

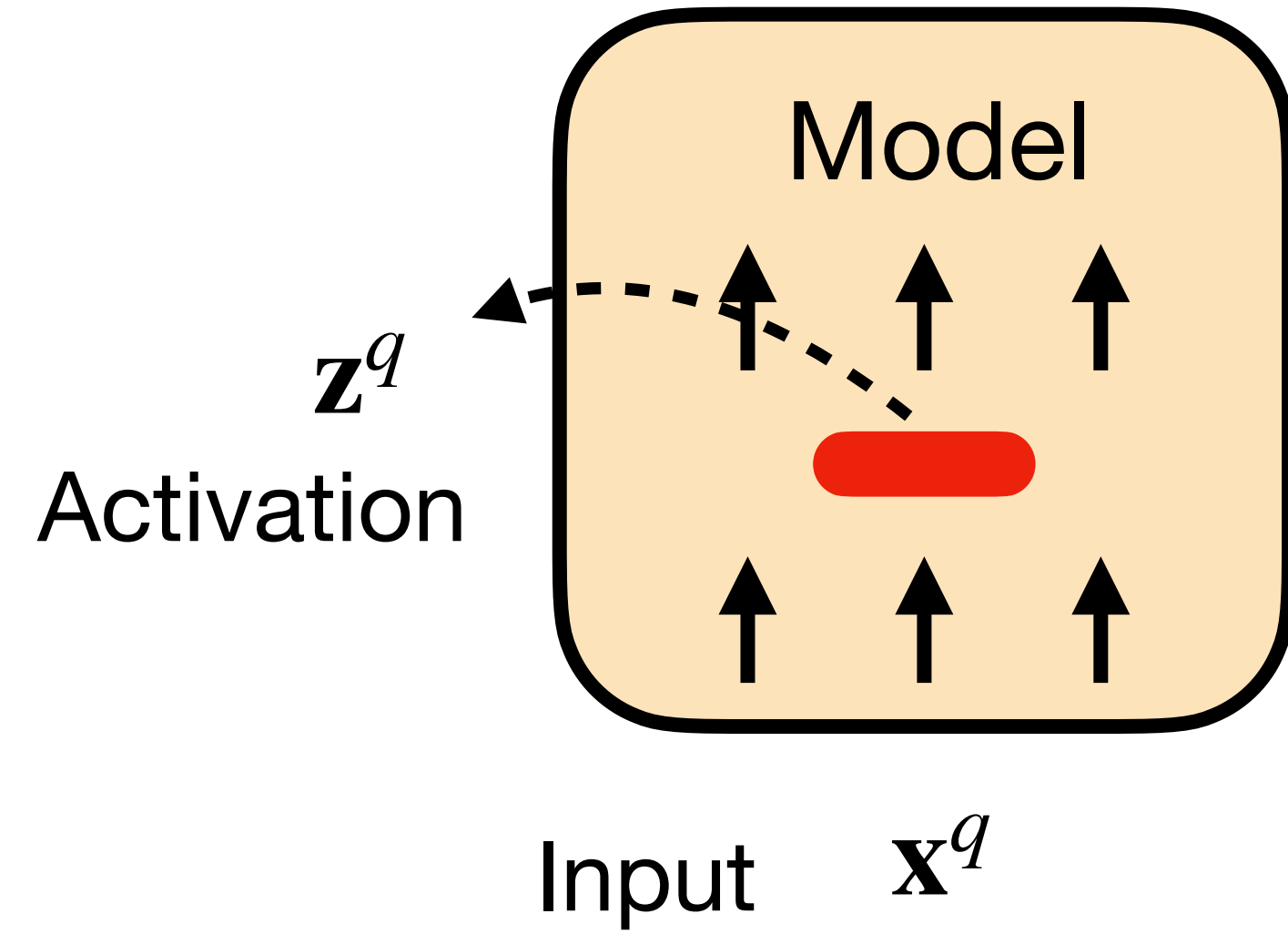
# **InversionView: A General-Purpose Method for Reading Information from Neural Activations**

**Xinting Huang, Madhur Panwar, Navin Goyal, Michael Hahn**

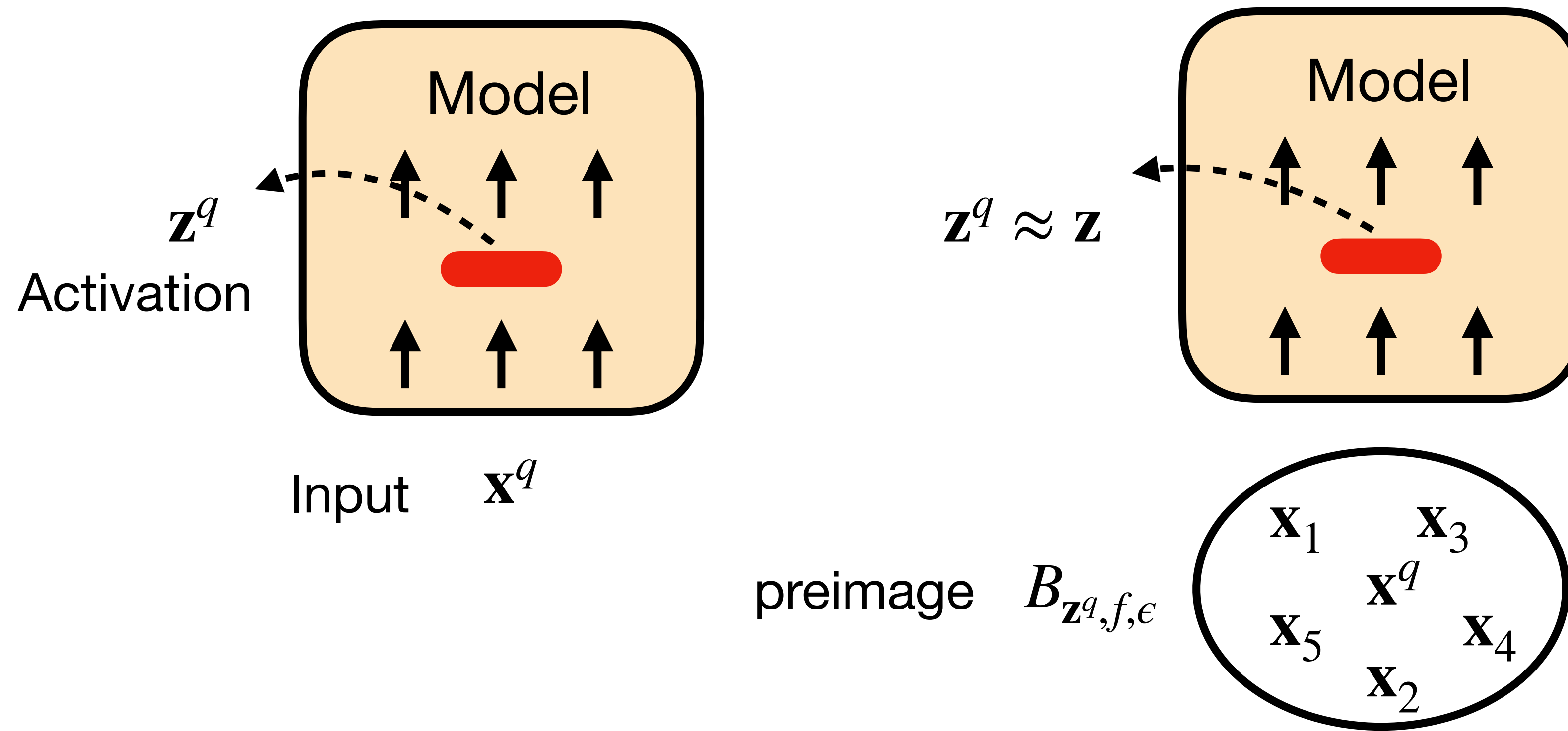


# Introduction & Methodology

**Research Question:** What information is encoded in neural activations?

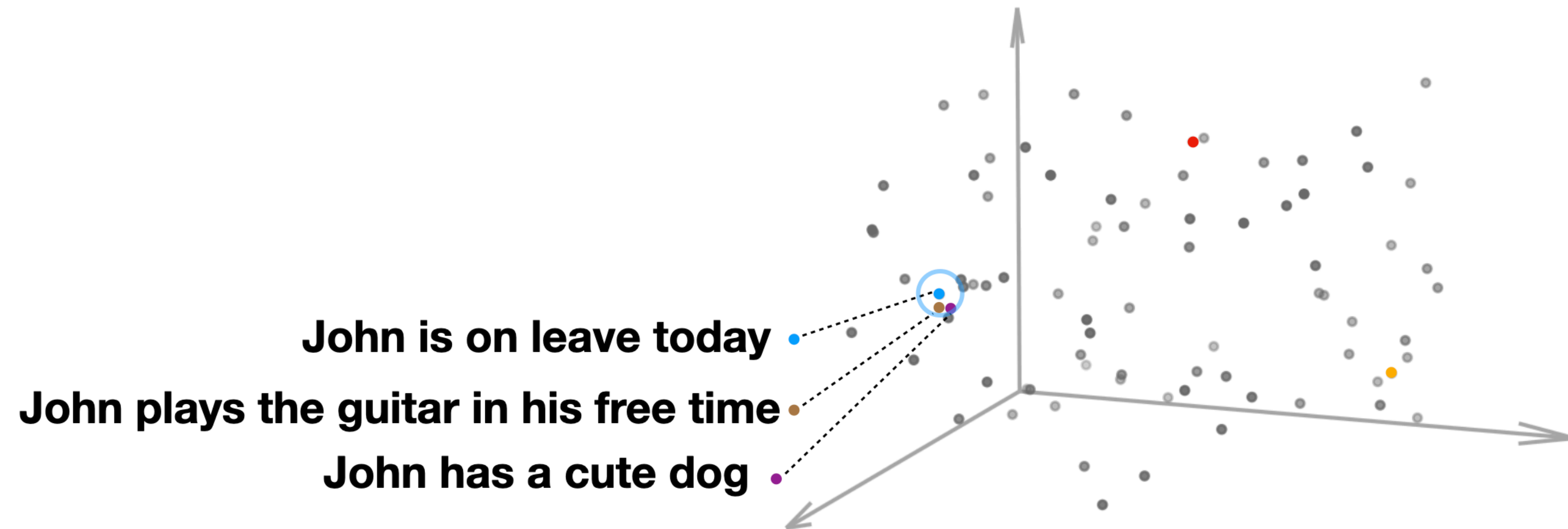


**Research Question:** What information is encoded in neural activations?

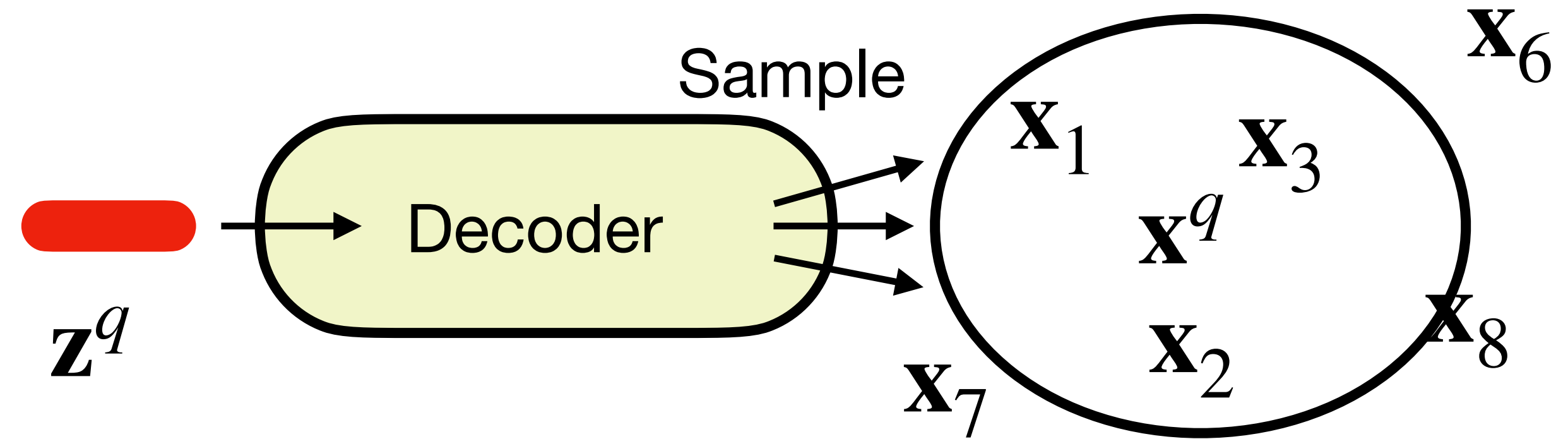


# Introduction & Methodology

Why does this idea make sense?



**New Question:** How to find the preimage  $B_{\mathbf{z}^q, f, \epsilon}$  efficiently ?



# Case study 1: Character Counting

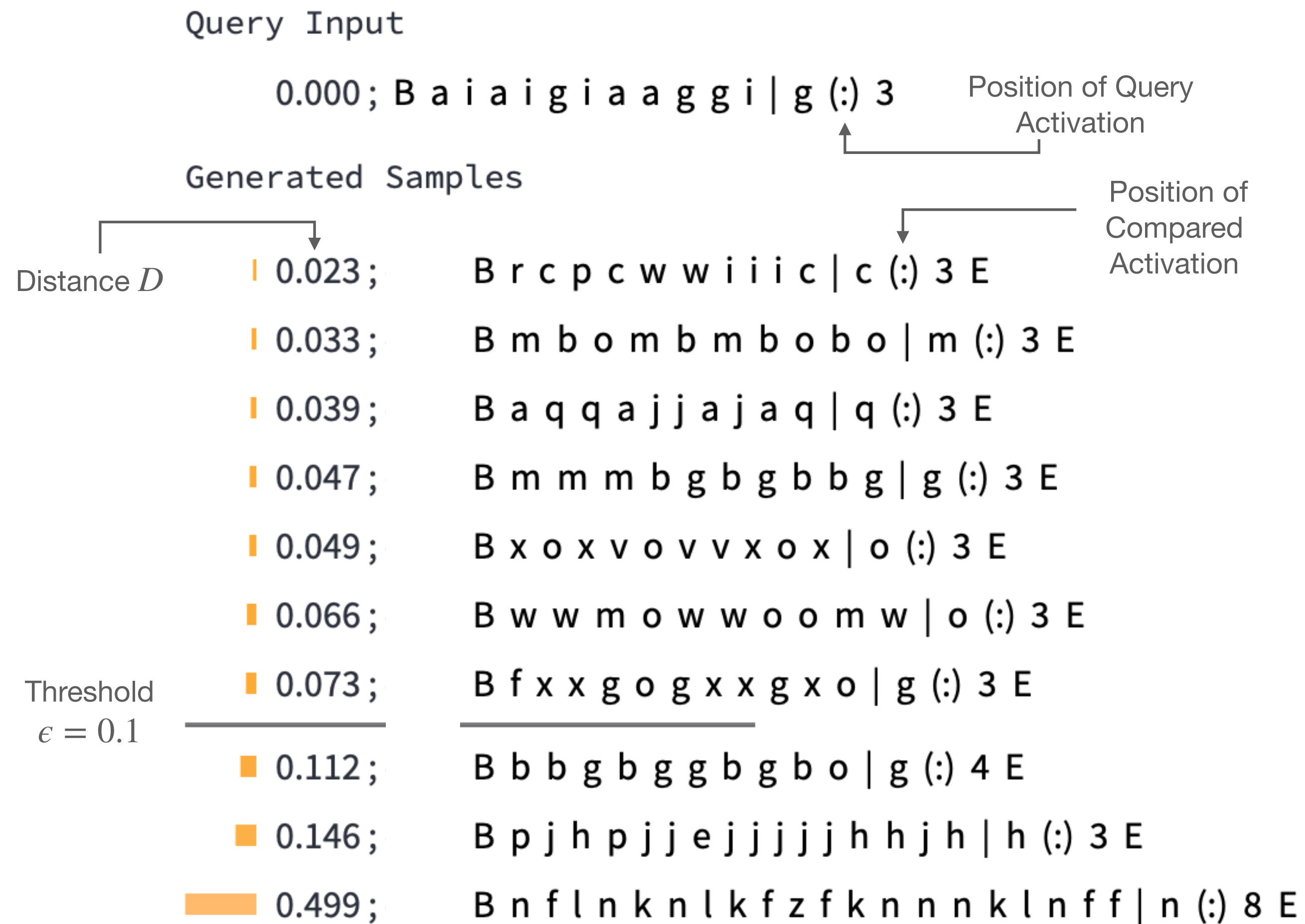
Query Input

B a i a i g i a a g g i | g (:) 3





# Case study 1: Character Counting





# Case study 1: Character Counting

Query Input

0.000; B a i a i g i a a g g i | g (:) 3

Position of Query  
Activation

Generated Samples

- | 0.023; #: 3; B r c p c w w i i i c | c (:) 3 E
- | 0.033; #: 3; B m b o m b m b o b o | m (:) 3 E
- | 0.039; #: 3; B a q q a j j a j a q | q (:) 3 E
- | 0.047; #: 3; B m m m b g b g b b g | g (:) 3 E
- | 0.049; #: 3; B x o x v o v v x o x | o (:) 3 E
- | 0.066; #: 3; B w w m o w w o o m w | o (:) 3 E
- | 0.073; #: 3; B f x x g o g x x g x o | g (:) 3 E

---

- 0.112; #: 4; B b b g b g g b g b o | g (:) 4 E
- 0.146; #: 4; B p j h p j j e j j j j j h h j h | h (:) 3 E
- 0.499; #: 7; B n f l n k n l k f z f k n n n k l n f f | n (:) 8 E

Position of  
Compared  
Activation

# Case study 2: IOI circuit in GPT-2 Small

## Query Input

0.000 ; <|endoftext|>After Erin and Justin went to the house, Erin gave a ring( to) Justin

## Generated Samples

- | 0.024 ; <|endoftext|>The station Sara and Justin went to had a kiss. Sara gave it( to) Justin[EOS]
- | 0.024 ; <|endoftext|>When Paul and Justin got a kiss at the school, Paul decided to give it( to) Justin[EOS]
- | 0.025 ; <|endoftext|>Then, Alicia and Justin had a long argument. Afterwards Alicia said( to) Justin[EOS]
- | 0.034 ; <|endoftext|>Then, Justin and Erin went to the garden. Erin gave a basketball( to) Justin[EOS]
- | 0.037 ; <|endoftext|>After the lunch in the afternoon, Justin and Kristen went to the station. Kristen gave a kiss( to) Justin[EOS]
- | 0.039 ; <|endoftext|>After taking a long break Kimberly and Justin went to the house, Kimberly gave a bone( to) Justin[EOS]
- | 0.043 ; <|endoftext|>While spending time together Justin and Alicia were working at the garden, Alicia gave a kiss( to) Justin[EOS]
- | 0.056 ; <|endoftext|>Then, Justin and Kristen went to the school. Kristen gave a bone( to) Justin[EOS]

---

- 0.506 ; <|endoftext|>Friends separated at birth Kristen and Justin found a snack at the garden. Justin gave it( to) Kristen[EOS]
- 0.598 ; <|endoftext|>While spending time together Michelle and Joshua were commuting to the restaurant, Alexander gave a ring( to) Michelle[EOS]

# Case study 3: 3-Digit Addition

Query Input

$$0.000; B\ 9\ 2\ 0 + 8\ 7\ 8 = (1)\ 7\ 9\ 8$$

Generated Samples

0.015; B 9 2 9 + 8 7 7 = (1) 8 0 6 E

0.041; B 9 6 4 + 8 3 2 = (1) 7 9 6 E

0.048; B 9 6 6 + 8 3 8 = (1) 8 0 4 E

0.063; B 9 1 1 + 8 8 5 = (1) 7 9 6 E

0.064; B 8 1 8 + 9 8 4 = (1) 8 0 2 E

0.066; B 8 1 6 + 9 8 0 = (1) 7 9 6 E

0.069; B 9 4 8 + 8 5 1 = (1) 7 9 9 E

---

0.120; B 8 0 5 + 9 9 6 = (1) 8 0 1 E

0.242; B 7 4 1 + 9 5 0 = (1) 6 9 1 E

0.406; B 8 3 4 + 9 7 7 = (1) 8 1 1 E

# Case study 4: Factual Recall

## Query Input

0.000 ; <|endoftext|>Joseph Schumpeter's domain of work is( the) economics

## Generated Samples

- 0.034 ; <|endoftext|>Joseph Schumpeter's domain of work( is) economics[EOS]
- 0.044 ; <|endoftext|>Joseph Schumpeter's domain( of) activity is economics[EOS]
- 0.050 ; <|endoftext|>Joseph Schumpeter('s) expertise is economics[EOS]
- 0.228 ; <|endoftext|>Adam Smith works( in) the area of economics[EOS]
- 0.242 ; <|endoftext|>Misesian economics consists of three aspects(.)[EOS]
- 0.246 ; <|endoftext|>Mises works( in) the field of economics[EOS]
- 0.248 ; <|endoftext|>Milton Friedman worked( in) the city of Paris[EOS]

---

- 0.442 ; <|endoftext|>".But I wouldn't let any economist, or indeed any historian, define Friedman fairly(.)[EOS]
- 0.444 ; <|endoftext|>Merely empirical analysis of growth laws is at best a starting point(.)[EOS]
- 0.487 ; <|endoftext|>Conrad Brehm took up( work) in Berlin[EOS]
- 0.508 ; <|endoftext|>An introduction to financial theory and the problems of finance( and) economic policy.[EOS]
- 0.528 ; <|endoftext|>They say that optimization is the year\_'s most important innovation(.)[EOS]
- 0.536 ; <|endoftext|>Foucault works( in) the field of history[EOS]

the samples are generated, so factual errors often occur

**Thank you**