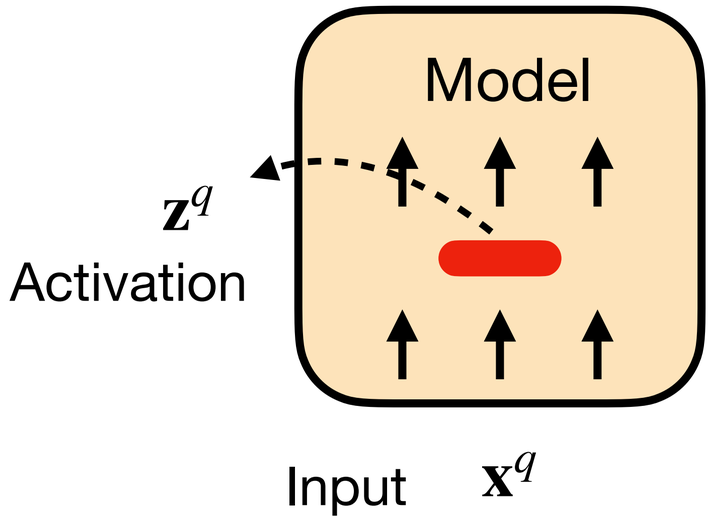
InversionView: A General-Purpose Method for Reading Information from Neural Activations

Xinting Huang, Madhur Panwar, Navin Goyal, Michael Hahn

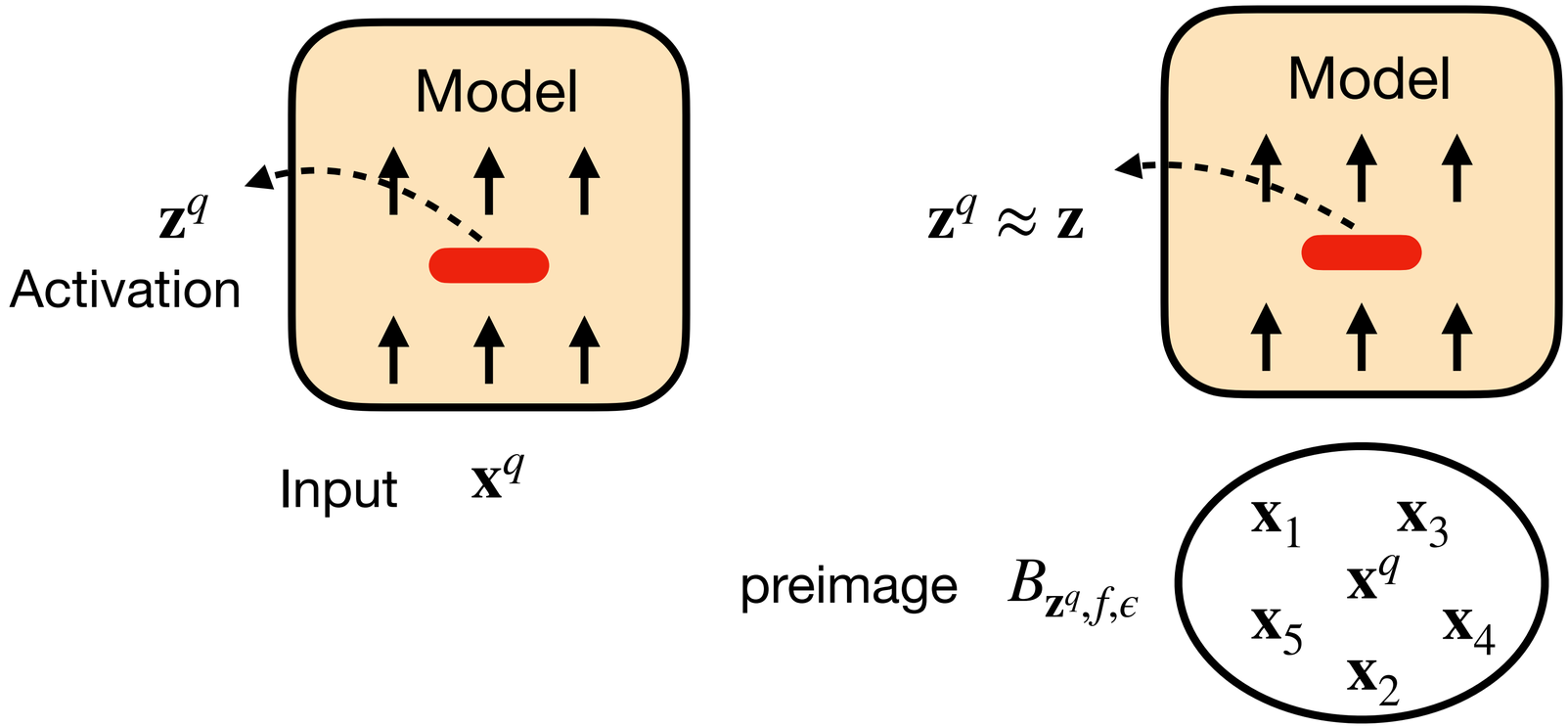



Xinting Huang, Nov 2024

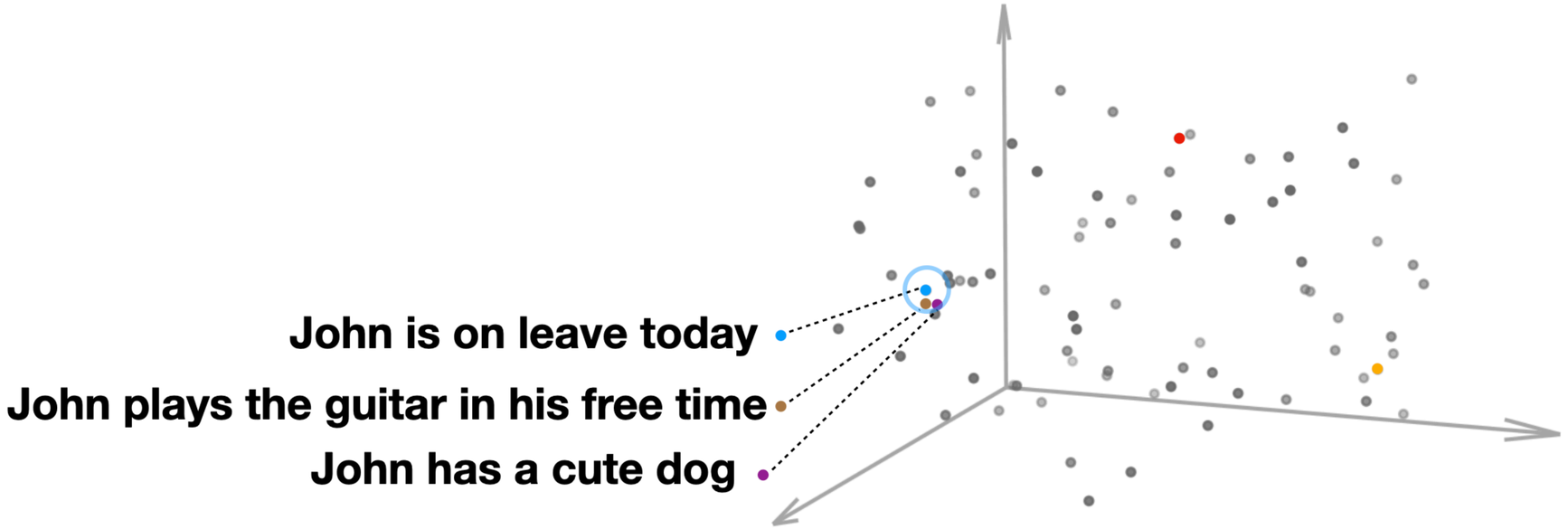
Research Question: What information is encoded in neural activations?



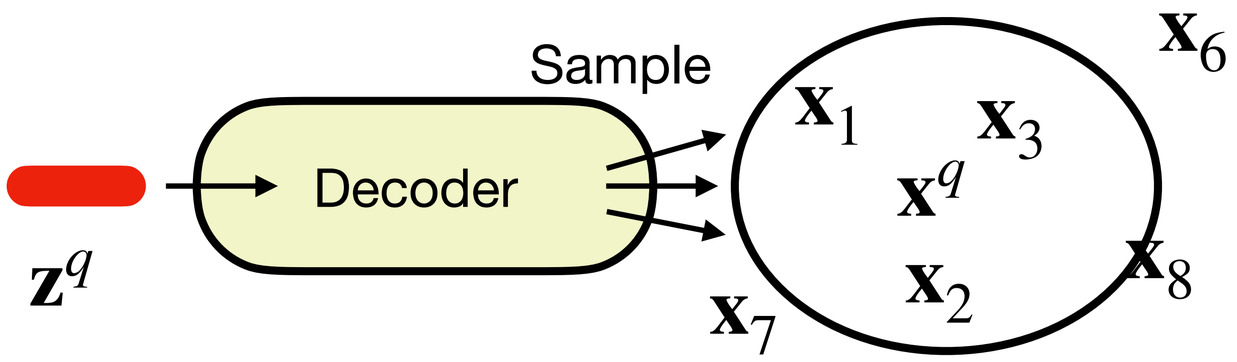
Research Question: What information is encoded in neural activations?



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

Case study 1: Character Counting

Query Input

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

Position of Query Activation

Generated Samples

B r c p c w w i i i c | c (:) 3 E

B m b o m b m b o b o | m (:) 3 E

B a q q a j j a j a q | q (:) 3 E

B m m m b g b g b b g | g (:) 3 E

B x o x v o v v x o x | o (:) 3 E

B w w m o w w o o m w | o (:) 3 E

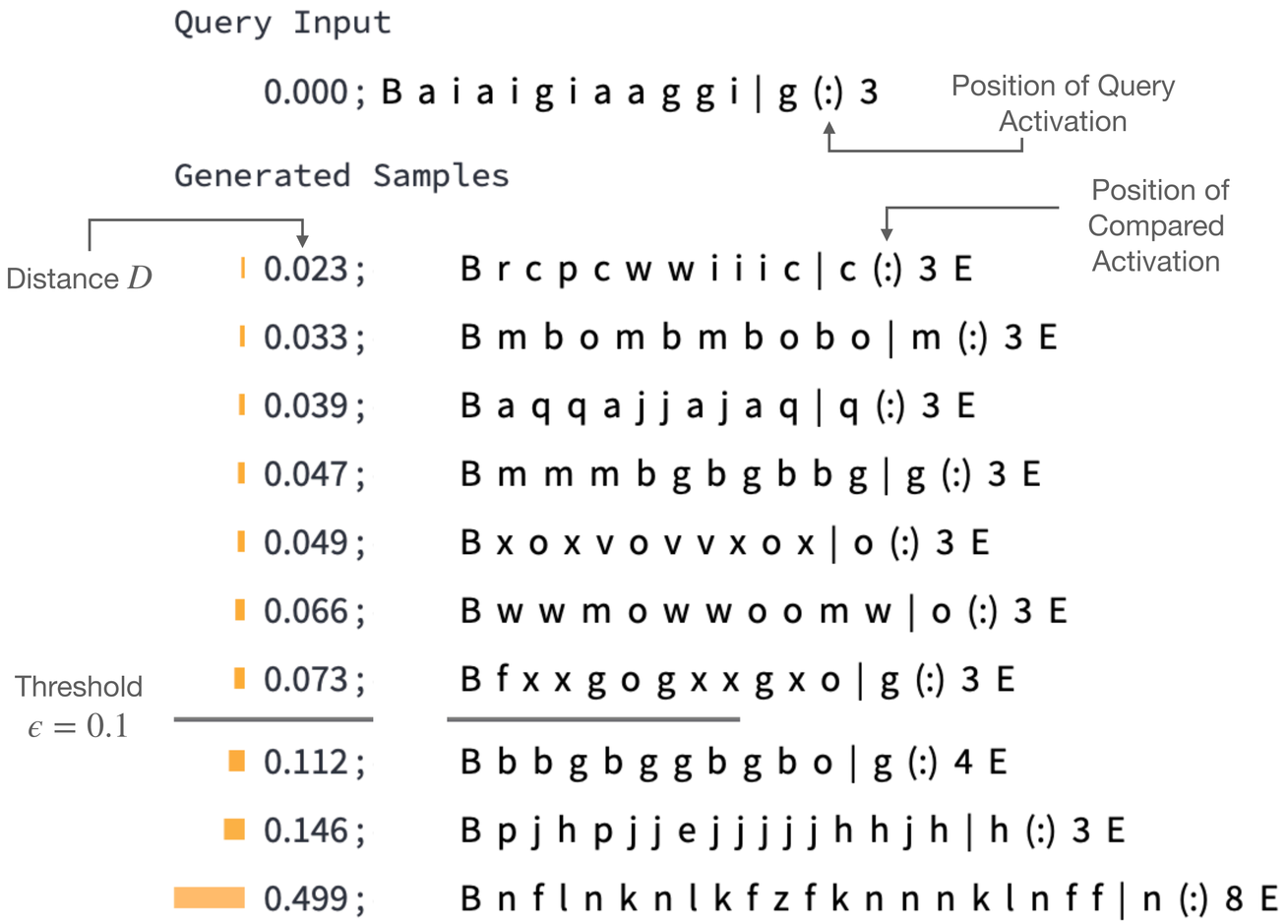
B f x x g o g x x g x o | g (:) 3 E

B b b g b g g b g b o | g (:) 4 E

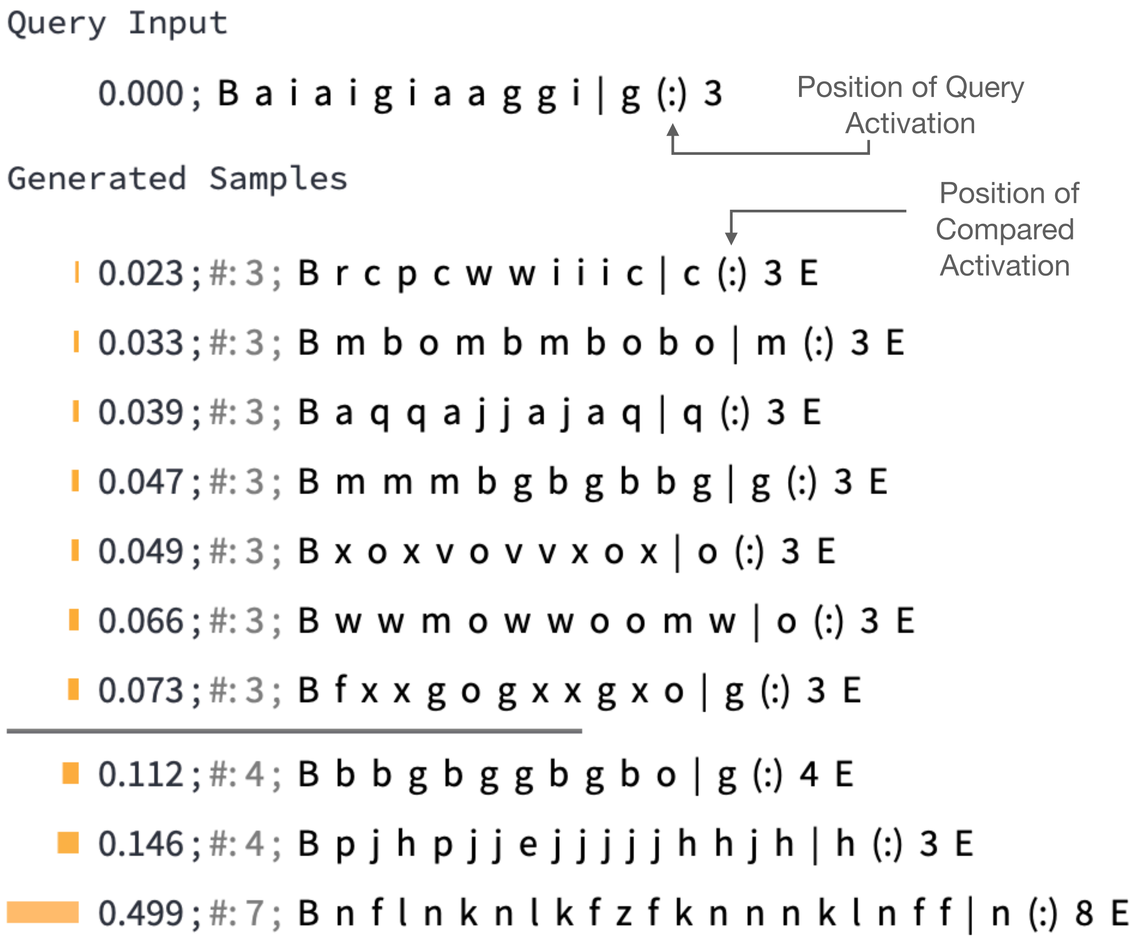
B p j h p j j e j j j j j h h j h | h (:) 3 E

B n f l n k n l k f z f k n n n k l n f f | n (:) 8 E

Case study 1: Character Counting



Case study 1: Character Counting



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]

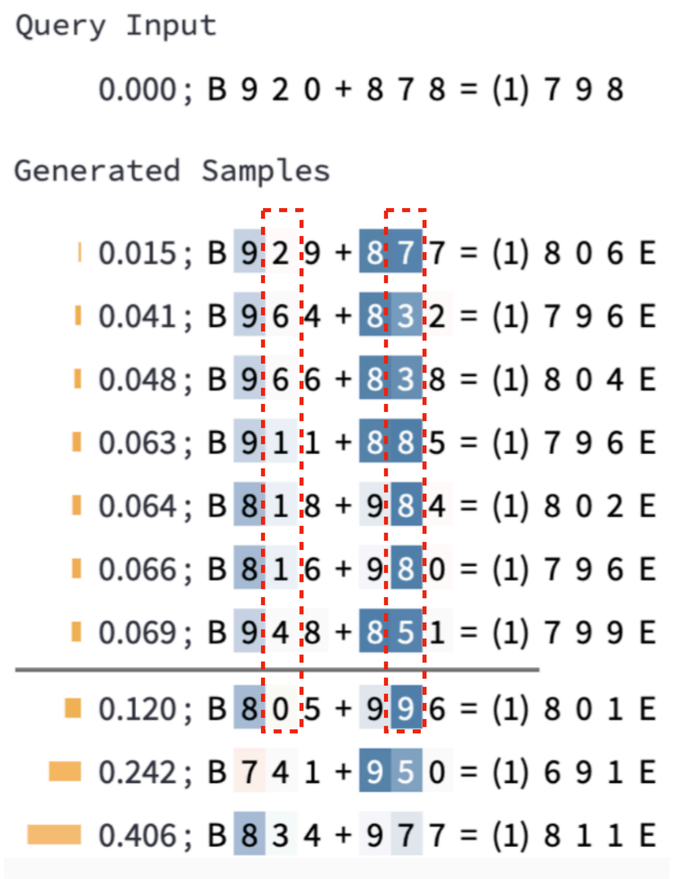
0.056 ; <|endoftext|>Then, Justin and Kristen went to the school. Kristen gave a bone(to) Justin[EOS]

0.506 ; <|endoftext|>Friends separated at birth Kristen and Justin found a snack at the garden. Justin gave it(to) Kristen[EOS]

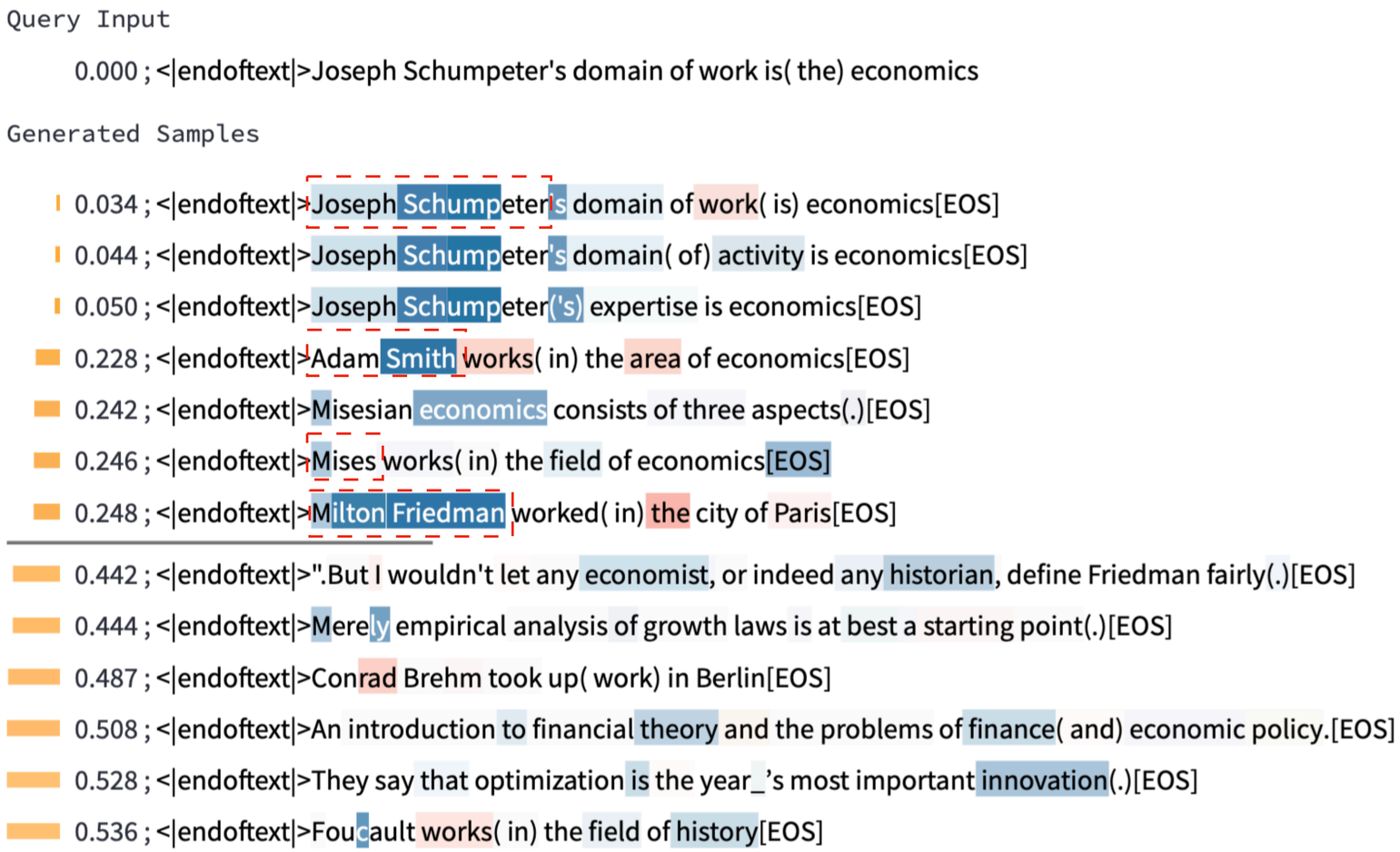
0.598 ; <|endoftext|>While spending time together Michelle and Joshua were commuting to the restaurant, Alexander gave a ring(to) Michelle[EOS]

Name Mover Head 9.9

Case study 3: 3-Digit Addition



Case study 4: Factual Recall



the samples are generated, so factual errors often occur

Thank you