Semantic Regexes

Automatically Interpreting LLM Features with a Structured Language

Angie Boggust, Donghao Ren, Yannick Assogba, Dominik Moritz,
Arvind Satyanarayan, Fred Hohman

ICLR 2026

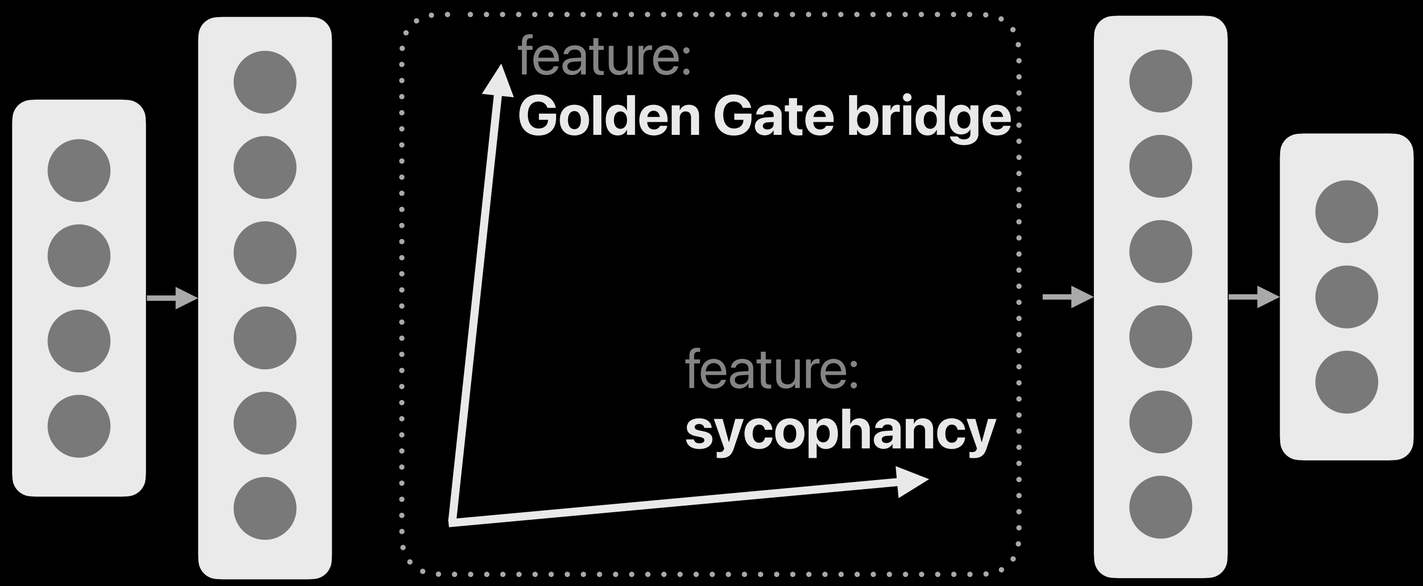
LLMs encode concepts as features in representation space

Understanding LLM features is important for...

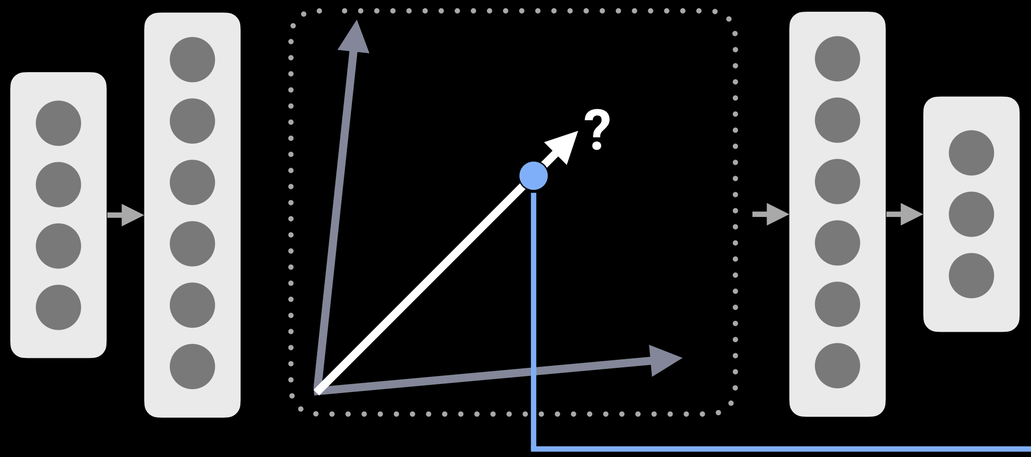
Interpretability
Understand medical suggestions and hallucination.

Safety
Detect deception and sycophancy.

Control
Steer models towards or away from concepts.



Feature descriptions explain models concepts



Natural Language Description

Verbs related to journalism that denote the act of covering or reporting events or issues, often in a critical context.

LLM

Activating Data

responsible	journalists	who	covered	the	story
foreign	journalists	covering	the	famine	at
his	channel	was	covering	the	issue	when

Could structured languages help us understand LLM features?

Natural language descriptions are ...

- **Convoluted** and long, even for conceptually simple features

Natural Language Description

Verbs related to journalism that denote the act of covering or reporting events or issues, often in a critical context.

Activating Data

responsible journalists who covered the story

foreign journalists covering the famine at

his channel was covering the issue when

Could structured languages help us understand LLM features?

Natural language descriptions are ...

- **Convoluted** and long, even for conceptually simple features
- Imprecise and contain **ambiguity**.

Natural Language Description

Verbs related to journalism that denote the act of covering or reporting events or issues, often in a critical context.

Activating Data

responsible journalists who covered the story

foreign journalists covering the famine at

his channel was covering the issue when

Could structured languages help us understand LLM features?

Natural language descriptions are ...

- **Convoluted** and long, even for conceptually simple features
- Imprecise and contain **ambiguity**.

What if we used structured language descriptions that ...

- **Match the patterns** of LLM features
- Combine **syntax** with conceptual **flexibility**

Natural Language Description

Verbs related to journalism that denote the act of covering or reporting events or issues, often in a critical context.

Activating Data

responsible journalists who covered the story

foreign journalists covering the famine at

his channel was covering the issue when

Semantic Regex Description

@{:context journalism:}[:lexeme cover:]

Could structured languages help us understand LLM features?

Natural language descriptions are ...

- **Convoluted** and long, even for conceptually simple features
- Imprecise and contain **ambiguity**.

What if we used structured language descriptions that ...

- **Match the patterns** of LLM features
- Combine **syntax** with conceptual **flexibility**

Natural Language Description

Verbs related to journalism that denote the act of covering or reporting events or issues, often in a critical context.

Activating Data

responsible journalists who covered the story

foreign journalists covering the famine at

his channel was covering the issue when

Semantic Regex Description

@{:context journalism:}[:lexeme cover:]

Semantic regex descriptions combine structure and flexibility

Primitives — atomic units designed to match common feature patterns

`symbol`

Match exact strings

SEMANTIC REGEX:

```
[:symbol color:]
```

DATA EXAMPLES:

a splash of **color**

color your world

`lexeme`

Match syntactically related variants

SEMANTIC REGEX:

```
[:lexeme color:]
```

DATA EXAMPLES:

color in a **coloring** book

her favorite **colors**

`field`

Match semantically related variants

SEMANTIC REGEX:

```
[:field color:]
```

DATA EXAMPLES:

blue jeans

pink skies at night

Modifiers — increase expressive power by refining or expanding scope

`context`

Match when appearing in the context

SEMANTIC REGEX:

```
@{:context politics:}
([:field color:])
```

DATA EXAMPLES:

turned **blue** in the election

Green Party candidate

`combinations`

Sequential or OR combinations

SEMANTIC REGEX:

```
[:field color:] ([:symbol and:]|
[:symbol or:]) [:field color:]
```

DATA EXAMPLES:

green **or** **yellow** bananas

it is **black** **and** **white**

`quantifiers`

Metacharacter to denote zero or one

SEMANTIC REGEX:

```
[:symbol a:] [:field color:]?
[:field flower:]
```

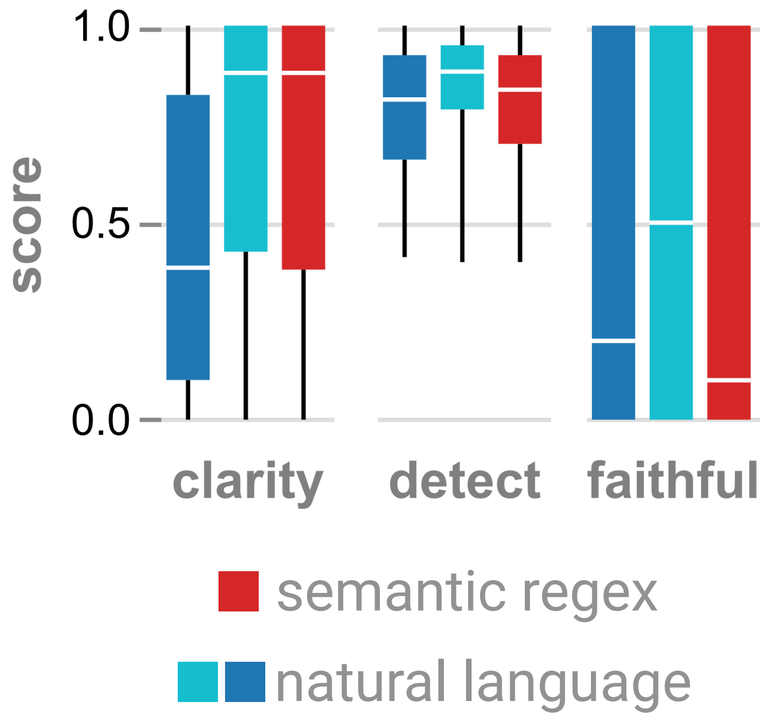
DATA EXAMPLES:

a **red** **rose** is blooming

it is **a** **da** **isy**

Semantic regexes benefit interpretability without losing expressivity

Semantic regexes **perform on par** with natural language.



Semantic regexes are **more concise and consistent** than natural language

Activating Data

They tried to **run** their usual scam

I mean he was **running** wild

Natural Language Description

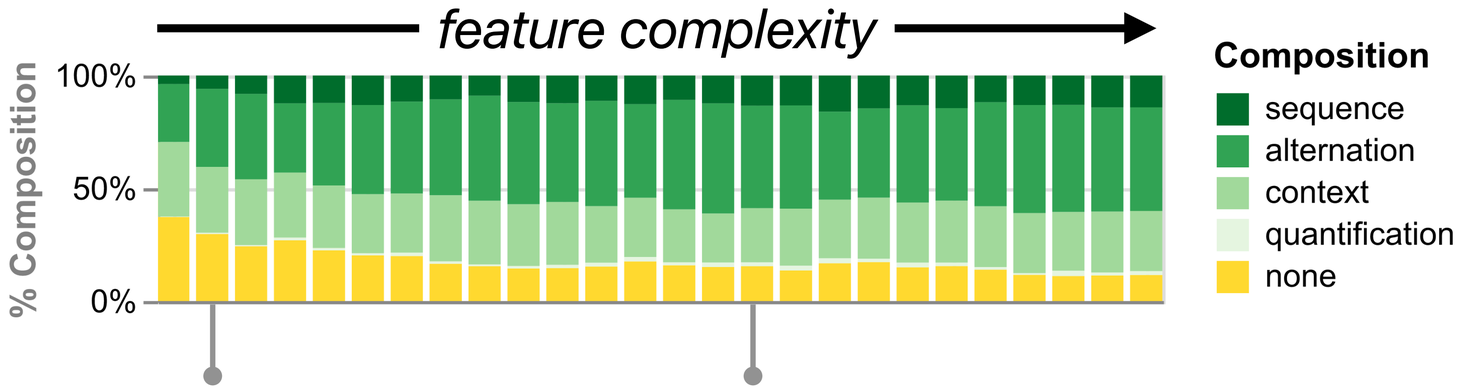
The verb “run” and its variations frequently describe physical movement or participation.

Semantic Regex Description

[:lexeme run:]

Semantic regexes afford model-wide feature interpretability

Semantic regexes encode feature complexity, revealing an increase in complex primitives and compositions across layers.



Simple feature: one low-level primitive

Semantic Regex Description

[:lexeme run:]

Complex feature: 3 composed primitives

Semantic Regex Description

[:symbol born:] [:field location:] [:field date:]

Semantic regexes help people understand LLM feature behavior

Semantic regexes help people build **accurate mental models** of feature behavior.

Natural Language Description

The phrase “expected to” is often used to indicate anticipation.

✅ Activating Example

it is expected to rain

❌ Counterfactual Example

He does not know the meaning of the phrase expected to

Semantic Regex Description

[:lexeme expect:] [:symbol to:]

✅ Activating Example

He expects to finish on time

✅ Counterfactual Example

She’s expecting in November

Semantic regexes set up future work on interpretability

Semantic regexes afford new types of interpretability analysis.

By structuring descriptions, semantic regexes **improve consistency**, **conciseness**, and **encode complexity**.

People easily adopt structured language descriptions

Challenges the assumption that that structured languages for interpretability come at the cost of “specialized training”.

Alternative structured languages could benefit other domains.

We could design analogous structured languages to define **safety guardrails** or **model steering.**

Semantic Regexes

Automatically Interpreting LLM Features with a Structured Language

Angie Boggust, Donghao Ren, Yannick Assogba, Dominik Moritz,
Arvind Satyanarayan, Fred Hohman

ICLR 2026