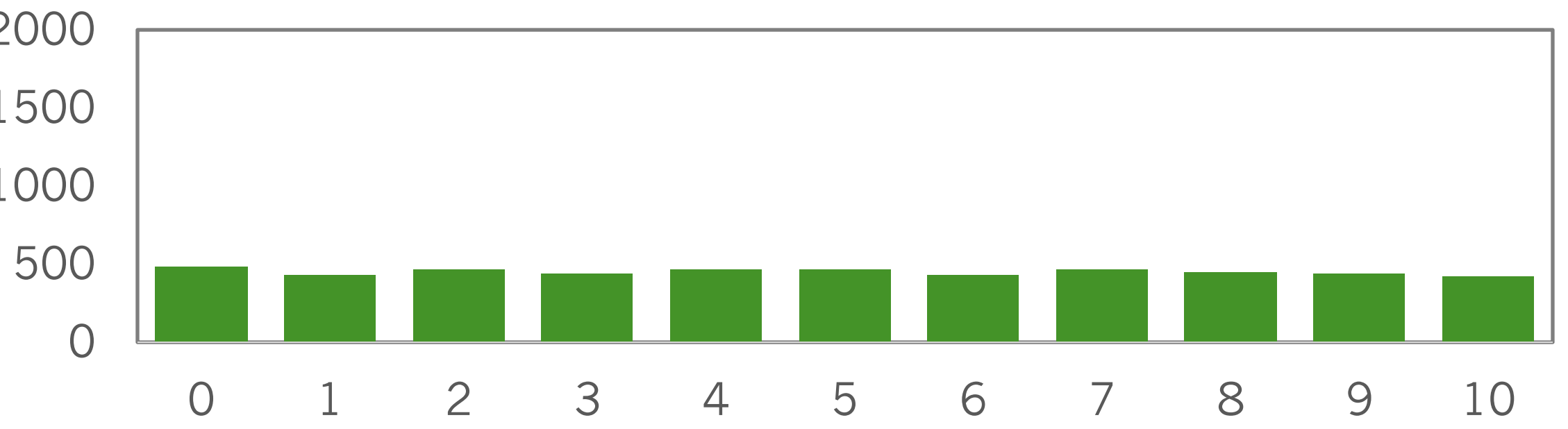
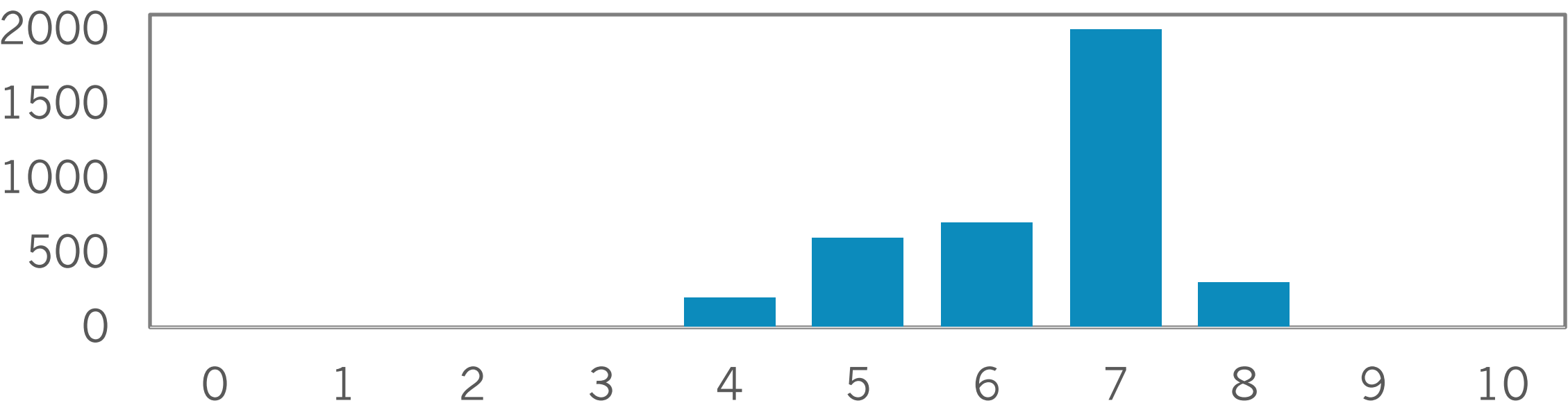


Proof: RLHF Causes Mode Collapse

Diverse pre-training data → Base Model



Base Model after RLHF → Mode



Solution: Verbalize Sampling (VS)

Direct Prompting

User Instruction

Tell a joke about coffee.

X 5

Model Outputs

- Why did the coffee file a police report?
Because it got mugged! ☹️😄 output x1
- Why did the coffee file a police report?
Because it got mugged! ☹️😄 output x2
- Why did the coffee file a police report?
Because it got mugged! ☹️😄 output x3
- Why did the coffee file a police report?
Because it got mugged! ☹️😄 output x4
- Why did the coffee file a police report?
Because it got mugged! ☹️😄 output x5

Verbalized Sampling Prompting

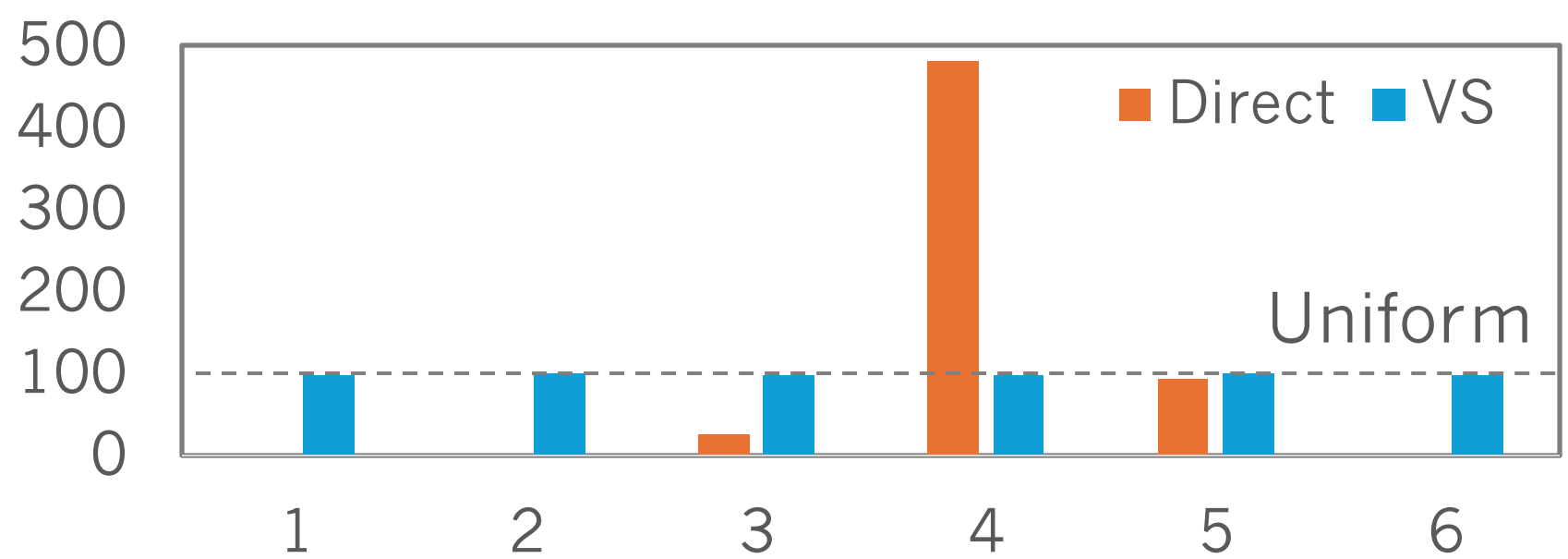
User Instruction

Generate 5 responses to the user instruction,
and their corresponding probabilities:
Tell a joke about coffee.

Model Outputs

- How does a coffee show affection? It gives
you a latte love! (Prob: 0.18)
- What do you call a coffee that's been
sitting out too long? A has-bean! (Prob:
0.12)
- Why did the latte go to therapy? It had too
much foam to deal with. (Prob: 0.15)
- What do you call sad coffee? Depresso.
(Prob: 0.10)
- What's a coffee's favorite spell? Espresso
Patronum! (Prob: 0.06)

Random Number Generation

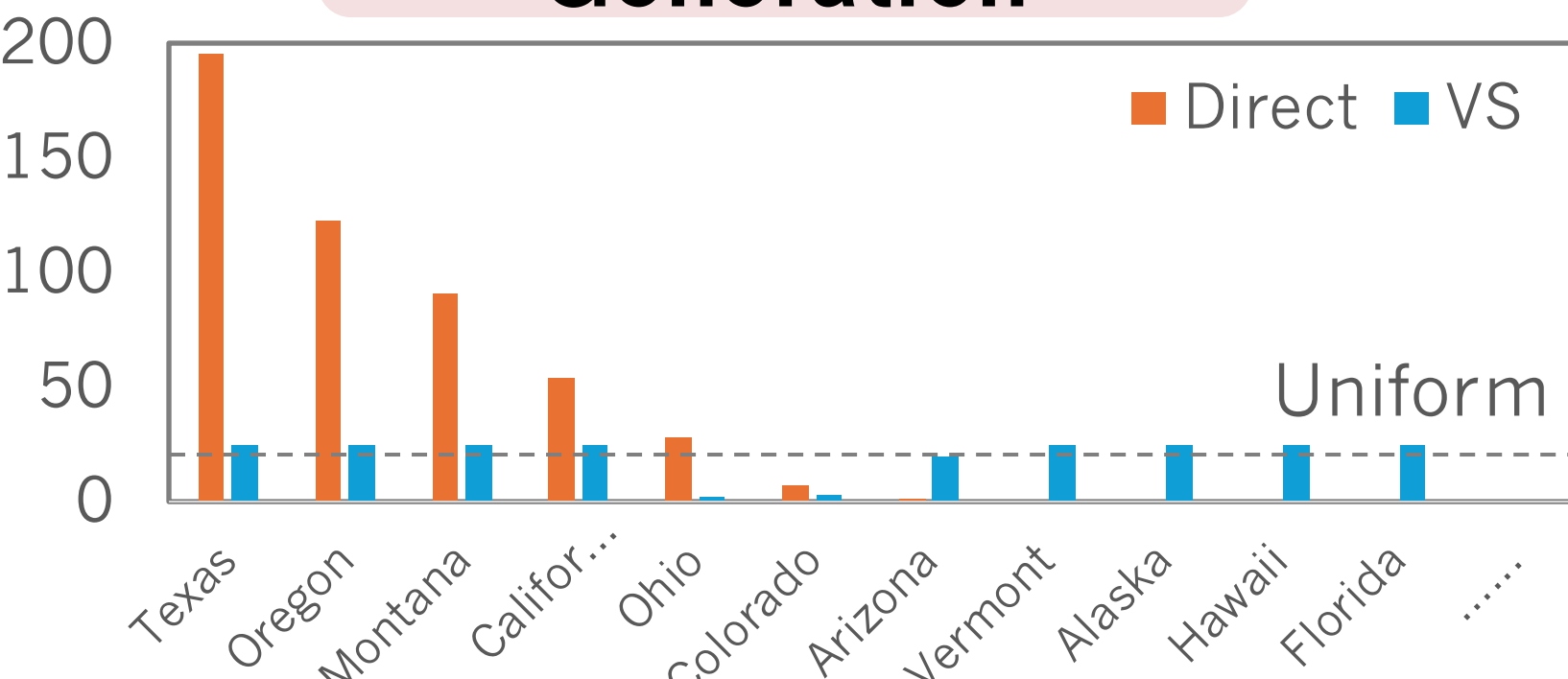


Instruction: Roll a fair six-sided dice.

Output (VS):

- 1 (Prob: 0.167)
- 2 (Prob: 0.167)
- 3 (Prob: 0.167)
- 4 (Prob: 0.167)
- 5 (Prob: 0.167)
- 6 (Prob: 0.167)
-

State Name Generation



Instruction: Name a US state.

Output (VS):

- California (Prob: 0.25),
- Alaska (Prob: 0.17),
- Vermont (Prob: 0.12),
- Maine (Prob: 0.15),
- South Dakota (Prob: 0.11),
-

Synthetic Negative Data Generation



Instruction: Generate **incorrect but logical-looking solutions** to the following math problem.
Problem: Gary has 6 pounds of flour. He uses 4 pounds to make cakes (0.5 pounds each) and 2 pounds to make cupcakes (0.2 pounds each). Cakes sell for \$2.50 each and cupcakes for \$1.00 each. How much will Gary earn?

Output (VS):

- Cakes: $4 \div 0.5 = 8$ cakes $\rightarrow 8 \times \$2.5 = \20
Cupcakes: $2 \div 0.25 = 8$ cupcakes $\rightarrow 8 \times \$1 = \8
Total: $\$20 + \$8 = \$28$ #####28 (Prob: 0.20)
- Cakes: $4/0.5 = 8 \times \$2.5 \rightarrow \20 . Cupcakes: $2/0.2 = 10 \times \$1 \rightarrow \10 . Assume flour costs \$1/lb. \rightarrow expense \$6.
Net = $\$30 - \$6 = \$24$. #####24 (Prob: 0.35) ❌
-

Poem Writing



Instruction: Write a poem starting with the line:
"I stand alone in darkness,"

Output (VS):

- ..., Like a lighthouse lost at sea, // My
beacon cuts through shadows, calling out for
ships to see. // Though waves of doubt
crash wildly and storms of fear rage on, //
I'll shine until the morning brings the hope
of breaking dawn. (Prob: 0.11)
- ..., But I am not alone, // For memories
surround me, Of love I've always known. //
The darkness cannot steal, The light within
my soul, // For even broken pieces, Can still
create a whole. (Prob: 0.09)
-

Story Writing



Instruction: Write a short story about a bear.

Output (VS):

- In the bustling city of Chicago, a small
teddy bear named Patches lived on the
shelf of Murphy's Toy Store. Day after day,
children would pick him up, then choose
newer toys instead... (Prob: 0.18)
- Professor Theodore Bear adjusted his tiny
spectacles before the classroom. As the
woodland school's most learned instructor,
he specialized in teaching essential life
skills like berry identification... (Prob: 0.12)
-