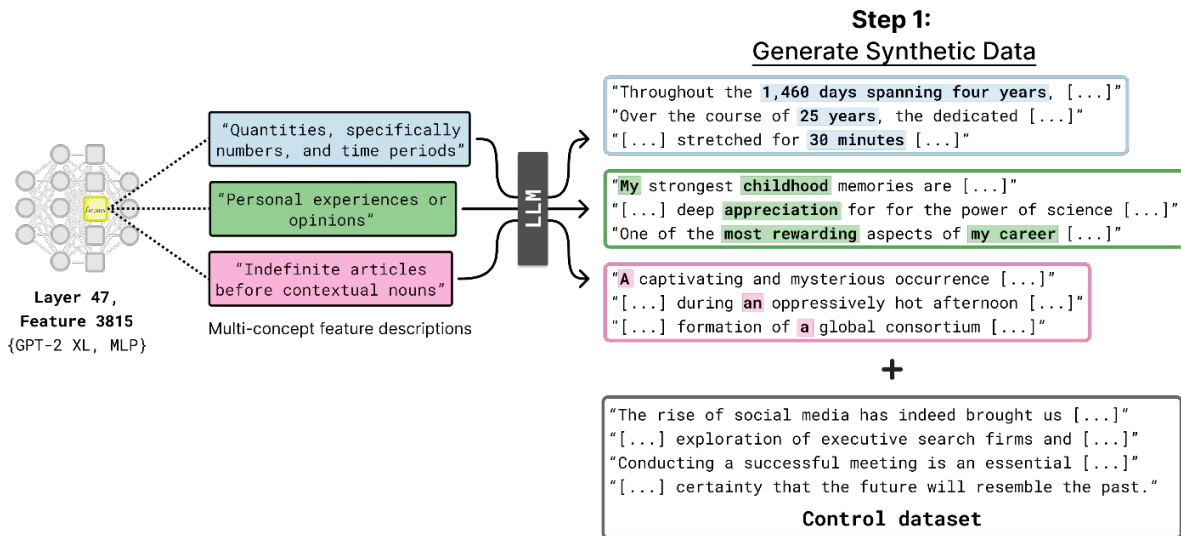
**Evaluation:** Assess how well each description aligns with a feature's activation distribution.

# Description Scoring



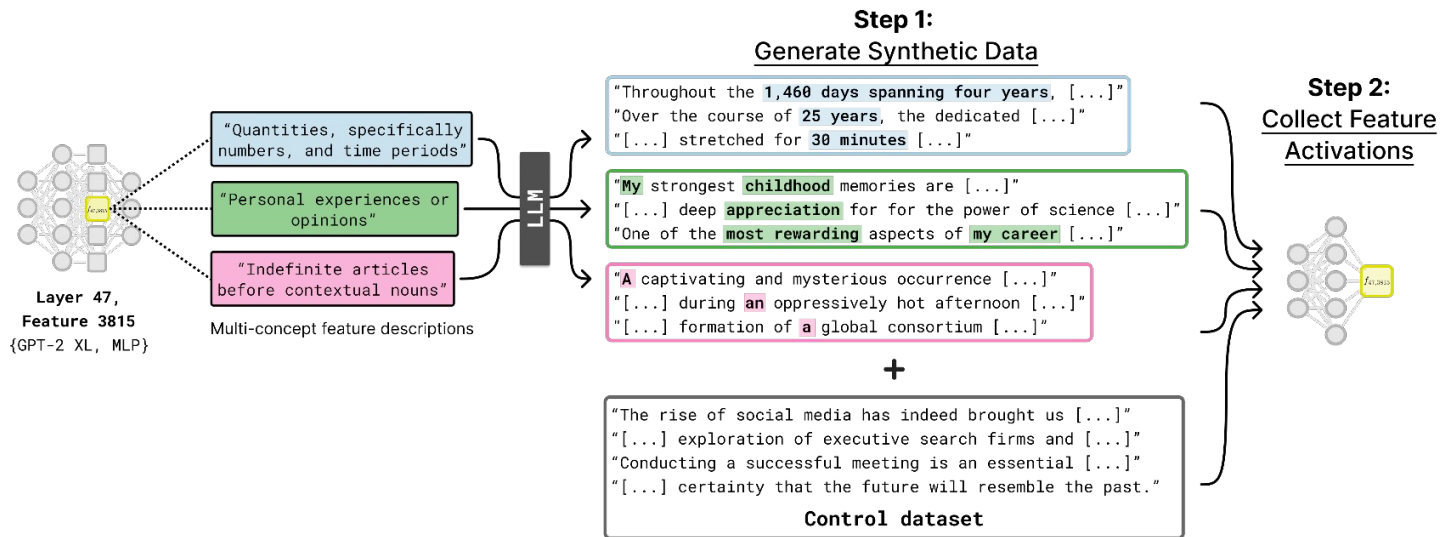
**Evaluation:** Assess how well each description aligns with a feature's activation distribution.

# Description Scoring



**Evaluation:** Assess how well each description aligns with a feature's activation distribution.

# Description Scoring



**Evaluation:** Assess how well each description aligns with a feature's activation distribution.

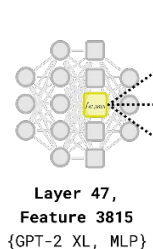
# Description Scoring

## Scoring Functions

$$\Psi_{\text{AUROC}}(\mathbb{A}_0, \mathbb{A}_1) = \frac{\sum_{a \in \mathbb{A}_0} \sum_{b \in \mathbb{A}_1} \mathbf{1}[a < b]}{|\mathbb{A}_0| \cdot |\mathbb{A}_1|}$$

$$\Psi_{\text{MAD}}(\mathbb{A}_0, \mathbb{A}_1) = \frac{\frac{1}{m} \sum_{b \in \mathbb{A}_1} b - \frac{1}{n} \sum_{a \in \mathbb{A}_0} a}{\sqrt{\frac{1}{n-1} \sum_{a \in \mathbb{A}_0} (a - \bar{a})^2}}$$

### Step 1: Generate Synthetic Data



"Quantities, specifically numbers, and time periods"  
"Personal experiences or opinions"  
"Indefinite articles before contextual nouns"  
Multi-concept feature descriptions

LLM

"Throughout the 1,460 days spanning four years, [...]"  
"Over the course of 25 years, the dedicated [...]"  
"[...] stretched for 30 minutes [...]"

"My strongest childhood memories are [...]"  
"[...] deep appreciation for for the power of science [...]"  
"One of the most rewarding aspects of my career [...]"

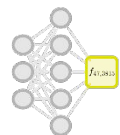
"A captivating and mysterious occurrence [...]"  
"[...] during an oppressively hot afternoon [...]"  
"[...] formation of a global consortium [...]"

+

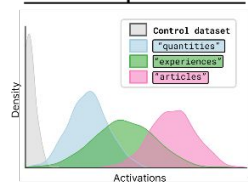
"The rise of social media has indeed brought us [...]"  
"[...] exploration of executive search firms and [...]"  
"Conducting a successful meeting is an essential [...]"  
"[...] certainty that the future will resemble the past."

Control dataset

### Step 2: Collect Feature Activations



### Step 3: Score Explanations



**Evaluation:** Assess how well each description aligns with a feature's activation distribution.

# Benchmark Results

Method	GPT-2 XL (MLP neuron)		Llama 3.1 8B Instruct (MLP neuron)		GPT-2 Small (resid. SAE feature)		Gemma Scope (resid. SAE feature)	
	AUROC (↑)	MAD (↑)	AUROC (↑)	MAD (↑)	AUROC (↑)	MAD (↑)	AUROC (↑)	MAD (↑)
MaxAct	0.53 (0.49-0.58)	11.86%	0.54 (0.46-0.63)	50.00%	0.53 (0.49-0.58)	11.86%	0.60 (0.50-0.69)	50.00%
GPT-Explain [1]	0.64 (0.56-0.73)	65.00%	—	—	—	—	—	—
Transluce-Explain [2]	—	—	0.59 (0.51-0.67)	63.33%	—	—	—	—
Neuronpedia [3]	—	—	—	—	0.54 (0.50-0.59)	18.97%	<b>0.62 (0.53-0.72)</b>	<b>63.33%</b>
Output-Centric [4]	—	—	0.55 (0.46-0.64)	58.33%	0.57 (0.53-0.62)	22.03%	0.58 (0.49-0.67)	46.67%
PRISM (mean)	0.65 (0.61-0.69)	66.33%	0.52 (0.48-0.55)	51.33%	0.51 (0.50-0.53)	13.22%	0.43 (0.39-0.46)	24.67%
PRISM (max)	<b>0.85 (0.78-0.91)</b>	<b>91.67%</b>	<b>0.71 (0.63-0.78)</b>	<b>81.67%</b>	<b>0.57 (0.53-0.61)</b>	<b>28.81%</b>	0.54 (0.45-0.62)	38.33%

PRISM (max) descriptions are more accurate and outperform the competitive approach of GPT-Explain.

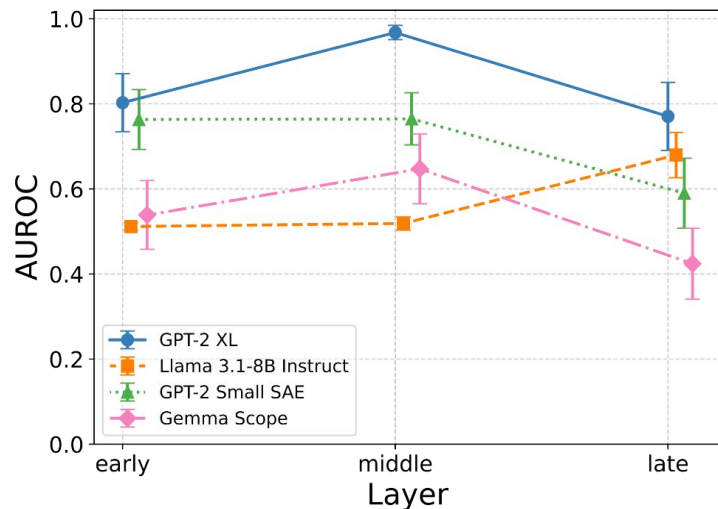
[1] Steven Bills et al. Language models can explain neurons in language models. OpenAI. 2023.

[2] Dami Choi et al. Scaling Automatic Neuron Description. Transluce AI. 2024.

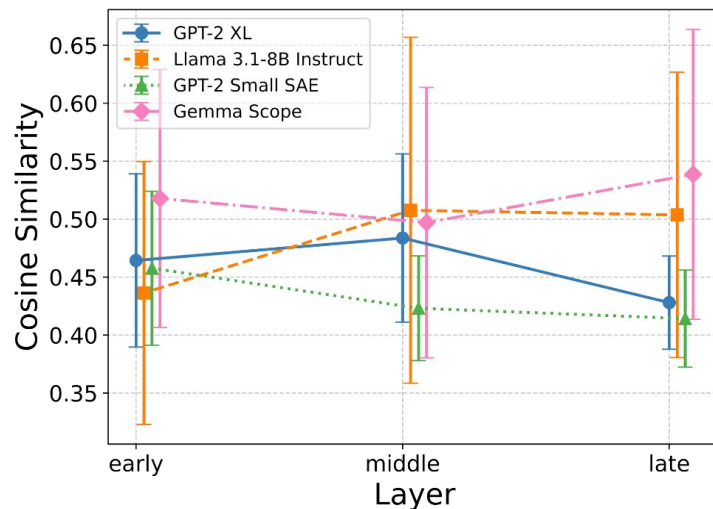
[3] Johnny Lin. Neuronpedia: Interactive Reference and Tooling for Analyzing Neural Networks. 2023.

[4] Yoav Gur-Arieh et al. Enhancing Automated Interpretability with Output-Centric Feature Descriptions. ACL. 2025.

# Evaluation across Models and Layers



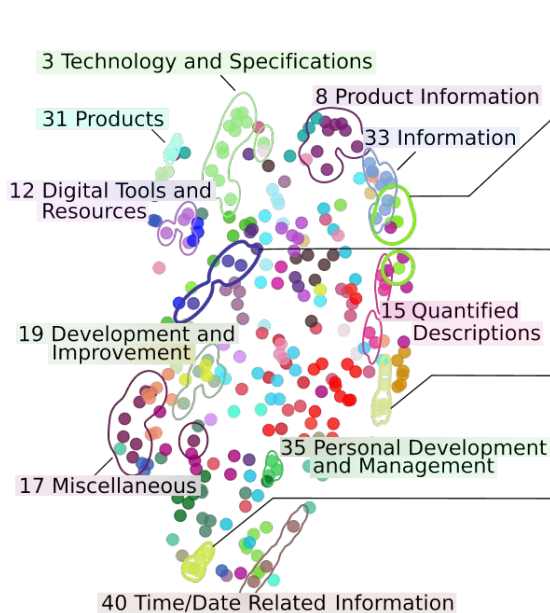
(a) Description scores.



(b) Polysemanticity scores.

- (a) Middle layers generally appear to be easier to interpret.
- (b) Gemma Scope SAE feature descriptions show high monosemanticity across layers.

# Meta-Level Concepts

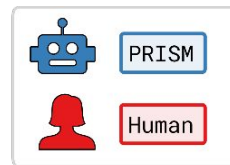


{GPT-2 XL, MLP}

id	Metalabel	Feature Descriptions
18	Structured Data	<ul style="list-style-type: none"> <li>- Numerical or quantitative values, including years, measurements, and counts, in advert-like text excerpts</li> <li>- Titles, names, locations, and dates in bibliographic entries or citations</li> <li>- Academic degrees, professional roles, locations, and years related to education or employment history</li> </ul>
45	Legal and Administrative Affairs	<ul style="list-style-type: none"> <li>- Division of assets/property, medical procedures/treatments, legal disputes/court proceedings, and organizational activities/events</li> <li>- Ownership of property or membership status</li> <li>- Commercial transactions, legal proceedings, and financial obligations</li> </ul>
29	Events and Activities	<ul style="list-style-type: none"> <li>- Achievements, awards, or special events, often including a specific person or group</li> <li>- Food, specific locations, and activities/routines, often involving a change in direction or state</li> <li>- Events, shows, or locations, often with a time or date, and sometimes including named people</li> </ul>
6	Positive Experiences	<ul style="list-style-type: none"> <li>- Expressions of excitement, sharing, or positive feedback</li> <li>- Expressions of gratitude, current time references, or positive descriptions</li> <li>- Experiences related to travel, leisure activities, meals, and events, particularly those with a temporal element (time, dates, or duration)</li> </ul>

**Metalabels:** Group feature descriptions to identify higher-level topics.

# Polysemanticity and Human Interpretation



## Sentence clusters

## Cluster descriptions

Layer 40, Feature 6067

{GPT-2 XL, MLP}

1. "New Year. Winter holiday. **Christmas** background."  
"Her **charm, confidence and** regal style **captured**"  
"can swing **making** it appear **like a Y shaped necklace.**"
2. "accustomed to seeing the **president deliver** the SOTU"  
"that was why President **Nixon** had to resign"  
"she was the first **presidential** spouse **to testify**"
3. "We will **revisit** these concepts in [...]."  
"a **free online verbalizing** English dictionary."  
"comma **separates the dialogue from the standard**"
4. "Python **subroutines to encode or decode** [...]"  
"With **lambdas-as-method-bodies** that [...]"  
"[...] **including addressing** language internals,"
5. "producer for the **evening of** musical entertainment."  
"**Additionally** another **special thanks** to sponsors"  
"to **create** the World **Premiere** of an **exciting new** ballet"

"Holiday or special occasion accessories/decorations"

"Product descriptions and features in the context of celebrations"

"Discussions of formal speeches by US presidents"

"Official titles of personalities or institutions on formal events"

"Recollection, usage, or discussion of educational resources, creative works, personal experiences, media, or specific items"

"Learning, using, or talking about language, literature, and creative work"

"Descriptions, explanations, or instructions related to methods or processes"

"Technical instructions, especially in the context of computing, language, and education"

"Events, often with a time and/or a date"

"Times of day, locations, events and performances"

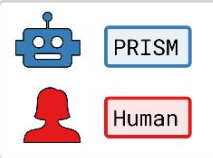
0.38

0.40

Lower Cosine Similarity  
→ high polysemanticity

PRISM aligns with the human interpretation of polysemanticity.

# Monosemanticity and Human Interpretation



## Sentence clusters

## Cluster descriptions

Layer 10, Feature 603

{Gemma Scope, resid}

1. “[...] make the **week** ahead a breeze [...]”  
“A **month or so** ago I found some purple [...]”  
“in **day time** (along with **night time**) [...]”
2. “packages include **year-round** promotion, [...]”  
“School during the academic **year** in [...]”  
“Half **Day** Courses, Full **Day** Courses [...]”
3. “the **season** opener [...] the **night** before”  
“[...] a **day or so** during their **weekly** diet.”  
“a similar call the group made one **year** ago”
4. “May is a busy **month**, with various key [...]”  
“their commander a **Night time** King as [...]”  
“most common questions we get every **year**.”
5. “In the **afternoon**, head out to enjoy [...]”  
“[...] that sounded like a lovely **day** [...]”  
“[...] a camping permit and stay **overnight**.”

- “Units of time (days, weeks, months, years, hours, minutes, evenings, nights, weekends, centuries, lifetimes)”
- “Time references such as days, weeks, months, years and hours in relation to events, routines or durations”
- “Units of time like days, months, and years”
- “Timeframes, events and offers, administration and regulation”
- “Units of time, often in the context of duration, repetition, or scheduling”
- “Time periods during days months or years, times of day, repetitive cycles, schedules”
- “Units of time”
- “Time, time periods”
- “Times of day, days of the week/year, or longer durations of time”
- “Timeframes, nature, travel and events”

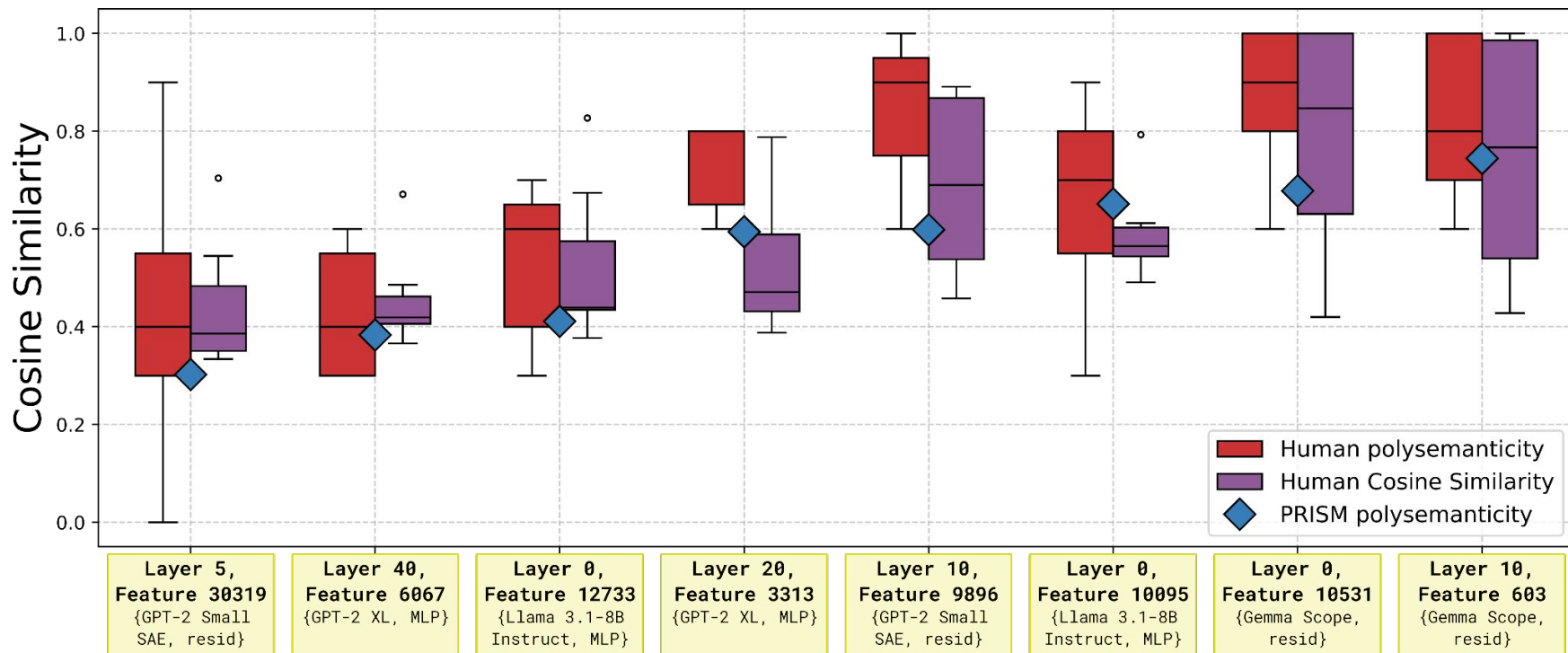
0.74

0.80

Higher Cosine Similarity  
→ high monosemanticity

PRISM aligns with the human interpretation of monosemanticity.

# Polysemanticity and Human Interpretation

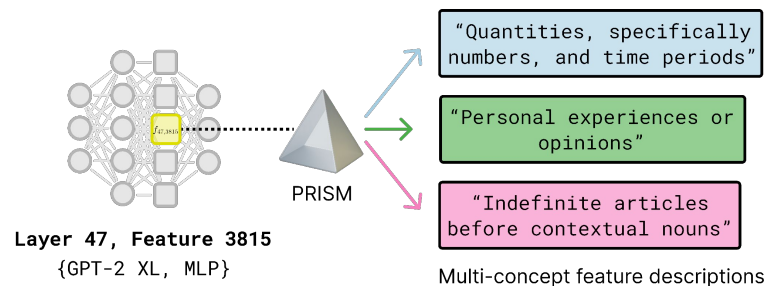


PRISM aligns with the human interpretation of polysemanticity.

# Conclusion

## PRISM

- generates **multi-concept descriptions** of features
- evaluates **polysemanticity** and **description quality**
- enables **multi-level concept analysis**

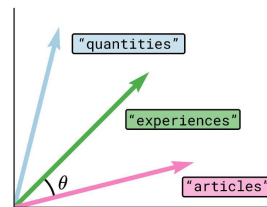


# Conclusion

## PRISM

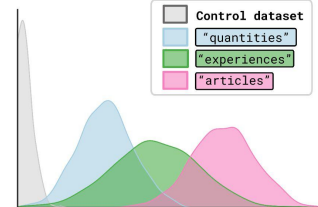
- generates **multi-concept descriptions** of features
- evaluates **polysemanticity** and **description quality**
- enables **multi-level concept analysis**

**Polysemanticity Scoring**



Lower Cosine Similarity  
→ high polysemanticity

**Description Scoring**

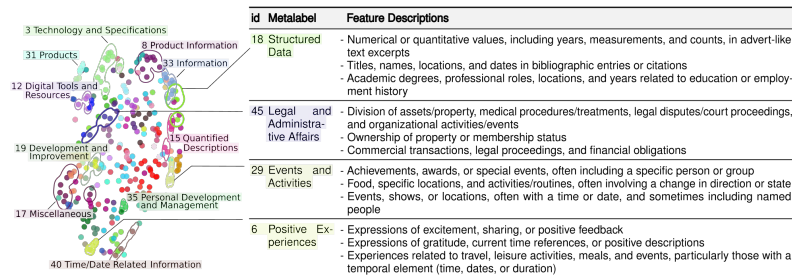


Higher activation  
→ more accurate description

# Conclusion

## PRISM

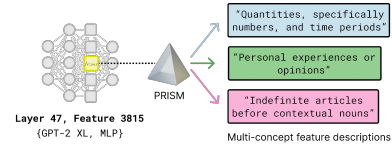
- generates **multi-concept descriptions** of features
- evaluates **polysemanticity** and **description quality**
- enables **multi-level concept analysis**



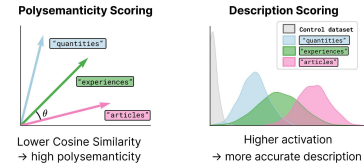
# Conclusion

## PRISM

- generates **multi-concept descriptions** of features



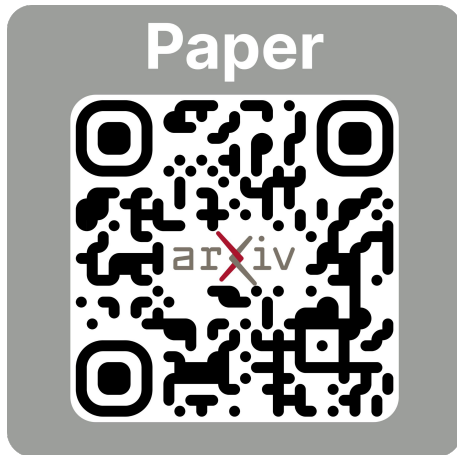
- evaluates **polysemanticity** and **description quality**



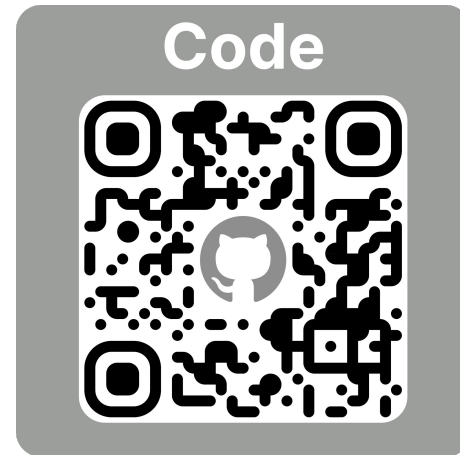
- enables **multi-level concept analysis**

33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100

# Get PRISM !



<https://arxiv.org/abs/2506.15538>



<https://github.com/lkopf/prism>